Real Analysis

Bruce Blackadar
Department of Mathematics and Statistics
University of Nevada, Reno
bruceb@unr.edu
Incomplete Preliminary Edition

March 1, 2026

THE CRITICS HAVE RAVED, AND I'M SURE THEY MUST HAVE BEEN TALKING ABOUT THIS BOOK:

THE PHENOMENAL #1 BEST SELLER!

THE BOOK EVERYBODY'S TALKING ABOUT!

NOW A MAJOR MOTION PICTURE

"Taut, fast, and relentless ... A ride worth taking. Keeps the reader hooked ... from the gripping first chapters ... to the vise-tightening midsection and on to the sensational finale."

Name withheld by request

A SAMIZDAT PUBLICATION

.1. Preface

> "[I] would have a man knockt in the head that should write anything in Mathematiques that had been written of before."
>
> *Edward Davenant*[1]

This manuscript represents my effort to produce a comprehensive text on real analysis, in a new organizational format possible only in an online book. The project is far from complete; indeed, it could be described as still being in its infancy. It will continue to evolve and grow in the future.

I have tried to take advantage of the increased organizational flexibility available for an online manuscript. A printed book must be organized linearly, but an online book need not be. Hyperlinks can make it possible to navigate through the various sections in a complicated and organic way. Supplemental material can be included without interfering with the flow of the core material (which can itself be varied according to the interests of the reader). While I am certainly not the first author to think of trying to take advantage of this flexibility in organizing an online book, I have been surprised to find that the possibilities have been largely unexplored so far, and the potentially revolutionary benefits of flexible organization online have been little exploited. Even TEX (or at least LATEX), which has become the universally-used text editing system for mathematics and related fields (and which I have used for this manuscript), is too inflexibly oriented toward producing linearly-organized printed documents. I hope my approach to organization of this manuscript will help encourage other attempts to develop the possibilities. See I.8. for more detailed comments about organization. The organization is not yet what I ultimately envision; I have added a topic index at the end of this preface as a partial first step.

This book is intended as a supplement, not a replacement, for the many fine books on this subject already available. I have consulted existing texts and references extensively, and they have collectively been very influential in both the general approach and specific details of my treatment. I have tried to be careful in giving credit where credit is due; however, in a subject as generally mature as real analysis, there is little originality in the actual mathematics of most books (including this one) – the primary contributions of each author are the exposition and organization, and decisions on what material to emphasize and what to leave out (the online organization of this book has made the decision criteria considerably different for me than for authors of printed books).

No part of this manuscript should be regarded as the last word, or even *my* last word, on the subject. My approach on many topics is pretty standard, and in a lot of cases will probably remain so; much of this subject has been quite refined for a long time. But my approach to some parts of the subject has changed already during the project (frankly, I have been forced to think critically about the treatment of some topics for the first time in my career), and this process will surely continue.

I regard myself as a competent mathematician who is reasonably qualified for a project of this sort, and I think I have learned something about the right way to present this subject in my 45 years of experience in teaching analysis at the university level. But I am in no way uniquely or even especially well qualified; I think there are many other mathematicians who could do at least as good a job, in some cases much better, on all or parts of the project. The only difference is that I actually have done it.

I expect and welcome comment and criticism concerning my efforts, either through personal communication or public statements. I will take all (informed) criticism to heart, and it will perhaps result in improvements in the manuscript. Certainly simple corrections to errors and misstatements (of which there are likely to be a substantial number despite my best efforts to be careful) will be greatly appreciated and

[1] J. Aubrey, *Brief Lives*; cf. [Bil95, p. v].

quickly acted on. Broader criticism of the text is also welcome. However, I issue the following challenge to anyone who is dissatisfied with what is here: don't just criticize me, try to do better yourself. I do not regard myself as being in competition with anyone; my goal is simply to have the best treatment of this subject available to the community at large. Criticism of my manuscript makes a small contribution to this goal, but anything you write which is better than what I write benefits us all much more.

Topical Index

This topical index is abbreviated for easier use. For a full guide to topics, see the Table of Contents or the Index at the end.

absolute continuity
algebraic topology
analytic function
axiom of choice
Banach space
Borel set
bounded variation
calculus of variations
cardinality
C*-algebra
category
compactness
complex analysis
complex numbers
continuity
convex function
convex set
countable set
curve
derivative
differential equations
differential forms
distribution
Dominated Convergence
exponential function
Fubini's Theorem
function
functional analysis
functor
Hilbert space
homological algebra
homotopy
Implicit Function Theorem
induction
infinite series
interpolation
Lebesgue integral
Lebesgue measure
l'Hôpital's Rule
Lie group
literature
logic
manifold
Mean Value Theorem
measurable set
measurable space
measure space
metric space
monotone function
multilinear algebra
multivariable calculus
nonstandard analysis
norm
partial order
power series
probability
Radon-Nikodým Theorem
real numbers
relation
Riemann integral
rigor
sequences
set theory
smooth function
Sobolev space
Taylor series
Taylor's Theorem
topological group
topological space
topological vector space
topology
transcendental number
trigonometric function
uncountable set
uniform space
vector space
Zorn's Lemma

Contents

.1. Preface ii

Contents **iv**

I Introduction **1**
I.1. Calculus and Analysis 1
I.2. Differential Calculus 2
I.3. Measure and Integration 2
I.3.1. Measure 2
I.3.2. Integration 6
I.4. Other Topics 8
I.5. What is Mathematics, and Why the Bad Reputation? 9
I.6. Rigor in Mathematics 13
I.6.1. Semantics: What is Obvious? 21
I.6.2. Set Theory and Logic 56
I.6.3. Proofs vs. Formal Proofs (or, How Mathematics Really Works) 57
I.6.4. How to Understand a Proof 61
I.6.5. Proof by Contradiction 62
I.6.6. Proofs and Counterexamples: Mathematics as Creative Activity 65
I.6.7. Scholarship vs. Practice 77
I.6.8. Is Mathematics Created or Discovered? 77
I.6.9. Analogies for the Structure of Mathematics 79
I.6.10. Pure vs. Applied Mathematics 83
I.6.11. Is Mathematics an Art or a Science? 86
I.7. The Foundations of Mathematics 91
I.7.1. Philosophies of Mathematics 91
I.7.2. Mathematical Theories 118
I.7.3. Languages 118
I.7.4. Logic 120
I.7.5. Mathematical Structures and Models 121
I.7.6. Exercises 121
I.8. Printed or Online? 122
I.8.1. How Much Generality? 124
I.8.2. History vs. Logic 125

II The Formalities of Mathematics **127**
II.1. Formal Statements, Negations, and Mathematical Arguments . . . 128
II.1.1. Formal Statements . . . 129
II.1.2. Meaning of Statements . . . 133
II.1.3. Quantification and Closed Statements . . . 135
II.1.4. Negations . . . 139
II.1.5. Uniqueness Statements . . . 142
II.1.6. Tautologies . . . 142
II.1.7. Definitions . . . 144
II.1.8. Languages and Relational Systems . . . 147
II.1.9. Mathematical Arguments and Formal Proofs . . . 151
II.1.10. Exercises . . . 151
II.2. Set Theory . . . 156
II.2.1. New Sets from Old . . . 156
II.3. Relations and Functions . . . 157
II.3.1. General Relations . . . 157
II.3.2. Functions . . . 159
II.3.3. Equivalence Relations . . . 163
II.3.4. Exercises . . . 165
II.4. Ordered Sets and Induction . . . 167
II.4.1. Partially Ordered Sets . . . 167
II.4.2. Well-Ordered Sets and Transfinite Induction . . . 169
II.4.3. Exercises . . . 174
II.4.4. Directed Sets . . . 177
II.4.5. Exercises . . . 179
II.5. Cardinality of Sets . . . 181
II.5.1. Cardinality . . . 181
II.5.2. Finite and Infinite Sets . . . 182
II.5.3. Countable and Uncountable Sets . . . 184
II.5.4. Properties of Countable Sets . . . 186
II.5.5. Additional Facts about Cardinality . . . 190
II.5.6. Exercises . . . 190
II.6. The Axiom of Choice and Zorn's Lemma . . . 195
II.6.1. The Axiom of Choice . . . 195
II.6.2. Equivalent Formulations . . . 196
II.6.3. Other Forms of Choice . . . 199
II.6.4. Exercises . . . 203
II.7. Axiomatic Set Theory . . . 208
II.7.1. The ZF Axioms . . . 209
II.7.2. The Axiom of Regularity . . . 219
II.7.3. Controversy . . . 220
II.7.4. The ZF Axioms (Recapitulation) . . . 221
II.7.5. Models of ZF(C) Set Theory . . . 222
II.7.6. The NBG Axioms . . . 223
II.7.7. The MK Axioms . . . 223
II.7.8. Set Theory with Atoms . . . 223

II.7.9. Exercises . . . 224
II.7.10. The Theory of Set Theories . . . 225
II.7.11. Are the Foundations of Mathematics Circular? . . . 226
II.8. Ordinals . . . 227
II.8.1. Basic Definitions . . . 227
II.8.2. Ordering on Ordinals . . . 227
II.8.3. Transfinite Induction . . . 229
II.8.4. Ordinals as Universal Models . . . 230
II.8.5. The Axiom of Choice and Zorn's Lemma . . . 230
II.8.6. Ordinal Arithmetic and Examples . . . 231
II.8.7. Induction and Recursion . . . 236
II.8.8. The Theory of Levels . . . 238
II.8.9. Exercises . . . 239
II.9. Cardinals . . . 244
II.9.1. Cardinals as Ordinals . . . 244
II.9.2. General Cardinals and Alephs . . . 246
II.9.3. Cardinal Arithmetic . . . 247
II.9.4. The Role of the Axiom of Choice . . . 248
II.9.5. Dedekind-Finite Sets . . . 251
II.9.6. Exercises . . . 257
II.9.7. The Continuum Hypothesis . . . 258
II.9.8. Large Cardinals . . . 259
II.9.9. Exercises . . . 260
II.10. Measurable Cardinals . . . 265
II.10.1. Introduction . . . 265
II.10.2. Real-Valued Measurable Cardinals . . . 266
II.10.3. Atomless Full Real-Valued Measures . . . 268
II.10.4. Measurable Cardinals . . . 270
II.10.5. Consistency and Equiconsistency Results . . . 274
II.11. Categories and Functors . . . 277
II.11.1. Mathematical Theories . . . 277
II.11.2. Categories . . . 278
II.11.3. Functors . . . 280
II.11.4. Natural Transformations . . . 282
II.11.5. Equivalence of Categories and Skeletons . . . 283
II.11.6. Adjoint Functors . . . 285
II.11.7. The Category of Small Categories . . . 285
II.11.8. Categories as a Foundation of Mathematics . . . 286
II.11.9. Exercises . . . 287

III The Numbers of Analysis **291**
III.1. The Real Numbers – Axiomatic Approach . . . 291
III.1.1. The Algebraic Axioms of $\mathbb{R}$ . . . 292
III.1.2. The Order Axioms for $\mathbb{R}$ . . . 295
III.1.3. Absolute Value . . . 298
III.1.4. The Natural Numbers . . . 298

III.1.5. Induction and Recursion . . . 300
III.1.6. The Integers . . . 305
III.1.7. The Rational Numbers . . . 309
III.1.8. The Number Line . . . 310
III.1.9. Suprema, Infima, and the Completeness Axiom . . . 311
III.1.10 The Archimedean Property and Density of $\mathbb{Q}$ . . . 316
III.1.11 Uniqueness of $\mathbb{R}$ . . . 317
III.1.12 Existence of Roots . . . 321
III.1.13 The Extended Real Line . . . 323
III.1.14 Exercises . . . 325
III.2. The Natural Numbers . . . 327
III.2.1. Introduction . . . 327
III.2.2. Axiomatic Characterization of $\mathbb{N}$ . . . 327
III.2.3. Induction and Recursion . . . 331
III.2.4. Algebraic Structure of $\mathbb{N}$ . . . 332
III.2.5. Algebraic Axiomatization of $\mathbb{N}$ . . . 335
III.2.6. Order Structure of $\mathbb{N}$ . . . 337
III.2.7. Subtraction and Division in $\mathbb{N}$ . . . 339
III.2.8. Axiomatization of $\mathbb{N}_0$ . . . 339
III.2.9. Further Properties of $\mathbb{N}$ . . . 341
III.2.10 Summary . . . 344
III.2.11 Exercises . . . 344
III.3. The Integers . . . 345
III.3.1. Introduction and Axiomatization of $\mathbb{Z}$ . . . 345
III.3.2. Models for $\mathbb{Z}$ . . . 346
III.3.3. Uniqueness of $\mathbb{Z}$ . . . 347
III.3.4. Ordering and Subtraction in $\mathbb{Z}$ . . . 349
III.3.5. Multiplication and Division in $\mathbb{Z}$ . . . 349
III.3.6. Ring Axiomatization of $\mathbb{Z}$ . . . 351
III.3.7. Exercises . . . 353
III.4. The Rational Numbers . . . 355
III.4.1. Introduction and Axiomatization of $\mathbb{Q}$ . . . 355
III.4.2. The Standard Model of $\mathbb{Q}$ . . . 356
III.4.3. Exercises . . . 357
III.5. The Real Numbers . . . 359
III.5.1. Construction of $\mathbb{R}$: the Dedekind Cut Approach . . . 359
III.5.2. Construction of $\mathbb{R}$: the Cauchy Sequence Approach . . . 368
III.5.3. Exercises . . . 374
III.6. The Complex Numbers . . . 380
III.6.1. Axiomatization of $\mathbb{C}$ . . . 380
III.6.2. The Standard Model for $\mathbb{C}$: the Complex Plane . . . 382
III.6.3. Exercises . . . 387
III.7. Fundamental Theorems About $\mathbb{R}$ . . . 389
III.7.1. Partitions . . . 389
III.7.2. Compactness and Continuity . . . 390
III.7.3. The Creeping Lemma . . . 393

III.7.4. Exercises . . . 394
III.8. Representation of Real Numbers . . . 395
III.8.1. Decimal Expansions . . . 395
III.8.2. Continued Fractions . . . 402
III.8.3. Measure Theory of Continued Fractions . . . 415
III.8.4. Exercises . . . 417
III.9. Irrational and Transcendental Numbers . . . 422
III.9.1. Introduction . . . 422
III.9.2. Irrationality of π . . . 422
III.9.3. Transcendence of e and Related Numbers . . . 424
III.9.4. Twentieth Century Results . . . 427
III.9.5. Exercises . . . 430
III.10 Hyperreal Numbers and Nonstandard Analysis . . . 434
III.10.1 Ordered Fields Containing $\mathbb{R}$ . . . 436
III.10.2 The Hyperreal Numbers via Ultrafilters . . . 438
III.10.3 Internal and External Sets and Functions . . . 441
III.10.4 The Transfer Principle . . . 445
III.10.5 Other Nonstandard Structures . . . 451
III.10.6 Proofs and Applications of Nonstandard Analysis . . . 451
III.10.7 Loeb Measures and a Nonstandard Approach to Measure Theory . . . 451
III.10.8 Exercises . . . 451
III.11 Linear Algebra . . . 455
III.11.1 Matrix Algebra . . . 455
III.11.2 Vector Spaces . . . 462
III.11.3 Linear Transformations . . . 475
III.11.4 The Matrix of a Linear Transformation . . . 480
III.11.5 Linear Functionals and Dual Spaces . . . 483
III.11.6 Covariant vs. Contravariant . . . 486
III.11.7 More on Matrices and Linear Equations . . . 487
III.11.8 Affine Subspaces and Affine Functions . . . 493
III.11.9 The Dot Product on $\mathbb{R}^n$ and Adjoints . . . 494
III.11.10 Determinants . . . 496
III.11.11 Eigenvalues and Eigenvectors . . . 510
III.11.12 Exercises . . . 512
III.11.13 Multilinear Algebra . . . 514
III.11.14 Tensor Products . . . 514
III.11.15 Modules . . . 522
III.11.16 The Tensor Algebra of a Vector Space . . . 527
III.11.17 The Exterior Algebra of a Vector Space . . . 530
III.11.18 Mixed Tensors and General Tensor Algebra . . . 540
III.11.19 Exercises . . . 541
III.12 The Banach-Tarski Paradox . . . 545
III.12.1 Rigid Motions . . . 545
III.12.2 Rigid Equivalence . . . 546
III.12.3 Equivalence by Dissection . . . 547
III.12.4 Equidecomposability . . . 547

III.12.5Free Groups and Paradoxical Decompositions . . . 550
III.12.6The Hausdorff Paradox . . . 553
III.12.7The Banach-Tarski Paradox, Basic Version . . . 555
III.12.8The Banach-Tarski Paradox, General Version . . . 557
III.12.9Exercises . . . 558

IV Sequences and Series **560**
IV.1. Sequences . . . 560
IV.1.1. Sequences of Numbers . . . 561
IV.1.2. Subsequences and Tails . . . 561
IV.1.3. Limits of Sequences . . . 561
IV.1.4. Infinite Limits . . . 571
IV.1.5. Monotone Sequences . . . 571
IV.1.6. Subsequences and Cluster Points of Sequences . . . 571
IV.1.7. The Bolzano-Weierstrass Theorem . . . 572
IV.1.8. Limit Superior and Limit Inferior . . . 572
IV.1.9. Cauchy Sequences . . . 578
IV.1.10Sequences in $\mathbb{R}^n$ . . . 579
IV.1.11Sequences in Metric Spaces . . . 579
IV.1.12Exercises . . . 579
IV.2. Infinite Series . . . 582
IV.2.1. Partial Sums and Convergence . . . 582
IV.2.2. Nonnegative Series . . . 590
IV.2.3. Tests for Convergence of Nonnegative Series . . . 596
IV.2.4. More Delicate Tests . . . 602
IV.2.5. Absolute and Conditional Convergence . . . 611
IV.2.6. Hypergeometric Series . . . 616
IV.2.7. Scaling of Series . . . 620
IV.2.8. Subseries and Rearrangements of Series . . . 624
IV.2.9. Unordered Summations . . . 630
IV.2.10Double Summations . . . 635
IV.2.11Products of Infinite Series . . . 644
IV.2.12Infinite Products . . . 648
IV.2.13Infinite Series of Complex Numbers . . . 653
IV.2.14Numerical Evaluation and Estimation of Series . . . 654
IV.2.15Summability Methods for Divergent Series . . . 654
IV.2.16Exercises . . . 655
IV.3. Convergence of Sequences of Functions . . . 664
IV.3.1. Pointwise and Uniform Convergence . . . 664
IV.3.2. U.C. Convergence . . . 665
IV.3.3. Mean and Mean-Square Convergence . . . 667
IV.3.4. Exercises . . . 667

V Calculus **669**
V.1. Limits of Functions . . . 670
V.1.1. Ordinary Limits . . . 670
V.1.2. The Sequential Criterion . . . 673
V.1.3. Limit Theorems . . . 674
V.1.4. Indeterminate Forms . . . 676
V.1.5. One-Sided Limits . . . 676
V.1.6. Extended Limits . . . 676
V.1.7. Limit Superior and Limit Inferior . . . 676
V.1.8. Exercises . . . 676
V.2. Continuity on $\mathbb{R}$ . . . 678
V.2.1. Continuity at a Point . . . 678
V.2.2. Continuity on a Set . . . 678
V.2.3. The Intermediate Value Theorem . . . 680
V.2.4. The Max-Min Theorem . . . 682
V.2.5. Continuity and Oscillation . . . 683
V.3. Differentiation and Derivatives . . . 685
V.3.1. The Derivative . . . 686
V.3.2. Elementary Rules of Differentiation . . . 693
V.3.3. The Product and Quotient Rules . . . 694
V.3.4. The Chain Rule . . . 698
V.3.5. Derivatives of Inverse Functions . . . 700
V.3.6. Critical Points and Local Extrema . . . 700
V.3.7. Higher Derivatives . . . 703
V.3.8. Continuous but Nondifferentiable Functions . . . 705
V.3.9. The Bush-Wunderlich-Swift Example . . . 707
V.3.10. Exercises . . . 712
V.3.11. Calculus of Finite Differences . . . 716
V.3.12. Exercises . . . 723
V.4. Extensions of the Derivative . . . 724
V.4.1. One-Sided Derivatives . . . 724
V.4.2. Dini Derivatives . . . 725
V.4.3. Strong Derivatives . . . 729
V.4.4. Symmetric First and Second Derivatives . . . 729
V.4.5. Approximate and Weak Derivatives . . . 729
V.4.6. Exercises . . . 729
V.5. Exponential and Trigonometric Functions . . . 730
V.5.1. Rational Exponents . . . 730
V.5.2. Irrational Exponents and Exponential Functions . . . 731
V.5.3. The Number e . . . 733
V.5.4. Derivatives of Exponential and Logarithm Functions . . . 735
V.5.5. Exercises . . . 736
V.6. Trigonometric Functions . . . 740
V.6.1. Arc Length of a Circle and the Arcsine Function . . . 740
V.6.2. The Sine and Cosine Functions . . . 741
V.6.3. Other Trigonometric Functions . . . 743

V.6.4. The Inverse Trigonometric Functions . . . 744
V.6.5. The Area of a Circle . . . 745
V.6.6. The Hyperbolic Functions . . . 746
V.6.7. Exercises . . . 750
V.7. Big-O and Little-o Notation . . . 752
V.7.1. Exercises . . . 757
V.8. The Mean Value Theorem . . . 758
V.8.1. Rolle's Theorem and the Mean Value Theorem . . . 758
V.8.2. Applications of the Mean Value Theorem . . . 761
V.8.3. Geometric Interpretation of the Second Derivative . . . 765
V.8.4. Refinements of the Mean Value Theorem . . . 767
V.8.5. Antiderivatives and Primitives . . . 770
V.8.6. Exercises . . . 774
V.9. L'Hôpital's Rule . . . 778
V.9.1. L'Hôpital's Rule, Version 1 . . . 778
V.9.2. L'Hôpital's Rule, Version 2 . . . 779
V.9.3. Applications . . . 784
V.9.4. Exercises . . . 787
V.10. Taylor's Theorem . . . 791
V.10.1. Taylor Polynomials . . . 791
V.10.2. Vanishing Order . . . 794
V.10.3. Remainder Formulas . . . 795
V.10.4. The Integral Remainder Formula . . . 797
V.10.5. The Morse Lemma . . . 798
V.10.6. Applications . . . 800
V.10.7. Exercises . . . 800
V.11. Interpolation and Extrapolation . . . 804
V.11.1. The Interpolating Polynomial . . . 804
V.11.2. Linear Interpolation and Extrapolation . . . 805
V.11.3. Higher-Order Interpolation and Extrapolation . . . 806
V.11.4. The Mean Value Theorem for Divided Differences . . . 807
V.11.5. Repeated Points . . . 808
V.11.6. Osculating Polynomials . . . 812
V.11.7. How Good are the Approximations? . . . 814
V.11.8. Exercises . . . 818
V.12. Miscellaneous Topics in Calculus . . . 821
V.12.1. How Discontinuous Can a Derivative Be? . . . 821
V.12.2. Other Properties of Derivatives . . . 823
V.12.3. Analysis on $\mathbb{Q}$ . . . 824
V.12.4. Exercises . . . 826
V.12.5. The Binomial Theorem . . . 829
V.12.6. Exercises . . . 832
V.13. The Gamma Function . . . 834
V.13.1. Definition of the Gamma Function . . . 834
V.13.2. Log-Convexity and a Characterization of the Gamma Function . . . 836
V.13.3. Miscellaneous Additional Results . . . 839

V.13.4. Exercises . . . 840
V.13.5. Stirling's Formula . . . 841
V.13.6. Exercises . . . 845
V.14. The Volume of the Unit Ball in $\mathbb{R}^n$ . . . 846
V.14.1. Radial Borel Sets and Functions . . . 846
V.14.2. Calculation of β_n . . . 848
V.14.3. Surface Area of a Sphere . . . 849
V.14.4. Exercises . . . 851
V.15. Infinite Series of Functions . . . 853
V.15.1. Infinite Series of Functions . . . 853
V.15.2. Exercises . . . 855
V.15.3. Power Series . . . 857
V.16. Analytic Functions . . . 867
V.16.1. Definitions and Basic Properties . . . 867
V.16.2. Rigidity of Analytic Functions . . . 871
V.16.3. Combinations of Analytic Functions . . . 874
V.16.4. The Analytic Morse Lemma . . . 883
V.17. Smooth Non-Analytic Functions . . . 885
V.17.1. Analyticity at a Point . . . 885
V.17.2. (C)-Points and Smooth Partitions . . . 886
V.17.3. (P)-Points and Borel's Theorem . . . 889
V.17.4. Everywhere Nonanalytic Functions . . . 893
V.17.5. Exercises . . . 900

VI The Topology of $\mathbb{R}$, $\mathbb{R}^n$, and Metric Spaces **901**
VI.1. The Topology of $\mathbb{R}$ . . . 901
VI.1.1. Open and Closed Sets in $\mathbb{R}$ . . . 901
VI.1.2. Limit Points and Closures . . . 903
VI.1.3. Compactness and Continuity . . . 904
VI.2. Metric Spaces . . . 908
VI.2.1. Definitions . . . 908
VI.2.2. Balls, Limit Points, and Closures . . . 909
VI.2.3. Limits and Continuity . . . 912
VI.2.4. Continuity Without the Countable AC . . . 916
VI.2.5. Extended Metrics . . . 916
VI.2.6. Cauchy Sequences . . . 918
VI.2.7. Normed Vector Spaces . . . 918
VI.2.8. Exercises . . . 920
VI.2.9. How Many Separable Metric Spaces Are There? . . . 924
VI.2.10 The Banach Fixed-Point Theorem . . . 926
VI.2.11 Exercises . . . 928
VI.3. Euclidean Space . . . 929
VI.3.1. Metric Axiomatization of Euclidean Space . . . 929
VI.3.2. $\mathbb{R}^n$ is n-Dimensional . . . 943
VI.3.3. Exercises . . . 943
VI.4. Length Structures and Metric Geometry . . . 947

VI.4.1. Arc Length in Metric Spaces . . . 947
VI.4.2. Length Metrics . . . 955
VI.4.3. Length Structures . . . 960
VI.4.4. Metric Geometry . . . 964
VI.4.5. Exercises . . . 964

VII Curves **965**
VII.1 Curves and Parametrized Curves in $\mathbb{R}^n$ . . . 966
VII.1.1 Parametrized Curves . . . 966
VII.1.2 Reparametrization . . . 967
VII.1.3 Paths . . . 968
VII.2 Smooth and Piecewise Smooth Curves in $\mathbb{R}^n$ . . . 968
VII.2.1 Smooth Functions from $\mathbb{R}$ to $\mathbb{R}^n$ . . . 968
VII.2.2 Smooth Curves in $\mathbb{R}^n$ . . . 970
VII.2.3 Exercises . . . 971
VII.3 Rectifiable Curves and Arc Length . . . 972
VII.3.1 Arc Length . . . 972
VII.3.2 Rectifiable Curves . . . 975
VII.3.3 Length of Arcs . . . 982
VII.3.4 Integrating a Function Along a Curve . . . 983
VII.3.5 Exercises . . . 986
VII.4 Curves in $\mathbb{R}^2$ . . . 986
VII.5 Curves in $\mathbb{R}^3$ . . . 986
VII.5.1 Exercises . . . 986

VIII Multivariable and Vector Calculus **987**
VIII.1 Limits and Continuity for Vector Functions . . . 988
VIII.1.1 Graphs of Functions of Several Variables . . . 988
VIII.1.2 Limits of Sequences in $\mathbb{R}^n$ . . . 988
VIII.1.3 Multivariable Limits . . . 989
VIII.2 Partial and Directional Derivatives . . . 995
VIII.2.1 Partial Derivatives . . . 995
VIII.2.2 Equality of Mixed Partials . . . 999
VIII.2.3 Directional Derivatives . . . 1004
VIII.2.4 Exercises . . . 1005
VIII.3 Differentiability of Multivariable Functions . . . 1006
VIII.3.1 Differentiability and the Derivative . . . 1006
VIII.3.2 Differentiability of Algebraic Combinations . . . 1009
VIII.3.3 A Criterion for Differentiability . . . 1011
VIII.3.4 The Chain Rule . . . 1013
VIII.3.5 The Gradient and Tangent Planes . . . 1017
VIII.3.6 Functional Independence . . . 1019
VIII.3.7 Exercises . . . 1021
VIII.4 Higher-Order Derivatives and Critical Point Analysis . . . 1024
VIII.4.1 Higher-Order Differentiability . . . 1024
VIII.4.2 Smooth Partitions of Unity and the Smooth Urysohn Lemma . . . 1026

VIII.4.3 Taylor's Theorem in Several Variables . . . 1029
VIII.4.4 Critical Point Analysis . . . 1033
VIII.4.5 The Morse Lemma . . . 1043
VIII.4.6 Exercises . . . 1043
VIII.5 The Implicit and Inverse Function Theorems . . . 1045
VIII.5.1 Statements of the Theorems . . . 1045
VIII.5.2 Twin Theorems and a Generalization . . . 1058
VIII.5.3 Proof of the Implicit Function Theorem, Case $n = 1$ . . . 1061
VIII.5.4 Proof of the General Implicit Function Theorem . . . 1071
VIII.5.5 Proof of the Inverse Function Theorem . . . 1075
VIII.5.6 How Do We Find The Implicit/Inverse Function? . . . 1077
VIII.5.7 How Much Can the Theorems be Generalized? . . . 1077
VIII.5.8 The Domain and Range Straightening Theorems . . . 1079
VIII.5.9 The Degenerate Case . . . 1084
VIII.5.10 Analytic Versions . . . 1086
VIII.5.11 Exercises . . . 1090
VIII.6 Constrained Extrema and Lagrange Multipliers . . . 1095
VIII.6.1 The Two-Dimensional Case . . . 1095
VIII.6.2 The Higher-Dimensional Case with One Constraint . . . 1095
VIII.6.3 The Higher-Dimensional Case with Multiple Constraints . . . 1095
VIII.6.4 Degenerate Cases . . . 1095
VIII.6.5 Exercises . . . 1095
VIII.7 Surface Area and Surface Integrals . . . 1096
VIII.7.1 Surface Area . . . 1096
VIII.8 Power Series and Analytic Functions of Several Variables . . . 1098
VIII.8.1 Multivariable Power Series . . . 1098
VIII.8.2 Analytic Functions of Several Variables . . . 1104

IX Riemann Integration **1106**
IX.1. Introduction . . . 1106
IX.1.1. Origin of Integration Theories . . . 1106
IX.1.2. The Cauchy Integral for Continuous Functions . . . 1108
IX.1.3. The Riemann Integral . . . 1111
IX.1.4. The Lebesgue Approach . . . 1112
IX.1.5. The Daniell Approach . . . 1117
IX.1.6. Direct Generalizations of Riemann Integration . . . 1118
IX.1.7. Oriented and Unoriented Integration . . . 1119
IX.2. The Riemann Integral . . . 1121
IX.2.1. Partitions, Tags, Riemann and Darboux Sums . . . 1121
IX.2.2. Step Functions and their Integrals . . . 1129
IX.2.3. Definitions of the Riemann Integral . . . 1129
IX.2.4. Comparing the Definitions . . . 1133
IX.2.5. Integrals of Continuous and Monotone Functions . . . 1140
IX.2.6. Properties of the Riemann Integral . . . 1140
IX.2.7. Null Sets and Almost Everywhere . . . 1147
IX.2.8. The Upper and Lower Envelope of a Riemann-Integrable Function . . . 1151

IX.2.9. Differentiation and the Fundamental Theorem of Calculus . . . 1153
IX.2.10 Integration by Substitution and Integration by Parts . . . 1159
IX.2.11 Improper Riemann Integration . . . 1163
IX.2.12 Exercises . . . 1164
IX.2.13 Numerical Integration and Simpson's Rule . . . 1167
IX.2.14 Exercises . . . 1185
IX.3. Jordan Content and Riemann Integration in $\mathbb{R}^n$ . . . 1188
IX.3.1. Integration over Rectangles . . . 1188
IX.3.2. Content and Jordan Regions . . . 1198
IX.3.3. Integration over Jordan Regions . . . 1204
IX.3.4. Fubini's Theorem . . . 1207
IX.3.5. Interchanging Differentiation and Integration . . . 1217
IX.3.6. Change of Variable . . . 1221
IX.3.7. Improper Integrals . . . 1221
IX.3.8. Step Functions . . . 1222
IX.3.9. The Monotone Convergence Theorem . . . 1226
IX.3.10 Characterization of Riemann Integrable Functions . . . 1228
IX.3.11 The Extended Riemann Integral . . . 1231
IX.3.12 Exercises . . . 1239
IX.4. The Jacobian and Change of Variable . . . 1241
IX.4.1. Inner Products on Exterior Powers . . . 1241
IX.4.2. The Jacobian(s) of a Linear Transformation . . . 1243
IX.4.3. The Jacobian(s) of a Differentiable Function . . . 1246
IX.4.4. Change of Variable . . . 1247

X Differential Forms and Vector Analysis **1248**
X.1. Vector Fields . . . 1249
X.1.1. The Tangent and Cotangent Bundle on $\mathbb{R}^n$ . . . 1249
X.1.2. Sections and Vector Fields . . . 1249
X.1.3. Change of Basis . . . 1251
X.1.4. Lie Differentiation . . . 1251
X.1.5. Integral Curves . . . 1252
X.1.6. The Local Flow . . . 1253
X.1.7. Exercises . . . 1254
X.2. Differential Forms . . . 1256
X.2.1. Differential Forms on $\mathbb{R}^1$ . . . 1256
X.2.2. Differential 1-Forms and Line Integrals on $\mathbb{R}^n$ . . . 1259
X.2.3. Higher Differential Forms on $\mathbb{R}^n$ . . . 1259
X.2.4. Integration of Differential Forms on $\mathbb{R}^n$ . . . 1260
X.2.5. The Differential of a Form . . . 1261
X.2.6. Stokes' Theorem for Rectangles in $\mathbb{R}^n$ . . . 1264
X.2.7. Pullbacks of Forms . . . 1266
X.2.8. Chains and the General Stokes' Theorem in $\mathbb{R}^n$ . . . 1268
X.2.9. Surfaces in $\mathbb{R}^3$ . . . 1268
X.2.10. Divergence and Curl . . . 1268
X.2.11. Analysis on Manifolds . . . 1268

X.2.12. Exercises . . . 1268
X.3. Differential Topology . . . 1268
X.3.1. De Rham Cohomology . . . 1268
X.3.2. Vector Bundles . . . 1268
X.3.3. Topological K-Theory . . . 1268

XI Differential Equations **1269**
XI.1. Existence and Uniqueness of Solutions to Ordinary Differential Equations . . . 1271
XI.1.1. Elementary Considerations and Terminology . . . 1271
XI.1.2. Existence and Uniqueness Theorem for First-Order ODEs . . . 1275
XI.1.3. Systems of First-Order Differential Equations . . . 1279
XI.1.4. Higher-Order Differential Equations and Systems . . . 1281
XI.1.5. Continuous Dependence on Initial Conditions . . . 1282
XI.1.6. Peano's Existence Theorem for First-Order ODEs . . . 1284
XI.1.7. Power Series and Analytic Solutions . . . 1285
XI.1.8. Linear ODEs . . . 1285
XI.1.9. Boundary-Value Problems . . . 1285
XI.1.10 Exercises . . . 1286
XI.2. Partial Differential Equations . . . 1293
XI.2.1. Elementary Theory . . . 1293
XI.2.2. Quasilinear Partial Differential Equations . . . 1294
XI.2.3. The Cauchy Problem . . . 1294
XI.2.4. The Cauchy-Kovalevskaya Theorem . . . 1295

XII Fundamentals of Complex Analysis **1297**
XII.1 Elementary Theory . . . 1297
XII.1.1 Complex Differentiation and Holomorphic Functions . . . 1297
XII.1.2 Rules of Differentiation . . . 1302
XII.1.3 Branch Cuts and Branches . . . 1304
XII.1.4 The Fundamental Theorem of Algebra . . . 1307
XII.2 Contour Integration . . . 1308
XII.2.1 Integration along a Contour . . . 1308
XII.3 Some Big Theorems . . . 1311
XII.3.1 Goursat's Theorem . . . 1312
XII.3.2 The Antiderivative Theorem and Cauchy's Theorem for Convex Sets . . . 1314
XII.3.3 The Cauchy Integral Formula . . . 1317
XII.3.4 The Maximum Modulus Theorem . . . 1322
XII.3.5 The Homotopy Theorem for Contour Integrals . . . 1323
XII.3.6 Homology of Contours . . . 1327
XII.3.7 Cauchy's Theorem, General Versions . . . 1327
XII.3.8 The Cauchy Integral Formula, General Versions . . . 1330
XII.3.9 Exercises . . . 1330
XII.4 Sequences and Series of Holomorphic Functions; Analytic Functions . . . 1331
XII.4.1 Sequences and Series of Holomorphic Functions . . . 1331
XII.4.2 Power Series . . . 1335
XII.4.3 Inverse Power Series . . . 1339

XII.4.4 Analytic Functions . . . 1341
XII.4.5 Normal Families and Montel's Theorem . . . 1348
XII.5 Zeroes and Singularities . . . 1350
XII.5.1 Zeroes of Holomorphic Functions . . . 1350
XII.5.2 Singularities . . . 1352
XII.5.3 Behavior at Infinity . . . 1358
XII.5.4 Meromorphic Functions . . . 1360
XII.5.5 Exercises . . . 1361
XII.6 Mapping Properties of Analytic Functions . . . 1361
XII.6.1 The Complex Morse Lemma . . . 1361
XII.6.2 Local Mapping of an Analytic Function . . . 1362
XII.7 Residues and the Principle of the Argument . . . 1363
XII.7.1 The Residue Theorem . . . 1363
XII.7.2 Principle of the Argument . . . 1365
XII.7.3 Rouché's Theorem . . . 1367
XII.7.4 Hurwitz's Theorem . . . 1368
XII.7.5 Exercises . . . 1369
XII.8 Conformal Equivalence and the Riemann Mapping Theorem . . . 1371
XII.8.1 The Schwarz Lemma . . . 1371
XII.8.2 Conformal Equivalence . . . 1371
XII.8.3 Simply Connected Sets in $\mathbb{C}$ . . . 1372
XII.8.4 The Riemann Mapping Theorem . . . 1375
XII.8.5 Extension to the Boundary . . . 1377
XII.8.6 Exercises . . . 1380
XII.9 Analytic Continuation and Riemann Surfaces . . . 1384
XII.9.1 Analytic Function Elements . . . 1384
XII.9.2 The Riemann Surface of a Total Analytic Function . . . 1385
XII.9.3 Analytic Continuation . . . 1389
XII.9.4 The Regular Riemann Surface of a Function . . . 1392
XII.9.5 Branch Points . . . 1393
XII.9.6 Points at Infinity and Meromorphic Functions . . . 1394
XII.9.7 Germs of Analytic Functions . . . 1396
XII.9.8 General Riemann Surfaces . . . 1397
XII.9.9 Exercises . . . 1398
XII.10 Analytic Functions of Several Variables . . . 1403
XII.10.1 Elementary Theory . . . 1403
XII.10.2 Exercises . . . 1406
XII.11 Harmonic Functions . . . 1407
XII.11.1 Definitions and Basic Properties . . . 1407
XII.11.2 Harmonic and Analytic Functions . . . 1408
XII.11.3 The Mean Value Property and Local Extrema . . . 1409
XII.11.4 The Dirichlet Problem . . . 1411
XII.11.5 Exercises . . . 1412
XII.12 Differential Algebra . . . 1414
XII.12.1 Derivations . . . 1414
XII.12.2 Differential Fields . . . 1417

XII.12.3 Integration in Finite Terms . . . 1422
XII.12.4 Exercises . . . 1427

XIII Topology **1429**
XIII.1 Topological Spaces . . . 1432
XIII.1.1 Definitions . . . 1432
XIII.1.2 Bases and Subbases . . . 1433
XIII.1.3 Covers and Subcovers . . . 1437
XIII.1.4 Relative Topology on a Subspace . . . 1438
XIII.1.5 Separated Union . . . 1440
XIII.1.6 Exercises . . . 1441
XIII.2 Closed Sets and Limit Points . . . 1443
XIII.2.1 Closed Sets . . . 1443
XIII.2.2 Limit Points of a Set . . . 1444
XIII.2.3 Closure of a Set . . . 1444
XIII.2.4 Specification of a Topology via Closure . . . 1448
XIII.2.5 Isolated Points and Derived Sets . . . 1450
XIII.2.6 Exercises . . . 1452
XIII.3 Convergence . . . 1453
XIII.3.1 Limits of Sequences . . . 1453
XIII.3.2 Nets and Limits . . . 1455
XIII.3.3 Specification of a Topology via Convergence . . . 1461
XIII.3.4 Convergence via Filters . . . 1465
XIII.3.5 Exercises . . . 1469
XIII.4 Continuous Functions . . . 1471
XIII.4.1 Definition and Properties . . . 1471
XIII.4.2 Topological Continuity vs. Metric Continuity . . . 1471
XIII.5 Comparison of Topologies . . . 1472
XIII.5.1 Stronger and Weaker Topologies . . . 1472
XIII.6 Products and the Product Topology . . . 1476
XIII.6.1 Finite Products . . . 1476
XIII.6.2 General Products . . . 1478
XIII.6.3 Exercises . . . 1480
XIII.7 Separation Axioms . . . 1482
XIII.7.1 The T-Axioms . . . 1482
XIII.7.2 T_0-Spaces . . . 1483
XIII.7.3 T_1-Spaces . . . 1484
XIII.7.4 Hausdorff Spaces . . . 1485
XIII.7.5 Regular Spaces . . . 1485
XIII.7.6 Normal Spaces . . . 1486
XIII.7.7 Fully Normal Spaces . . . 1489
XIII.7.8 Completely Regular Spaces . . . 1491
XIII.7.9 The Tietze Extension Theorem . . . 1492
XIII.7.10 Exercises . . . 1493
XIII.8 Quotient Spaces and the Quotient Topology . . . 1502
XIII.8.1 The Quotient Topology . . . 1502

XIII.8.2 Semicontinuous Decompositions . . . 1504
XIII.8.3 Separation Axioms for Quotient Spaces . . . 1506
XIII.8.4 Quotient Metrics . . . 1509
XIII.8.5 Exercises . . . 1511
XIII.9 Identification Spaces . . . 1515
XIII.9.1 Collapsing a Subset to a Point . . . 1515
XIII.9.2 Gluing and Attaching . . . 1517
XIII.9.3 Group Actions on Spaces . . . 1527
XIII.9.4 Local Homeomorphisms and Covering Maps . . . 1530
XIII.9.5 Exercises . . . 1530
XIII.10 Metrizable Spaces . . . 1536
XIII.10.1 Metrizability . . . 1536
XIII.10.2 Continuity and Oscillation . . . 1541
XIII.10.3 Completely Metrizable Spaces . . . 1542
XIII.10.4 Universal Space . . . 1544
XIII.10.5 Polish Spaces . . . 1548
XIII.10.6 Exercises . . . 1549
XIII.11 Compactness . . . 1554
XIII.11.1 Basic Definitions . . . 1554
XIII.11.2 Compactness and Convergence . . . 1556
XIII.11.3 Compact Metrizable Spaces . . . 1558
XIII.11.4 Separation Properties . . . 1561
XIII.11.5 Automatic Continuity and Comparison of Topologies . . . 1562
XIII.11.6 Products of Compact Spaces and Tikhonov's Theorem . . . 1562
XIII.11.7 Variations on Compactness . . . 1566
XIII.11.8 Locally Compact Spaces . . . 1568
XIII.11.9 Compactifications . . . 1574
XIII.11.10 Paracompactness and Partitions of Unity . . . 1578
XIII.11.11 Compactly Generated Spaces . . . 1583
XIII.11.12 Proper Maps . . . 1585
XIII.11.13 Local Compactness . . . 1586
XIII.11.14 Exercises . . . 1587
XIII.11.15 Dini's Theorem . . . 1595
XIII.12 Baire Spaces and the Baire Category Theorem . . . 1596
XIII.12.1 Meager Sets and Baire Spaces . . . 1596
XIII.12.2 The Baire Category Theorem . . . 1597
XIII.12.3 Meager Sets vs. Null Sets . . . 1599
XIII.12.4 The Baire Property . . . 1600
XIII.12.5 Exercises . . . 1602
XIII.13 Connectedness . . . 1605
XIII.13.1 Connected Spaces . . . 1605
XIII.13.2 Continuous Images of Connected Sets . . . 1607
XIII.13.3 Unions and Intersections . . . 1608
XIII.13.4 Components of a Space . . . 1609
XIII.13.5 Products of Connected Spaces . . . 1610
XIII.13.6 Totally Disconnected and Zero-Dimensional Spaces . . . 1611

XIII.13 Locally Connected Spaces . . . 1614
XIII.13 Path-Connected Spaces . . . 1614
XIII.13 Exercises . . . 1618
XIII.14 Direct and Inverse Limits . . . 1622
XIII.14 Direct Systems and Direct Limits of Sets . . . 1622
XIII.14 Inverse Systems and Inverse Limits of Sets . . . 1626
XIII.14 Direct Limits in Topology . . . 1630
XIII.14 Inverse Limits in Topology . . . 1631
XIII.14 Exercises . . . 1637
XIII.15 Uniform Spaces . . . 1643
XIII.15 Topologies and Uniform Structures Defined by Pseudometrics . . . 1643
XIII.15 General Uniform Structures . . . 1650
XIII.15 The Covering Approach . . . 1656
XIII.15 Comparing the Definitions . . . 1658
XIII.15 Pullbacks and Fine Uniformities . . . 1660
XIII.15 Completeness . . . 1661
XIII.15 Subspace Uniformities and Completeness . . . 1664
XIII.15 Completions of Uniform Spaces . . . 1666
XIII.15 Metric Spaces . . . 1668
XIII.15 Compact and Precompact Uniform Spaces . . . 1671
XIII.15 Topologically Complete Spaces . . . 1674
XIII.15 Exercises . . . 1674
XIII.16 Order Topology . . . 1678
XIII.16 Definition and Properties . . . 1678
XIII.16 Completion of an Order Topology . . . 1678
XIII.17 Function Spaces . . . 1679
XIII.17 Function Space Notation . . . 1679
XIII.17 U.C. Convergence and the Compact-Open Topology . . . 1680
XIII.17 The Arzelà-Ascoli Theorem . . . 1684
XIII.17 Exercises . . . 1691
XIII.18 Cantor Sets . . . 1692
XIII.18 The Cantor Set . . . 1692
XIII.18 Other Cantor Sets . . . 1692
XIII.18 Exercises . . . 1695
XIII.18 Space-Filling Curves . . . 1696
XIII.18 Exercises . . . 1697
XIII.18 A Plane Curve With Positive Area . . . 1698
XIII.19 The Topology of $\mathbb{R}^n$ . . . 1703
XIII.19 Invariance of Domain and the Jordan Curve Theorem . . . 1703
XIII.19 Orientation . . . 1707
XIII.19 $\mathbb{R}^n$ is n-Dimensional . . . 1708
XIII.19 When is $\mathbb{R}^n$ "Similar" to $\mathbb{R}^m$? . . . 1721
XIII.19 Compact Subsets of $\mathbb{R}^n$ . . . 1724
XIII.19 Exercises . . . 1753
XIII.20 Manifolds . . . 1764
XIII.20 Topological Manifolds . . . 1764

XIII.20.2 Topological Manifolds With Boundary . . . 1767
XIII.20.3 New Manifolds From Old . . . 1769
XIII.20.4 Classification of Topological Manifolds . . . 1769
XIII.20.5 Exercises . . . 1770
XIII.21 The Brouwer Fixed-Point Theorem . . . 1777
XIII.21.1 Statements of the Results . . . 1777
XIII.21.2 The Case $n = 2$ . . . 1783
XIII.21.3 The General Case . . . 1785
XIII.21.4 Exercises . . . 1786
XIII.21.5 Proof of the No-Retraction Theorem . . . 1788
XIII.21.6 The Fixed-Point Theorem for Cubes . . . 1791
XIII.22 Topological Dimension Theory . . . 1792
XIII.22.1 Essential Dimension . . . 1792
XIII.22.2 Invariance of Domain . . . 1795
XIII.22.3 Other Dimension Functions . . . 1796
XIII.22.4 Comparing the Dimensions . . . 1809
XIII.22.5 More Dimension Theories . . . 1813
XIII.22.6 Embedding in Euclidean Space . . . 1816
XIII.22.7 Infinite-Dimensional Spaces . . . 1821
XIII.22.8 Exercises . . . 1822
XIII.22.9 Lebesgue's Differentiation Theorem . . . 1827
XIII.22.10 Exercises . . . 1829
XIII.23 Sard's Theorem . . . 1831
XIII.23.1 Sets of Measure Zero . . . 1832
XIII.23.2 Statement of Sard's Theorem . . . 1836
XIII.23.3 The Case $n < m$ . . . 1837
XIII.23.4 The Case $n = m = 1$ . . . 1837
XIII.23.5 The Case $n = m > 1$ . . . 1838
XIII.23.6 The Case $n > m$, r Large . . . 1838
XIII.23.7 The Case $n > m$, Sharp r Bound . . . 1840
XIII.23.8 The Category Version . . . 1840
XIII.23.9 Counterexamples . . . 1841
XIII.23.10 Applications . . . 1842
XIII.23.11 Refined Versions . . . 1843
XIII.23.12 Exercises . . . 1844
XIII.24 Miscellaneous Aspects of Topology . . . 1846
XIII.24.1 The Countable Chain Condition . . . 1846
XIII.25 Piecewise-Linear Topology . . . 1849
XIII.25.1 Polyhedra, Cells, and Simplexes . . . 1849
XIII.25.2 Complexes . . . 1856
XIII.25.3 Triangulations . . . 1859
XIII.25.4 Piecewise-Linear Maps . . . 1861
XIII.25.5 Exercises . . . 1863
XIII.26 Absolute Retracts and Absolute Neighborhood Retracts . . . 1864
XIII.26.1 Retractions . . . 1864
XIII.26.2 Absolute Retracts . . . 1865

XIII.26.3 Absolute Neighborhood Retracts . . . 1871
XIII.26.4 The Noncompact Case . . . 1877
XIII.26.5 Affine Structure in ARs and ANRs . . . 1881
XIII.26.6 Exercises . . . 1882

XIV Differential Geometry and Topology **1885**
XIV.1 Smooth Manifolds . . . 1885
XIV.1.1 Concrete Manifolds . . . 1886
XIV.1.2 Smooth Maps and Diffeomorphisms . . . 1888
XIV.1.3 Tangent Space and Tangent Bundle . . . 1890
XIV.1.4 Derivatives . . . 1891
XIV.1.5 Immersions, Submersions, and Embeddings . . . 1891
XIV.1.6 Transversality . . . 1891
XIV.2 Abstract Smooth Manifolds . . . 1891
XIV.2.1 Charts and Atlases . . . 1891

XV Algebraic Topology **1892**
XV.1 Homotopy and the Fundamental Group . . . 1894
XV.1.1 Homotopy . . . 1894
XV.1.2 The Fundamental Group . . . 1899
XV.1.3 Covering Spaces . . . 1899
XV.1.4 Other Homotopy Groups . . . 1899
XV.1.5 Exercises . . . 1900
XV.2 Homology . . . 1901
XV.2.1 Singular Homology . . . 1901
XV.2.2 Simplicial Homology . . . 1901
XV.3 Homological Algebra . . . 1902
XV.3.1 Preliminaries . . . 1902
XV.3.2 Chain Complexes . . . 1909
XV.3.3 Exercises . . . 1909
XV.4 Cohomology . . . 1912
XV.4.1 General Cohomology . . . 1912
XV.4.2 Cup and Cap Products . . . 1912
XV.4.3 Čech Cohomology . . . 1912
XV.4.4 De Rham Cohomology . . . 1912
XV.4.5 Poincaré Duality . . . 1912
XV.4.6 K-Theory . . . 1912

XVI Measurable Spaces and Subset Structures **1913**
XVI.1 Algebras and σ-Algebras of Sets . . . 1913
XVI.1.1 Set Operations . . . 1914
XVI.1.2 Rings, Semirings, and Semialgebras . . . 1916
XVI.1.3 Algebras and σ-Algebras . . . 1919
XVI.1.4 σ-Rings and δ-Rings . . . 1922
XVI.1.5 π-Systems, λ-Systems, and Monotone Classes . . . 1925
XVI.1.6 Measurable Spaces and Measurable Functions . . . 1927

XVI.1.7 Exercises . . . 1930
XVI.1.8 Filters and Ultrafilters . . . 1932
XVI.1.9 Exercises . . . 1936
XVI.2 Borel Sets and Borel Measurable Functions . . . 1938
XVI.2.1 Borel Sets . . . 1938
XVI.2.2 Real-Valued Measurable Functions . . . 1940
XVI.2.3 Real-Valued Measurable Functions on δ-Rings . . . 1943
XVI.2.4 Borel Measurable Functions . . . 1944
XVI.2.5 Another Characterization of Borel Sets and Borel Measurable Functions . . . 1947
XVI.2.6 Algebraic Properties of $\bar{\mathfrak{C}}$ . . . 1948
XVI.2.7 The Borel Sets Are Not Countably Constructible . . . 1952
XVI.2.8 The Multiplicity Function . . . 1955
XVI.2.9 Standard Borel Spaces . . . 1957
XVI.2.10 Exercises . . . 1959
XVI.3 Analytic Sets . . . 1962
XVI.3.1 Standard Borel Spaces . . . 1962
XVI.3.2 Analytic Sets . . . 1963
XVI.3.3 Suslin Sets and the Suslin Operation . . . 1966
XVI.3.4 Addison's Notation and the Projective Hierarchy . . . 1968
XVI.3.5 Exercises . . . 1969

XVII Measure Theory **1970**
XVII.1 Measures . . . 1970
XVII.1.1 Definitions . . . 1970
XVII.1.2 Elementary Properties . . . 1972
XVII.1.3 Null Sets and Completeness . . . 1975
XVII.1.4 Finite, σ-Finite, and Semifinite Measures . . . 1978
XVII.1.5 Discrete and Continuous Measures . . . 1981
XVII.1.6 Linear Combinations, Restrictions, and Extensions of Measures . . . 1981
XVII.1.7 Measures Defined on Rings and σ-Rings . . . 1985
XVII.1.8 Egorov's Theorem . . . 1985
XVII.1.9 Measures on δ-Rings . . . 1986
XVII.1.10 Exercises . . . 1986
XVII.2 The Metric Space of a Measure Space . . . 1989
XVII.3 Probability . . . 1991
XVII.3.1 Introduction . . . 1991
XVII.3.2 The Dictionary . . . 1993
XVII.3.3 Exercises . . . 1993
XVII.4 Measures Defined by Extensions . . . 1995
XVII.4.1 Outer Measures . . . 1996
XVII.4.2 Lebesgue Outer Measure on $\mathbb{R}$ . . . 1998
XVII.4.3 The Measure Induced by an Outer Measure . . . 2000
XVII.4.4 Lebesgue Measure on $\mathbb{R}$ . . . 2003
XVII.4.5 Extending Premeasures from Semirings . . . 2004
XVII.4.6 The Suslin Operation . . . 2018
XVII.4.7 Local Measures on σ-Rings and δ-Rings . . . 2022

XVII.4.2 Radon Measures on $\mathbb{R}$ . . . 2024
XVII.4.3 Lebesgue Measure and Radon Measures on $\mathbb{R}^n$ . . . 2024
XVII.4.4 Nonmeasurable Subsets of $\mathbb{R}$ . . . 2024
XVII.5 Measurable Functions . . . 2027
XVII.5.1 Measurable Functions on a Measure Space . . . 2027
XVII.5.2 Lebesgue Measurable Functions . . . 2028
XVII.5.3 Exercises . . . 2031
XVII.6 Measures on Metric Spaces . . . 2034
XVII.6.1 Metric Outer Measures and Borel Measures . . . 2034
XVII.6.2 Method II Construction of Outer Measures . . . 2042
XVII.6.3 Hausdorff Measure . . . 2044
XVII.6.4 Hausdorff Measure on $\mathbb{R}^n$ . . . 2049
XVII.6.5 Hausdorff Measure and Hausdorff Dimension for Self-Similar Sets . . . 2050
XVII.6.6 Packing Measure and other Geometric Measures . . . 2050
XVII.6.7 Metric Dimensions . . . 2055
XVII.6.8 Radon Measures on Metric Spaces . . . 2060
XVII.6.9 Exercises . . . 2060
XVII.7 Measures on Product Spaces . . . 2062
XVII.7.1 Product σ-Algebras . . . 2062
XVII.7.2 Product Measures . . . 2063
XVII.7.3 Lebesgue Measure on $\mathbb{R}^n$ . . . 2066
XVII.7.4 Infinite Product Spaces . . . 2066
XVII.7.5 Exercises . . . 2068
XVII.8 The Besicovitch Covering Theorem . . . 2070
XVII.8.1 Controlled Metrics . . . 2070
XVII.8.2 The Besicovitch Partition Theorem . . . 2074
XVII.8.3 The Besicovitch Covering Theorem . . . 2077
XVII.8.4 Exercises . . . 2080
XVII.9 Signed and Complex Measures . . . 2083
XVII.9.1 Signed Measures . . . 2083
XVII.9.2 Complex Measures . . . 2088
XVII.9.3 Integration With Respect to Signed and Complex Measures . . . 2090
XVII.9.4 Another Approach via Local Measures . . . 2092
XVII.9.5 Exercises . . . 2094
XVII.9.6 Signed and Complex Measures on σ-Rings . . . 2097
XVII.10 Limits of Measures . . . 2098
XVII.10.1 General Considerations . . . 2098
XVII.10.2 Monotone and Dominated Convergence . . . 2099
XVII.10.3 The Nikodým and Vitali-Hahn-Saks Theorems . . . 2101
XVII.10.4 Exercises . . . 2104

XVIII Modern Integration **2106**
XVIII.1 The Lebesgue Integral . . . 2107
XVIII.1.1 Properties of the Integral . . . 2107
XVIII.1.2 Integral of Simple Functions . . . 2108
XVIII.1.3 The Integral of a Nonnegative Measurable Function . . . 2113

XVIII.4 The Monotone Convergence Theorem . . . 2117
XVIII.5 Consequences of the Monotone Convergence Theorem . . . 2118
XVIII.6 The Lebesgue Integral and Riemann Integral . . . 2125
XVIII.7 Continuity and Fatou's Lemma . . . 2130
XVIII.8 The Integral For General Functions . . . 2134
XVIII.9 Exercises . . . 2137
XVIII.10 The Dominated Convergence Theorem . . . 2139
XVIII.10.1 Applications of the DCT . . . 2139
XVIII.10.2 Uniform Integrability and Vitali's Convergence Theorem . . . 2141
XVIII.10.3 Exercises . . . 2149
XVIII.11 Integration on Product Spaces . . . 2150
XVIII.11.1 The Fubini-Tonelli Theorems . . . 2150
XVIII.11.2 Lebesgue Integration on $\mathbb{R}^n$ . . . 2159
XVIII.11.3 Exercises . . . 2159
XVIII.12 The Henstock-Kurzweil Integral . . . 2161
XVIII.12.1 Definition and Basic Properties . . . 2161
XVIII.12.2 The Fundamental Theorem of Calculus . . . 2162
XVIII.13 Convolution . . . 2164
XVIII.13.1 Convolution of Functions on $\mathbb{R}$, Continuous Case . . . 2164
XVIII.14 The Radon-Nikodým Theorem . . . 2173
XVIII.14.1 Density Functions . . . 2173
XVIII.14.2 Statement and Proof of the Theorem . . . 2176
XVIII.14.3 The Lebesgue Decomposition Theorem . . . 2178
XVIII.14.4 Signed and Complex Measures . . . 2178
XVIII.14.5 Conditional Expectations . . . 2178
XVIII.14.6 Change of Variable . . . 2184
XVIII.14.7 Exercises . . . 2187

XIX Advanced Analysis on Euclidean Spaces **2188**
XIX.1 Borel Measures on $\mathbb{R}$ and Cumulative Distribution Functions . . . 2189
XIX.1.1 Cumulative Distribution Functions . . . 2189
XIX.1.2 Density Functions and Absolute Continuity . . . 2190
XIX.1.3 Finite Borel Measures on Intervals . . . 2192
XIX.1.4 Stieltjes Integrals . . . 2193
XIX.1.5 Unbounded Distribution Functions and Positive Radon Measures . . . 2193
XIX.1.6 Signed Measures and Nonincreasing Distribution Functions . . . 2194
XIX.1.7 Borel Measures and Distribution Functions on $\mathbb{R}^n$ . . . 2194
XIX.1.8 Integration By Parts . . . 2194
XIX.2 Monotone Functions . . . 2196
XIX.2.1 Definitions and Basics . . . 2196
XIX.2.2 Differentiability of Monotone Functions . . . 2198
XIX.2.3 Fubini's Series Theorem . . . 2199
XIX.2.4 Exercises . . . 2203
XIX.3 Convex Functions . . . 2204
XIX.3.1 Definitions . . . 2204
XIX.3.2 Midpoint-Convex Functions . . . 2212

XIX.3.3 Convergence of Sequences of Convex Functions . . . 2216
XIX.3.4 Jensen's Inequality . . . 2221
XIX.3.5 Log-Convex Functions . . . 2223
XIX.3.6 Higher-Dimensional Versions of Convexity . . . 2225
XIX.3.7 Exercises . . . 2225
XIX.4 Functions of Bounded Variation . . . 2228
XIX.4.1 The Total Variation of a Function . . . 2228
XIX.4.2 The Total Variation Functions . . . 2233
XIX.4.3 Banach's Multiplicity Theorem . . . 2235
XIX.4.4 Extensions of Bounded Variation . . . 2236
XIX.4.5 Exercises . . . 2238
XIX.5 Absolute Continuity and Lebesgue Density . . . 2240
XIX.5.1 Luzin's Property (N) . . . 2240
XIX.5.2 Absolutely Continuous Functions . . . 2245
XIX.5.3 Signed and Complex Borel Measures on $\mathbb{R}$ . . . 2259
XIX.5.4 The Bounded Variation Norm . . . 2262
XIX.5.5 Integration by Substitution and Integration by Parts . . . 2266
XIX.5.6 L^p Versions, $p > 1$ . . . 2271
XIX.5.7 The Lebesgue Density Theorem . . . 2273
XIX.5.8 Exercises . . . 2275
XIX.6 Approximate Limits and Continuity . . . 2278
XIX.6.1 The Radon-Nikodým Theorem in $\mathbb{R}^n$ . . . 2278
XIX.6.2 Lebesgue Density in $\mathbb{R}^n$ . . . 2282
XIX.6.3 Approximate Limits . . . 2283
XIX.6.4 Approximate Continuity . . . 2285
XIX.6.5 Approximate Differentiability . . . 2286
XIX.7 Lipschitz Functions . . . 2288
XIX.7.1 Lipschitz Functions Between Metric Spaces . . . 2288
XIX.7.2 Locally Lipschitz Functions . . . 2288
XIX.7.3 Lipschitz Functions Between Euclidean Spaces . . . 2289
XIX.7.4 Hölder Continuity . . . 2291

XX Functional Analysis **2292**
XX.1 Preliminaries . . . 2293
XX.1.1 Properties of Subsets . . . 2293
XX.1.2 Convex Sets . . . 2294
XX.1.3 Faces and Extreme Points . . . 2295
XX.1.4 Minkowski Functionals . . . 2298
XX.2 Normed Vector Spaces . . . 2299
XX.2.1 Completeness . . . 2299
XX.2.2 Isometries of Normed Vector Spaces . . . 2299
XX.2.3 Exercises . . . 2301
XX.3 Banach Spaces . . . 2302
XX.4 Topological Vector Spaces . . . 2303
XX.4.1 Basic Definitions and Properties . . . 2303
XX.4.2 Locally Convex Topological Vector Spaces . . . 2307

XX.4.3 Paranorms and Fréchet Spaces . . . 2308
XX.4.4 Metrizability and Normability . . . 2314
XX.4.5 Continuous Linear Functionals and the Dual Space . . . 2317
XX.4.6 Duality and the Weak and Weak-* Topologies . . . 2320
XX.4.7 Uniform Topologies on X^* and the Bidual . . . 2322
XX.4.8 Exercises . . . 2325
XX.5 Bounded Operators . . . 2327
XX.5.1 Bounded Operators on Normed Vector Spaces . . . 2327
XX.5.2 Bounded Linear Functionals . . . 2330
XX.5.3 Operators on Finite-Dimensional Normed Spaces . . . 2331
XX.5.4 Bounded Operators on Hilbert Spaces . . . 2331
XX.5.5 Algebras of Operators . . . 2331
XX.6 The Hahn-Banach Theorem . . . 2333
XX.6.1 Sublinear Functionals and Version 1 . . . 2333
XX.6.2 Extension of Bounded Linear Functionals . . . 2337
XX.6.3 The Complex Hahn-Banach Theorem . . . 2338
XX.6.4 Separation of Convex Sets and Version 3 . . . 2340
XX.6.5 Exercises . . . 2343
XX.7 Dual Spaces and Weak Topologies . . . 2344
XX.7.1 Bounded Linear Functionals . . . 2344
XX.7.2 The Weak Topology . . . 2346
XX.7.3 The Second Dual . . . 2350
XX.7.4 The Weak-* Topology . . . 2351
XX.7.5 Exercises . . . 2355
XX.8 The Open Mapping Theorem, Closed Graph Theorem, and Uniform Boundedness Theorem . . . 2358
XX.8.1 Open Mappings . . . 2358
XX.8.2 The Open Mapping Theorem . . . 2359
XX.8.3 The Closed Graph Theorem . . . 2362
XX.8.4 The Uniform Boundedness Theorem . . . 2363
XX.8.5 Exercises . . . 2366
XX.9 Compact Convex Sets . . . 2367
XX.9.1 Convex Sets in Topological Vector Spaces . . . 2367
XX.9.2 Faces and Extreme Points . . . 2367
XX.9.3 Compact Convex Sets in Finite-Dimensional Spaces . . . 2369
XX.9.4 The Krein-Milman Theorem . . . 2371
XX.9.5 Choquet's Theorem . . . 2372
XX.9.6 Regular Embeddings, Cones, and Affine Functions . . . 2379
XX.9.7 Choquet Simplexes . . . 2379
XX.9.8 Exercises . . . 2380
XX.10 The Stone-Weierstrass Theorem . . . 2384
XX.10.1 The Weierstrass Approximation Theorem . . . 2384
XX.10.2 The Stone-Weierstrass Theorem . . . 2386
XX.10.3 Exercises . . . 2390
XX.11 Hilbert Spaces . . . 2395
XX.11.1 Inner Products . . . 2395
XX.11.2 The CBS Inequality and the Norm . . . 2399

XX.11.3 Completeness and Hilbert Spaces . . . 2404
XX.11.4 Best Approximation . . . 2407
XX.11.5 Orthogonality . . . 2408
XX.11.6 Projections . . . 2419
XX.11.7 Dual Spaces and Weak Topology . . . 2421
XX.11.8 Direct Sums and Amplifications . . . 2422
XX.11.9 Tensor Products of Inner Product Spaces . . . 2422
XX.11.10 Real Hilbert Spaces . . . 2423
XX.11.11 Exercises . . . 2424
XX.11.12 Sesquilinear Forms and Operators . . . 2429
XX.12 Banach Algebras and Spectral Theory . . . 2432
XX.12.1 Banach Algebras . . . 2432
XX.12.2 Invertibility, Ideals, and Quotients . . . 2434
XX.12.3 Spectrum . . . 2437
XX.12.4 Holomorphic Functional Calculus . . . 2446
XX.12.5 Spectral Theory for Bounded Operators . . . 2449
XX.12.6 Exercises . . . 2454
XX.13 Fredholm Operators and Fredholm Index . . . 2457
XX.13.1 Banach Space Preliminaries . . . 2457
XX.13.2 Compact Operators . . . 2458
XX.13.3 Fredholm Operators . . . 2460
XX.13.4 Fredholm Operators of Index 0 and the Fredholm Alternative . . . 2463
XX.13.5 Relation between Compact and Fredholm Operators . . . 2463
XX.13.6 Exercises . . . 2467
XX.14 The Spectral Theorem . . . 2468
XX.14.1 The Finite-Dimensional Spectral Theorem . . . 2468
XX.15 L^p-Spaces . . . 2470
XX.15.1 Conjugate Exponents . . . 2470
XX.15.2 $\mathcal{L}^p$-Spaces, Norms, and Inequalities . . . 2470
XX.15.3 L^p-Spaces . . . 2474
XX.15.4 $\mathcal{L}^\infty$- and L^∞-Spaces . . . 2476
XX.15.5 Comparison Between $\mathcal{L}^p$-Spaces and L^p-Spaces . . . 2476
XX.15.6 L^p-spaces for Borel Measures . . . 2481
XX.15.7 The Dual of L^p-Spaces . . . 2481
XX.15.8 Conditional Expectations on L^p-Spaces . . . 2488
XX.15.9 Exercises . . . 2490
XX.16 C*-Algebras . . . 2495
XX.16.1 Banach *-Algebras and C*-Algebras . . . 2495
XX.16.2 Commutative C*-Algebras and Continuous Functional Calculus . . . 2499
XX.16.3 Algebraic and Order Theory of C*-Algebras . . . 2507
XX.16.4 Representations of Banach *-Algebras and C*-Algebras . . . 2512
XX.16.5 Universal and Enveloping C*-Algebras . . . 2524
XX.16.6 Exercises . . . 2524

XXI Nonlinear Functional Analysis **2525**
XXI.1 Calculus on Banach Spaces . . . 2525
XXI.1.1 Differentiable Functions Between Banach Spaces . . . 2525
XXI.1.2 The Implicit and Inverse Function Theorems . . . 2528
XXI.1.3 Integration in Banach Spaces . . . 2531
XXI.2 Calculus of Variations . . . 2532
XXI.2.1 The Isoperimetric Problem . . . 2532

XXII Measure and Topology; Generalized Functions and Distributions **2534**
XXII.1 Measures on Locally Compact Spaces . . . 2534
XXII.1.1 Regular Measures . . . 2534
XXII.1.2 The Riesz Representation Theorem . . . 2534
XXII.2 Distributions . . . 2534
XXII.2.1 Generalized Functions on $\mathbb{R}$ . . . 2534
XXII.2.2 Order of Generalized Functions and a Sequential Approach . . . 2552
XXII.2.3 Distributions on $\mathbb{R}^n$ . . . 2561
XXII.2.4 Tempered Distributions . . . 2561
XXII.3 Sobolev Spaces . . . 2562
XXII.3.1 Definition of Sobolev Spaces . . . 2562
XXII.3.2 Sobolev Inequalities and Embedding Theorems . . . 2563
XXII.4 Differential Equations and Distributions . . . 2563
XXII.4.1 Ordinary Differential Equations and Generalized Functions . . . 2563
XXII.4.2 Partial Differential Equations and Generalized Functions . . . 2563
XXII.4.3 Exercises . . . 2563

XXIII Topological Groups and Abstract Harmonic Analysis **2565**
XXIII.1 Topological Groups . . . 2565
XXIII.1.1 Groups and Homomorphisms . . . 2565
XXIII.1.2 Topological Groups . . . 2569
XXIII.1.3 Uniformities on Topological Groups . . . 2576
XXIII.1.4 Locally Compact Topological Groups . . . 2578
XXIII.1.5 Polish Groups and Baire Groups . . . 2582
XXIII.1.6 Open Mappings and Automatic Continuity . . . 2583
XXIII.1.7 Exercises . . . 2584
XXIII.2 Haar Measure . . . 2587
XXIII.2.1 Definition and Existence . . . 2587
XXIII.2.2 Haar Integrals . . . 2588
XXIII.2.3 Invariant Measures for Group Actions . . . 2590
XXIII.2.4 Properties of Haar Measure . . . 2590
XXIII.2.5 The Modular Function . . . 2590
XXIII.2.6 Measures on Homogeneous Spaces . . . 2590
XXIII.2.7 Extending Haar Measure . . . 2590
XXIII.2.8 Measurable Groups . . . 2590
XXIII.2.9 Exercises . . . 2590
XXIII.3 Group Algebras and Unitary Representations . . . 2592
XXIII.3.1 The Group Algebra . . . 2592

XXIII.3.2 The L^1 Algebra and the C*-Algebra of a Locally Compact Group . . . 2595
XXIII.4 Lie Groups . . . 2597
XXIII.4.1 The Classical Groups . . . 2597
XXIII.4.2 Lie Groups and Lie Algebras . . . 2597
XXIII.4.3 Locally Euclidean Groups . . . 2597
XXIII.5 Locally Compact Abelian Groups and Pontryagin Duality . . . 2599
XXIII.5.1 The Dual Group . . . 2599
XXIII.5.2 Exercises . . . 2602

XXIV Addenda **2604**
XXIV.1 Mathematical Names . . . 2605
XXIV.1.1 The Greek Alphabet . . . 2614
XXIV.2 Origin of Terms and Notation . . . 2615
XXIV.3 Miscellany . . . 2620
XXIV.3.1 Age and Mathematics . . . 2620
XXIV.3.2 The Mathematical Organ . . . 2622
XXIV.3.3 Mathematics and Mental Illness . . . 2624
XXIV.4 Literature Review . . . 2627

Bibliography **2673**

Index **2710**

Chapter I

Introduction

> "I regard as quite useless the reading of large treatises of pure analysis: too large a number of methods pass at once before the eyes."
>
> *Joseph Louis Lagrange*[1]

This chapter is intended as an informal overview of the subject developed in this book. We will describe in general terms what the theory of real analysis is all about, and how it relates to the rest of mathematics.

I.1. Calculus and Analysis

> "By 1800 mathematicians and physicists had developed calculus into an indispensable tool for the study of the natural world, and the problems that arose from this connection led to a wealth of new concepts and methods – for example, ways to solve differential equations – that made calculus one of the richest and hottest research areas in the whole of mathematics. The beauty and power of calculus had become undeniable. However, Bishop Berkeley's criticisms of its logical basis remained unanswered, and as people began to tackle more sophisticated topics, the whole edifice started to look decidedly wobbly. The early cavalier use of infinite series, without regard to their meaning, produced nonsense as well as insights. The foundations of Fourier analysis were non-existent, and different mathematicians were claiming proofs of contradictory theorems. Words like 'infinitesimal' were being bandied about without being defined; logical paradoxes abounded; even the meaning of the word 'function' was in dispute. Clearly these unsatisfactory circumstances could not go on indefinitely.
>
> Sorting it all out took a clear head, and a willingness to replace intuition by precision, even if there was a cost in comprehensibility. The main players were Bernard Bolzano, Cauchy, Niels Abel, Peter Dirichlet, and, above all, Weierstrass. Thanks to their efforts, by 1900 even the most complicated manipulations of series, limits, derivatives and integrals could be carried out safely, accurately and without paradoxes. A new subject was created: analysis. Calculus became one core aspect of analysis, but more subtle and more basic concepts, such as continuity and

[1] http://www.math.okstate.edu/~wli/teach/fmq.html

limits, took logical precedence, underpinning the ideas of calculus. Infinitesimals were banned, completely."

Ian Stewart[2]

Calculus, as learned by undergraduates, roughly develops into two distinct, but interconnected, branches of mathematics. One branch, differential calculus, leads to the subject of differential geometry; the other, integral calculus, is the starting point for the subject of measure and integration. Subjects such as Complex Analysis and Ordinary and Partial Differential Equations are also based on calculus, but cannot be cleanly categorized in either branch. The catchall term *analysis* is used in mathematics to loosely include all subjects which are outgrowths of calculus, covering such a broad and diverse spectrum of mathematics as to defy a good definition, although it roughly means those parts of mathematics which involve infinite processes and limits; it is also often described as the mathematics of change or of continuous processes. (The word "analysis" in nontechnical English has a different and even broader meaning, which would include most of mathematics.)

I.2. Differential Calculus

I.3. Measure and Integration

Measure and integration, besides being a rich and beautiful theory in its own right, is an indispensable tool in most parts of pure and applied mathematics. To the working mathematician, measure and integration are as fundamental a tool as calculus and linear and abstract algebra. Whole parts of mathematics, including Probability Theory and Functional Analysis, are based primarily on the theory of measure and integration; other subjects, such as the modern theory of Ordinary and Partial Differential Equations, also owe their foundations to this theory.

Either measure or integration can be taken as the fundamental concept, and the other developed from it. It is somewhat a matter of taste which way to do this, but it is valuable to be able to think both ways since each approach is natural in some important contexts.

I.3.1. Measure

I.3.1.1. Measure theory, developed in the early twentieth century, gave the first mathematically satisfactory resolution to two foundational problems of analysis and geometry, one which caused great controversy especially in the sixteenth and seventeenth centuries in Europe (see [Ale15] for a study), and one which was only realized much later. (Measure theory later turned out to be good for a lot of other things too, such as formulating a rigorous theory of probability.)

The Paradox of Indivisibles

I.3.1.2. Should a line be regarded as a set of points? Any discussion of two lines intersecting at a point, drawing line through two given points, or especially considering the points between two given points on a line, all basic notions of Euclidean geometry, would seem to implicitly entail that certain points are at least "located" on a given line, almost (but not necessarily quite the same thing as) saying that the line is

[2] [?, p. 149].

composed of a collection of points; indeed, EUCLID refers to points being "on" a line and lines being "on" a plane. Thus it is hard to imagine that it could be controversial that a line is (or can be thought of as) a collection of points, but it was. (The question is actually more subtle than may appear, for deeper reasons than those discussed in previous centuries; see I.6.1.73..)

I.3.1.3. Actually, we must first consider a more basic question: is a mathematical line (or plane, or three-dimensional space) something which exists in a well-defined and uniquely determined form, reflecting properties of physical space, as apparently believed by everyone until the nineteenth century, or is it a human mental construct which can be made in various ways, with properties depending on the way it is defined? If one believes the former, then observed properties of physical space place limits on what can be believed or contemplated about mathematical space. (Our modern point of view is that mathematical space comes in many forms, some of which provide good, if imperfect [so far], *models* of physical space; indeed, our perception of physical space is via these mathematical models, not vice versa.)

I.3.1.4. Once we have a well-defined notion of what a line is (say a Euclidean line), the next problem to be considered is whether the number of points ("indivisibles") on a line segment is finite or infinite. If there are only finitely many points on a line segment, either the points must have a positive diameter and touch each other, or there must be gaps between them. Either hypothesis seems untenable, and few mathematicians or even nonmathematicians have taken this possibility seriously; a mathematical line, or mathematical space, is very different in character from physical objects made up of a large but finite number of atoms.

I.3.1.5. The other possibility is that a line segment is made up of infinitely many points (indivisibles). The next question then is whether for each point there is a "next" point, and if so whether the points "touch" or whether there is a gap between them. GALILEO thought there were infinitely many points, each of infinitesimal size, with gaps of infinitesimal length between them (he did not attempt to define what he meant by "infinitesimal"). Nowadays (and gradually since DESCARTES) we identify points on a (Euclidean) line with real numbers, and it is easily seen that between any two numbers there are infinitely many more, and in particular, there is no such thing as two consecutive numbers. In fact, in this picture of a line, every subinterval of a line segment, no matter how tiny, looks just like the whole segment, so a line is "infinitely divisible"; but contrary to what most if not all people from the sixteenth century believed, there is no logical contradiction in saying a line is infinitely divisible but that it is made up of points which are indivisible. In our picture of a line as the set of (real) numbers, the questions of whether points on a line "touch" or whether there are "gaps" become very subtle if not meaningless (see III.1.9.).

I.3.1.6. Whether or not there are consecutive points, if it is assumed that a line segment contains infinitely many points there is a basic paradox, which was one of the roots of the controversy over the theory of indivisibles. The length of the interval should be the sum of the sizes (diameters) of the points (and gaps, if any). Presumably all the points have the same character, and in particular have the same size. If the diameter of each is positive, the length of the interval is the sum of infinitely many equal positive numbers, which is infinite. And if the diameter of each is zero, then the length of the interval is the sum of infinitely many 0's, so should be zero. How can the length be a finite positive number?

I.3.1.7. Mathematicians (and nonmathematicians) worried about this paradox for a long time, but through the development of Calculus it was glossed over, or if considered was resolved using the vague

(at the time) notion of "infinitesimal numbers," numbers which were positive but equal to 0 for practical purposes. It was not until measure theory was developed in the early twentieth century that a satisfactory and rigorous solution to the paradox was found (and not until the 1960s was a rigorous theory of infinitesimals developed, a separate story; cf. III.10.).

I.3.1.8. The key idea was the sensational discovery by CANTOR in the late nineteenth century that there are different sizes of infinity, i.e. infinite sets could have different sizes. The smallest infinity is "countable infinity": a set is countably infinite if its elements can be listed in a sequence. There are also "uncountable infinities" (actually many different ones!): a set is uncountable if it has "too many" elements to be listed in a sequence, i.e. any sequence of elements from the set necessarily leaves out some ("most") elements of the set. CANTOR showed not only that uncountable sets exist, but even that every interval of numbers is uncountable, i.e. every line segment contains uncountably many points. The resolution of the paradox is that while a sensible definition of the sum of countably many numbers can be made (this is the theory of infinite series), no reasonable definition of the sum of uncountably many numbers is possible. Thus it cannot be sensibly said that the length of an interval is the (uncountable) sum of the (zero) diameters of the points.

I.3.1.9. The theory of Lebesgue measure is built on this resolution. A notion of "size" of subsets of $\mathbb{R}$ is defined, which has the property that the size of an interval is its length, and in particular the size of a point (singleton) is zero, and the size of a *countable* disjoint union of sets is the sum of the sizes of the sets. There is no strict numerical relationship (only an inequality) between the size of an uncountable union of disjoint sets and the sizes of the sets themselves, so in particular there is no inconsistency between the size of an interval, which is an uncountable disjoint union of single points, and the sizes of the points. (Actually, it turns out that not all subsets of $\mathbb{R}$ have a well-defined "size" (Lebesgue measure), but the ones which do not are so bizarre they cannot even be explicitly described, and never come up in practice.)

I.3.1.10. A variation of the theory of indivisibles is to regard a plane figure as being "made up" of a collection of parallel line segments, or a solid region in $\mathbb{R}^3$ being made up of a collection of parallel slices, each a plane figure, although these line segments or slices are technically not "indivisible." One has the same paradox as in the one-dimensional case: the area of the plane figure should be the sum of the areas of the line segments, and similarly in the solid case. The paradox is resolved in the same way via measure theory. But even before this can be done, another foundational problem must be resolved.

What Do Area and Volume Mean?

I.3.1.11. The area of a plane region, or the volume of a solid, seem to have a self-evident meaning. The volume of a physical object seems to be a well-defined concept, and can sometimes be measured, for example, by immersing it in water and measuring the amount of water displaced (under the assumption that a fixed quantity of water keeps the same volume no matter what shape it is put into), or by weighing it (assuming it is made of homogeneous material of a fixed known density). Area is less obviously physically meaningful, although the area of an agricultural field, for example, can be computed or estimated by the amount of crop it yields. But it is a leap of faith to assume that these notions carry over to abstract regions in mathematical space.

It apparently did not occur to anyone before the twentieth (or perhaps late nineteenth) century – not EUCLID, not ARCHIMEDES, not NEWTON, not EULER, not GAUSS, not CAUCHY – that area or volume had to be *defined*: they were just *there*, and merely needed to be calculated (although ARCHIMEDES did realize that surface area of a curvilinear surface is not inherently well defined, not even as well defined as arc length

of a curve, and took the trouble to axiomatize surface area; see I.6.1.8. and ()). This issue is on top of the more basic question discussed earlier of whether lines, planes, or space are themselves well defined. Not only was it implicitly assumed that area (volume) was a meaningful concept, it was also assumed that it had some "obvious" properties which the physical versions evidently satisfied: congruent regions have the same area, the area of a disjoint union is the sum of the areas of the components, the area of a rectangle is $length \times width$, the area of a curve is 0, and hence the area of the boundary of a region is 0, so the area of the union of "nonoverlapping" regions is the sum of the areas, etc. See I.6.1.3. for a more extensive discussion of the issue.

I.3.1.12. So how do we define the area of a region, and how do we show that area has the expected properties? Measure theory gives a satisfactory answer. We first define the area of a rectangle to be $length \times width$, and then by a somewhat involved process extend the definition to sets which can be built up from rectangles by countably many set operations (*Borel sets*). As in the case of Lebesgue measure on the line, not every subset of the plane has a well-defined area by this process, but all "reasonable" subsets do, and in particular all the sets considered during the long development of geometry and calculus (going back at least to Euclid and Archimedes) have an area by this definition, which agrees with the intuitive areas already calculated by algebraic and calculus methods. The same procedure gives a rigorous definition of volume for subsets of $\mathbb{R}^3$, agreeing numerically with the informal or assumed "volume" calculated by earlier methods. The expected properties of area must be proved from the definition, but can be shown to hold (although there are some subtleties, e.g. the area of the boundary of a region is not necessarily zero if the boundary is sufficiently complicated).

The Nature of Measure Theory

I.3.1.13. The goal of the subject of measure theory is to associate a numerical "size" to subsets of a set X in a precise, consistent, and well-behaved manner. The principal property that such a notion of "size" should have is that the "size" of a union of disjoint subsets should be the sum of the sizes of the sets.

I.3.1.14. It turns out that in most applications it is both natural and necessary to restrict attention only to certain well-behaved subsets of X; frequently some subsets are too bizarre to allow a reasonable notion of "size." In addition, to get a reasonable theory one must only allow countably many operations. The technical term is that the collection of "measurable sets" (subsets with a "size" or *measure*) form a *σ-algebra* (see () for a precise definition.)

I.3.1.15. In many instances there is a natural notion of size for a collection of subsets which is not large enough to form a σ-algebra, and the task is then to extend this size notion to a larger σ-algebra of sets.

There are two fundamental examples of this process, which have motivated most of the development of the theory:

I.3.1.16. Example. As discussed above, let X be the real numbers $\mathbb{R}$. There is a natural size for certain nice subsets, the intervals: if I is the interval from a to b (with or without endpoints), the measure of I should be the length $b-a$. It turns out that this measure can be extended to a σ-algebra of subsets of $\mathbb{R}$ in such a way that congruent sets have the same measure. The resulting measure is called *Lebesgue measure* on $\mathbb{R}$, and gives by far the most natural and important example of a measure space. It turns out that not

every subset of $\mathbb{R}$ can be Lebesgue measurable, but every subset which can be explicitly described, and in particular any set which arises in any reasonable way in applications, is measurable.

Similarly, the plane $\mathbb{R}^2$ has a size notion ("area"), as does three-dimensional Euclidean space $\mathbb{R}^3$ ("volume"), and more generally $\mathbb{R}^n$ for any n. These size notions are only defined at first for nice subsets such as rectangular solids, but can be extended in the same manner as in $\mathbb{R}$. These cases are a little harder intuitively than the one-dimensional case: for example, even finding the area of a set like a disk requires a nontrivial limiting process. However, the general theory applies in just the same manner as in the one-dimensional case to give n-dimensional Lebesgue measure.

There are other natural measures on Euclidean space. For example, there is a measure on $\mathbb{R}^2$ which assigns to a smooth curve its arc length. This measure, called one-dimensional Hausdorff measure, would give infinite size to any two-dimensional set. Hausdorff measures giving fractional dimensions to fractal sets can also be defined. In fact, measure theory underlies the whole study of fractals.

I.3.1.17. EXAMPLE. The basic setup of probability theory is to analyze a process or "experiment" to numerically describe the possible results. The set of all possible outcomes is called the *sample space*, which we will denote Ω. A subset E of Ω is called an *event*. In the simplest situation, Ω is just a finite set, and each point has a certain probability of occurring; the sum of all the individual probabilities is 1. The probability that an event E occurs is the sum of the probabilities of the points in E.

However, the set Ω is often infinite, and individual points have probability 0, but infinite events may have positive probability. For example, suppose the process is to choose a number x randomly from $[0, 1]$. The sample space Ω is thus $[0, 1]$. If x_0 is any fixed number, the probability that the chosen x is exactly equal to x_0 is 0. However, the probability that $0 \leq x \leq 1/2$ ($E = [0, 1/2]$) is $1/2$. More generally, the probability that $a \leq x \leq b$ is $b - a$.

The proper setup for probability theory is to take the set of allowable events to be a σ-algebra of subsets of Ω, and the probability of events is given by a measure on this σ-algebra with the measure of the whole space Ω equal to 1. In the above example, the probability is given by Lebesgue measure on $[0, 1]$.

In fact, most (but not all!) of the elementary concepts of probability theory are measure theory concepts with different terminology. A simple dictionary is all that is needed to make the translation.

I.3.2. Integration

I.3.2.1. Integration is the other basic component of this part of real analysis. The goal of integration is to systematically assign to a real-valued function f on a set X, and a subset A of X, a number $\mathcal{I}_A(f)$ called the *integral of f over A* (the notation $\int_A f$, or some variation, is usually used; the symbol $\int$ is called the *integral sign*), which is usually thought of as a "normalized sum" of the values of f on the points of the subset A. This number has various interpretations in applied situations as total distance (regarding f as giving instantaneous velocity), total mass (regarding f as giving density), etc. Integration in analysis is thus almost always (and always in this book) a generalization or analog of what in elementary calculus is called the "definite integral." (We will also have occasion to consider analogs of the "indefinite integral" of calculus, but these will have different names and will not be called integrals.)

I.3.2.2. All modern theories of integration are in a sense outgrowths of the so-called Riemann integral (actually developed first by CAUCHY for continuous functions and generalized by RIEMANN to also

handle moderately discontinuous functions), the type of integration treated in elementary calculus. The Riemann integral was the first reasonably satisfactory theory of integration, and is still by far the theory most widely used in applications; Riemann integrals can in many cases be exactly and efficiently computed using antiderivatives and the Fundamental Theorem of Calculus, and there are good numerical techniques for approximating Riemann integrals that cannot be computed exactly. In fact, more general integrals are rarely actually computed directly; in practice they are usually approximated by Riemann integrals which are then computed or estimated using standard techniques.

For theoretical purposes, Riemann integration has two serious inadequacies:

1. The class of Riemann integrable functions is too small. For a function to be Riemann integrable on an interval, it must be bounded and "almost continuous" (see () for a precise statement). It is particularly useful to be able to integrate certain unbounded functions. This is reasonably, but not entirely, satisfactorily achieved by using various notions of improper Riemann integrals and principal values. It is also important to be able to integrate over unbounded intervals, which can be done through other types of improper integrals, although not in an entirely satisfactory way. At least for theoretical purposes, it is also important to be able to integrate functions which are too discontinuous to be Riemann integrable, and to be able to integrate functions over subsets of the real numbers more general than intervals or finite unions of intervals (although the practical value of this type of integration is sometimes somewhat overemphasized as an inadequacy of Riemann integration).

2. It is crucial in many settings to have limit and approximation theorems, i.e. to know that under reasonable conditions, if $f_n \to f$ in some sense, we have $\mathcal{I}_A(f_n) \to \mathcal{I}_A(f)$. It is quite difficult to establish good theorems of this sort for Riemann integration which are applicable even if we know that the limit function is Riemann integrable (which is often not the case due to the limitations described above). It turns out that to get reasonable convergence theorems, more general integration theories are needed.

Another limitation of Riemann integration is that it only (directly) works on $\mathbb{R}$, or on $\mathbb{R}^n$. It is very important to have integration theories which work on more general spaces, and it turns out that modern approaches give this flexibility with essentially no additional work.

> "Lebesgue's definition enables us to integrate functions for which Riemann's method fails; but this is only one of its advantages. The new theory gives us a command over the whole subject which was previously lacking. It deals, so to speak, automatically with many of the limiting processes which present difficulties in the Riemann theory."
>
> *E. C. Titchmarsh*[3]

There are three principal modern approaches to integration, each with many variations. We first discuss the basics of Riemann integration in a form allowing the various generalizations to be made naturally.

[3] [?, p. 318].

I.4. Other Topics

"The object to be attained by the theory of functions of a real variable consists then largely in the precise formulation of necessary and sufficient conditions for the validity of the limiting processes of Analysis. ... In the literature of the subject, errors are not infrequent, largely owing to the fact that spatial intuition affords an inadequate corrective of the theories involved, and is indeed in some cases almost misleading."

E. W. Hobson[4]

"If you have an apple and I have an apple and we exchange apples then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas."

George Bernard Shaw

"Mathematics is a huge and highly interconnected structure. It is not linear. As one reads mathematics, one needs to have an active mind, asking questions, forming mental connections between the current topic and other ideas from other contexts, so as to develop a sense of the structure, not just familiarity with a particular tour through the structure. ... Think of a Tinkertoy set. The key is the pieces which have holes, allowing you to join them with rods to form interesting and highly interconnected structures. No interesting mathematical topic is self-contained or complete: rather, it is full of 'holes,' or natural questions and ideas not readily answered by techniques native to the topic. These holes often give rise to connections between the given topic and other topics that seem at first unrelated."

W. Thurston[5]

[4] [**?**].
[5] [Thu97, p. ix].

I.5. What is Mathematics, and Why the Bad Reputation?

> "Despite these [previously described] by no means modest contributions to our life and thought, educated people almost universally reject mathematics as an intellectual interest. This attitude toward the subject is, in a sense, justified. School courses and books have presented 'mathematics' as a series of apparently meaningless technical procedures. Such material is as representative of the subject as an account of the name, position, and function of every bone in the human skeleton is representative of the living, thinking, and emotional being called man. Just as a phrase either loses meaning or acquires an unintended meaning when removed from its context, so mathematics detached from its rich intellectual setting in the culture of our civilization and reduced to a series of techniques has been grossly distorted. Since the layman makes very little use of technical mathematics, he has objected to the naked and dry material usually presented. Consequently, a subject that is basic, vital, and elevating is neglected and even scorned by otherwise highly educated people. Indeed, ignorance of mathematics has attained the status of a social grace."
>
> *Morris Kline*[6]

It is socially acceptable, even fashionable, to be ignorant of mathematics. This is curious: no one would brag about being illiterate. Why is mathematics different?

The answer may be that most members of the public, even those who take university mathematics through calculus, have never actually seen real mathematics, despite many years of dreary study of mechanical techniques of elementary mathematics. Mathematics no more consists of learning and applying the mechanical techniques of arithmetic and algebra (or even calculus) than English consists of learning to spell words and combine them into sentences or history consists of learning a mass of dates and names. The difference may be that most people move on from these necessary basics in these other subjects at a fairly early age, while in mathematics most never get past them. Or maybe it is because logical thinking and reasoning is hard and intellectually challenging work which many people are unaccustomed to and uncomfortable with.

This is not a new phenomenon:

> "When I first applied my mind to Mathematics I read straight away most of what is usually given by the mathematical writers, and I paid special attention to Arithmetic and Geometry because they were said to be the simplest and so to speak the way to all the rest. But in neither case did I then meet with authors who fully satisfied me. I did indeed learn in their works many propositions about numbers which I found on calculation to be true. As to figures, they in a sense exhibited to my eyes a great number of truths and drew conclusions from certain consequences. But they did not seem to make it sufficiently plain to the mind itself why these things are so, and how they discovered them. Consequently I was not surprised that many people, even of talent and scholarship, should, after glancing at these sciences, have either given them up as being empty and childish or, taking them to be very difficult and intricate, been deterred from the very outset from learning them ... But when I afterwards bethought myself how it could be that the earliest pioneers of Philosophy in bygone ages refused to admit to the study of wisdom any one who was not well versed in Mathematics ... I was confirmed in my suspicion that they had knowledge of a species of Mathematics very different from that which passes current in our time."
>
> *René Descartes*[7]

[6] [**?**, p. vii-viii].
[7] cf. [**?**, p. 1].

The problem is in no small measure with the way mathematics is taught:

> "[M]any people grow up feeling great antipathy towards math, probably because of how they were taught it at school, as a set of facts you're supposed to believe, and a set of rules you have to follow. You're not supposed to ask why, and when you're wrong, you're wrong, end of story. The important stage between the belief and the rules has been omitted: the illuminating reasons. An illuminated approach is much less baffling, much less autocratic, and much less frightening. ... Mathematics can seem like an autocratic state with strict, unbending rules that seem arbitrary to the citizens of this 'state': the pupils and students. Schoolchildren try to follow the rules but are sometimes abruptly told that they have broken a rule. They didn't do it deliberately. Most students who get some math questions wrong didn't do it on purpose – they really thought they had the right answer. Yet they're told they have broken the law and will be punished – being marked wrong feels like a punishment to them. Perhaps it is never really explained to them what they did wrong, or perhaps it was not explained to them in an illuminating way that could actually make sense to them. As a result, they don't know when they will next be found to have broken a rule, and they will creep around in fear. Eventually they'll simply want to escape to a more 'democratic' place, a subject in which many different views are valid."
>
> *Eugenia Cheng*[8]

Unfortunately, we have degenerated into a vicious circle: many math teachers (as well as parents), especially at the elementary level, grew up with a fear and loathing of mathematics themselves and have never properly understood the subject, and pass on to their students, either consciously or subconsciously, a feeling that it is all right and even admirable for a "normal" person to dislike mathematics and be unable to understand it. Explanations are not part of the subject, and they are incapable of providing adequate ones anyway. This attitude is deeply engrained in the education establishment: the "math education specialist" in the College of Education at our university actually once told us they like prospective math teachers who struggled with the subject themselves, since they can "relate to the students"! (Do they like prospective English teachers who have trouble reading and writing?)

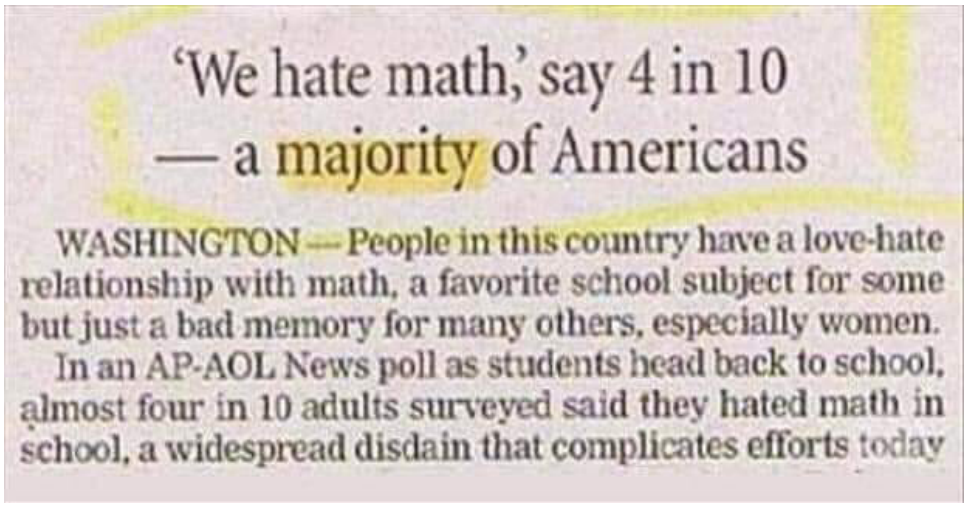
> "Math is about getting the right answer."
>
> *Ron DeSantis, Governor of Florida*[9]

> "Math might seem like it's about getting the right answers, but really it's about the process of discovering, the process of exploration, the journey toward mathematical truth, and how to recognize when we've found it. That journey starts with curiosity, and curiosity makes itself known in the form of questions.."
>
> *Eugenia Cheng*[10]

[8] [Che15, p. 277-278].

[9] Trying to explain why he rejected 41% of the proposed elementary math textbooks despite having no background in mathematics. The books were rejected because they included social context and tried to address self-image and math anxiety, and promoted logical reasoning instead of cookbook procedures.

[10] [**?**, p. 11]. An entire chapter of this book, and indirectly the entire book, is a refutation of the phrase "Math is about getting the right answer." On p. 52: "[T]oo often [math] is presented as if it is only about getting the right answers to things. It is presented as if there are facts that you have to know. The facts are announced by the teacher, and the students' role is to learn those facts, not question them. As a result, the role of the teacher is to announce those facts, not justify or explain them."



'We hate math,' say 4 in 10 — a majority of Americans

WASHINGTON — People in this country have a love-hate relationship with math, a favorite school subject for some but just a bad memory for many others, especially women.

In an AP-AOL News poll as students head back to school, almost four in 10 adults surveyed said they hated math in school, a widespread disdain that complicates efforts today

Figure I.1: Source mercifully unknown.

It does not have to be this way:

> "If you've ever spent time with a three-year-old, you'll know that children that age never stop asking why. Ever. ... Unfortunately, we suppress this instinct in children, possibly just because it gets rather tiresome after a while. Possibly because we quite quickly get to the point where we don't know the answer, and we don't like having to say 'I don't know.' Or we don't like reaching the end of our own understanding of things. But this natural instinct in children is beautiful. It's the difference between *knowledge* and *understanding*. ... At the heart of math is the desire to understand things rather than just know them." [Emphasis original]
>
> *Eugenia Cheng*[11]

The idea that mathematics is beyond the ability of the average person is as preposterous as the notion that the average person cannot learn to read. When children struggle to learn to read, our society goes to great lengths to remedy the problem. But when children struggle with math, the tendency is to shrug and accept the difficulty as "normal." While it is true that most people are not capable of doing research-level mathematics, just as most people are not capable of writing novels or composing symphonies, mathematical literacy is well within the abilities of almost everyone.

So what is mathematics, really? This is actually a tough question, with no simple answer. Just as English really consists of understanding how to express ideas and emotions in comprehensible form, and history consists of understanding events and the development of social structures (these are obviously inadequate definitions), mathematics essentially consists of understanding logical reasoning and drawing necessary conclusions, and developing the abstract structure to accomplish these goals. Mathematicians make precise definitions of (almost) all terms and concepts within mathematics, but remarkably it is virtually impossible to give a precise definition of mathematics itself. The definition that mathematics is the study of pattern and

[11] [Che15, p. 61].

form, while not entirely adequate, seems to be as good as any. We will elaborate on this nutshell definition over the next several sections.

> "A mind which rests content with a collection of facts is no scientific mind. The formulas in a mathematical handbook are no more mathematics than are the words in a dictionary a literary masterpiece. Until some unifying principle is conceived by which an amorphous mass of details can be given structure, neither science nor mathematics has begun."
>
> *E. T. Bell*[12]

> "In its broadest aspect mathematics is a spirit, the spirit of rationality. It is this spirit that challenges, stimulates, invigorates, and drives human minds to exercise themselves to the fullest."
>
> *Morris Kline*[13]

> "Math suffers from a reputation that it is dry and difficult. Often it is. Yet math can also be inspiring, accessible and, above all, brilliantly creative. Abstract mathematical thought is one of the great achievements of the human race, and arguably the foundation of all human progress. ⋯ The world of mathematics is a remarkable place. I would recommend a visit."
>
> *Alex Bellos*[14]

[12] [Bel92, p. 45].
[13] [**?**, p. 10].
[14] [**?**, p. xi].

I.6. Rigor in Mathematics

> "A simple man believes every word he hears; a clever man understands the need for proof."
>
> *Proverbs 14:15*[15]

> "After deserting for a time the old Euclidean standards of rigour, mathematics is now returning to them, and even making efforts to go beyond them. ... Later developments ... have shown more and more clearly that in mathematics a mere moral conviction, supported by a mass of successful applications, is not good enough. Proof is now demanded of many things that formerly passed as self-evident. Again and again the limits to the validity of a proposition have been in this way established for the first time. ... In all directions these same ideals can be seen at work – rigour of proof, precise delimitation of extent of validity, and as a means to this, sharp definition of concepts."
>
> *Gottlob Frege, 1884*[16]

A student, particularly a nonmathematics student, may well wonder about two questions:

1. Why is it necessary to spend the time and effort to develop a precise, rigorous, and complete theory of measure and integration (or any other mathematical topic)? Isn't this just an idle intellectual exercise for mathematicians with nothing better to do, without practical benefit to those using the theory? Does there have to be an elegant general theory behind every useful piece of mathematics?

2. Even given that *someone* ought to carefully work out the details, isn't it enough that someone already *has*? Why can't we just use the theory without having to immerse ourselves in the details?

Both questions are legitimate ones, and deserve serious answers.

Let us discuss the second question first. Here there is a good analogy. It is quite possible for a person to use a machine without understanding all the technical details of its construction and operation. However, it is quite important that the user, in addition to simply knowing how to operate the machine, understand enough about it to know what it can and cannot do, and what dangers it may pose; failure and even disaster can result from careless operation or trying to use the machine beyond its capabilities. Also, the user might eventually want to have the machine do something it was not designed for, and the machine may have to be modified to accomplish the new task. It would likely be impossible to figure out how to adapt the machine to a new task without a good understanding of its construction.

In addition, the existing theory is so vast that it is often easier to work out what is needed in a particular situation instead of doing a possibly long and involved search for exactly the right "off-the-shelf" result that applies. One needs a good understanding of the theory to do this (and indeed even to recognize when an off-the-shelf result applies and suffices!)

> "The problem may be viewed as one of information retrieval; we wish to find a particular piece of information, not knowing for sure that it has ever been recorded, and instead of trying to provide a catalog of known results (which would necessarily be finite and to a great extent represent past

[15] New English Bible

[16] G. Frege, *The Foundations of Arithmetic*, translated by J. L. Austin (Oxford, 1950), §1.

interests), we prefer a method of creating it as needed. *Instead of information retrieval we prefer information regeneration.* Of course this attitude will not work in all areas of knowledge, but in a mature science the principles should show how to recreate the information on demand."

R. W. Hamming[17]

The first question is almost asking: What is mathematics? To begin with the last part of Question 1, it is certainly not true that every bit of mathematics has to be backed by a rigorous general theory to be useful. For example, a lot of analysis beginning with calculus was developed and extensively used for about 200 years before its logical underpinnings were carefully worked out in the nineteenth century. Even today, where standards of rigor are high and mathematicians rarely if ever use techniques and results whose validity has not been carefully established, there are many heuristic and *ad hoc* methods not backed up by general theory, especially in applied mathematics. This is in no way a criticism of this work, or of applied mathematics in general; if applied mathematicians had to wait for general theories underlying their techniques before using them, progress would grind nearly to a halt (and the general theories might never come.)

A good example is fractal geometry. Some mathematicians question whether this subject should even be regarded as part of mathematics, since there are almost no theorems and the work is almost entirely descriptive rather than analytical. I personally do not take such a narrow view of mathematics, and I think this subject is unquestionably part of mathematics; after all, did calculus only become part of mathematics after Weierstrass?

Perhaps an even more striking example is the "theory" of quantum groups. Mathematicians seem to know a quantum group when they see one, and there has been a great deal of work on quantum groups over the last several decades, some of which is widely regarded as important, first-rate mathematics. But there is still as yet no generally accepted *definition* of a quantum group!

Nonetheless, the usefulness and even necessity of rigorous general theories is not diminished by such work, and indeed the desirability and urgency of a general theory explaining and connecting specific results and techniques is greatly enhanced by the applications.

Authority in Mathematics

"Mathematics is fundamentally an egalitarian subject. It should in principle be possible for anyone to check the correctness of any argument; nothing needs to be taken on trust."

J. K. Truss[18]

(See III.1. for a continuation of this quote.)

"Believe nothing on hearsay. Do not believe in traditions because they are old, or in anything on the mere authority of myself or any other teacher."

Buddha[19]

[17] [**?**, p. 244]
[18] [Tru97, p. v]
[19] [Bel92, p. xi].

> "No man can worthily praise Ptolemye ... yet muste ye and all men take heed, that both in him and in all mennes workes, you be not abused by their autoritye, but evermore attend to their reasons, and examine them well, ever regarding more what is saide, and how it is proved, than who saieth it, for autorite often times deceaveth many menne."
>
> *Robert Recorde, 1556*[20]

> "In questions of science the authority of a thousand is not worth the humble reasoning of a single individual."
>
> *Galileo*[21]

We should never rely entirely on an "authority" to tell us what is true in mathematics, although we should give credence to the guidance of knowledgable people. For one thing, any "authority" can be wrong. Nobody has a corner on the truth in mathematics; some just have a somewhat better handle on it than others.

A simple illustration is the story told [**?**, p. 73] by John Paulos, a mathematician and author of the popular book *Innumeracy*, who at age ten calculated the earned run average of a certain baseball pitcher to be 135 (a pitcher's earned run average, or ERA, is the average number of earned runs allowed per nine innings pitched, calculated by the formula

$$\text{ERA} = \frac{9 \cdot (\text{number of earned runs allowed})}{\text{innings pitched}} .$$

This pitcher had allowed 5 earned runs in 1/3 inning pitched.) He showed his calculation to his teacher, who authoritatively replied that he was wrong, that an ERA could never be higher than 27 (the number of outs in 9 innings). When the published official statistics confirmed his calculation, he felt vindicated, and recalls:

> "I remember thinking of mathematics as an omnipotent protector. You could prove things to people and they would have to believe you whether they liked you or not."

Unfortunately, his teacher still did not believe him even when presented with the published proof!

Students, particularly at lower levels, are often far too willing to defer to the "authority" of instructors in mathematics. I once had a calculus student who came to my office because he could not get his answer to an online homework problem to match what the computer demanded (a serious ongoing problem with online homework, although things are slowly improving). Instead of the "right" answer of 2, he kept getting $\frac{3+\sqrt{1}}{2}$. I told him to simplify the square root to 1. He thought a minute, then said, "Wait a minute. Is $\sqrt{1}$ 1?" When I assured him it was, he thought a minute more, then said, "Wait a minute. Is $\sqrt{1}$ *always* 1, or just in this course?" I refrained from suggesting he ask his Calculus II instructor the next semester what $\sqrt{1}$ is in that course. (Incidentally, inability of the computer to recognize alternate forms of an expression is not the only difficulty with online homework. Another student once complained to me that the computer told him that his answer to an online problem was "incorrect, although it is equal to the correct answer.")

[20] *The Castle of Knowledge*; cf. J. Borwein and D. Bailey, *Mathematics by Experiment: Plausible Reasoning in the 21st Century*, A K Peters, 2004, p. 12.

[21] Quoted in Arago's Eulogy on Laplace; cf. [Mor93, 1528].

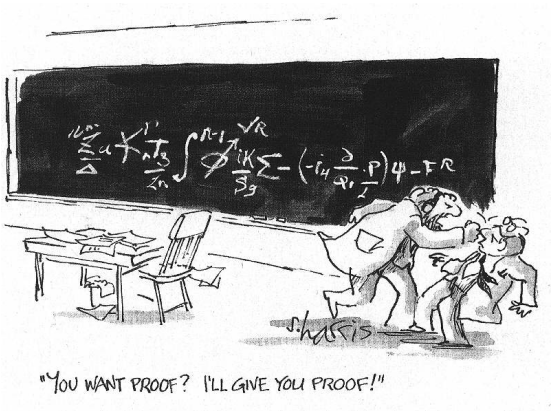


Figure I.2: Proof by intimidation.

Abstraction, Interpretation, and Point of View

> "The more comprehensive the subject is they treat, and the more general and abstract the manner is in which they are considered, the more are [mathematical] principles free of darkness and the easier it is to grasp them."
>
> *Jean d'Alembert, 1743*[22]

> "The purpose of axiomatization in mathematics is to reduce to a minimum the presuppositions on which a given theory depends, and at the same time to make every presupposition fully explicit."
>
> *B. Rotman and G. T. Kneebone*[23]

> "The most striking characteristic of the written language of algebra and of the higher form of the calculus is the sharpness of definition, by which we are enabled to reason upon the symbols by the mere laws of verbal logic, discharging our minds entirely of the meaning of the symbols, until we have reached a stage of the process where we desire to interpret our results. The ability to attend to the symbols, and to perform the verbal, visible changes in the position of them permitted by the logical rules of the science, without allowing the mind to be perplexed with the meaning of the symbols until the result is reached which you wish to interpret, is a fundamental part of what is called analytical power. Many students find themselves perplexed by a perpetual attempt to interpret not only the result, but each step of the process. They thus lose much of the benefit of the labor-saving machinery of the calculus and are, indeed, frequently incapacitated for using it."
>
> *Thomas Hill*[24]

[22] J. d'Alembert, *Traité de dynamique*, Paris, 1743.
[23] B. Rotman and G. T. Kneebone, *The Theory of Sets and Transfinite Numbers*, American Elsevier, 1968, p. 62.
[24] Uses of Mathematics, *Bibliotheca Sacra*, v. 32, p. 505; cf. R. Moritz, *Memorabilia Mathematica*, 1209.

> "Mathematics is a way of thinking that can help make muddy relationships clear. It is a language that allows us to translate the complexity of the world into manageable patterns. In a sense, it works like turning off the houselights in a theater the better to see a movie. Certainly, something is lost when the lights go down; you can no longer see the faces of those around you or the inlaid patterns on the ceiling. But you gain a far better view of the subject at hand."
>
> *K. C. Cole*[25]

One of the characteristic and essential features of mathematics is the fact that most parts of the subject can be approached or interpreted from various points of view. Not only does this variety of viewpoints increase the usefulness of mathematical theories, but the very progress of mathematics often depends on this flexibility: problems which appear obscure or difficult from one point of view often have obvious solutions when viewed in a different way.

In fact, modern mathematicians deliberately try to create theories which have flexibility and variety of viewpoints built in. This is one of the principal reasons for the abstraction of today's mathematics. The usual procedure is to base the entire theory on a few basic and generally abstract assumptions called *axioms*, and logically deduce all features of the theory from the axioms without making additional assumptions about the "meaning" of the axioms or the terminology used in them. Abstraction serves several purposes:

1. It simplifies the range of information which must be simultaneously considered and used in developing the theory, including the notation and terminology, to a manageable level.
2. It allows mathematicians to focus their thinking on the essential structural features of the theory without being tied to one particular interpretation of the theory which may constrain creative thought by introducing unimportant or even extraneous elements.
3. It allows the precise limits of the validity of the theory to be described.
4. It allows, even invites, new interpretations of the theory to be identified, leading not only to broader applicability of the theory and results already obtained, but also to new ways of looking at and solving problems arising in the development of the theory.
5. It makes it easier to notice similarities and connections with other parts of mathematics, leading both to a better understanding of the big picture of mathematics, and to cross-fertilization: importation of ideas and techniques from one part of mathematics into another.

> "Abstraction gives us the 'less is more' approach – fewer assumptions leave us in touch with more mathematical objects (and thus the theorems we prove apply in more instances). More importantly, however, it gives us the 'less is clearer' approach. Because we haven't made a myriad of assumptions, our mathematical choices are more clearly defined, our mathematical arguments are cleaner, and the conclusions we draw are less likely to be cluttered up by irrelevant issues. We can, therefore, sort out the essential mathematical ingredients that support a particular fact."
>
> *Carol Schumacher*[26]

[25] K. C. Cole, *The Universe and the Teacup*, 1998.
[26] [**?**, p. 5].

There are few if any coincidences in mathematics: when the same phenomenon occurs in different contexts, there is almost always a common explanation.

> "[A] single simple principle can masquerade as several difficult results; the proofs of many theorems involve merely stripping away the disguise."
>
> *M. Spivak*[27]

The Greek word "axiom", as used in mathematics, originally meant "self-evident truth." In formalist mathematics, axioms are regarded as "arbitrary assumptions." To most modern mathematicians, they are neither. The axioms used in a theory are a matter of choice and are not necessarily obvious. But they are not arbitrary; they are chosen to be a reasonable general starting point for a theory with at least one interpretation or application of interest. The terminology used in the axioms and throughout the theory is often, but not always, suggestive of a particular interpretation; such suggestive terminology can be beneficial in guiding a mathematician's thoughts, but also confining by restricting the mathematician's ability to "think outside the box" and to find and use other interpretations of the concepts. Sometimes an entirely equivalent theory can be developed from a completely different set of equally reasonable axioms, often using different terminology; when this occurs, the theory tends to be richer because different results can be more easily or naturally obtained in the two setups (and the different axiom schemes are also evidence of broader applicability of the theory). Measure theory is an excellent example of this; various approaches beginning with measure, probability, and integration lead to essentially the same theory, with many varied tools available from the different approaches.

> "Mathematics is the art of giving the same name to different things. It is enough that these things, though differing in matter, should be similar in form, to permit of their being run in the same mould. When language has been well chosen, one is astonished to find that all demonstrations made for a known object apply immediately to many new objects: nothing requires to be changed, not even the terms, since the names have become the same."
>
> *H. Poincaré*[28]

> "[T]he creation of a word or notation for a class of ideas may be, and often is, a scientific fact of very great importance, because it means connecting these ideas together in our subsequent thought."
>
> *J. Hadamard*[29]

[27] [Spi65, p. ix].
[28] [**?**, p. 34].
[29] [Had47, p. 38].

But a cautionary note must be expressed about runaway abstraction:

> "The real difficulty is to decide in the first place what aspects of a theory can be profitably abstracted. It is indeed easy to abstract almost anything in technical mathematics, but without insight or luck the result is a barren set of postulates and a sterile list of formulas. The abstract algebraists, the abstract analysts, and the abstract geometers of the twentieth century have at least been lucky."
>
> *E. T. Bell*[30]

Abstraction for its own sake is rarely worthwhile in mathematics. Even though it is usually impossible to foresee the full implications of a new theory, significant mathematical developments almost always occur when mathematicians have some interpretation or use in mind for the new theory; these can range from direct applications to important problems outside mathematics to simple curiosity about a specific mathematical problem that just seems interesting.

> "The axiomatic method has provided deep insights into mathematics, disclosing identities where none had been suspected. In the hands of mathematicians of genius this method has been used to strip away exterior details that seem to distinguish two subjects and to disclose an identical structure whose properties can be studied once for all and applied to the separate subjects."
>
> *Mina Rees*[31]

> "[T]he users of [mathematics] naturally enough take a somewhat skeptical view of this development [abstraction and generalization] and complain that the mathematicians now are doing mathematics for its own sake. As a mathematician my reply must be that the abstraction process that goes into functional analysis [and other parts of mathematics] is necessary to survey and to master the enormous material we have to handle. ... Our critics, especially those well-meaning pedagogues, should come to realize that mathematics becomes simpler only through abstraction. The mathematics that represented the conceptual limit for Newton and Leibniz is taught regularly in our high schools, because we now have a clear (i.e. abstract) notion of a function and of the real numbers."
>
> *G. Pedersen*[32]

[30] [Bel87, p. 162].
[31] "The nature of mathematics," *Science*, Oct. 5, 1962, p. 10.
[32] [Ped89, p. vii-viii].

Rigor is Not Always Appreciated

I.6.0.3. Nonmathematicians do not always understand or appreciate the efforts mathematicians make to rigorously prove results in analysis (or other parts of mathematics), and to specify the exact hypotheses under which results are valid; there is frequently even downright hostility. For example, there is the following quote from [Eag25, p. 139] (this book contains other mathematical nonsense; cf. [Bar00]):

> "... [S]o many books dealing with Fourier's series continually repeat the condition that the function must not have an infinite number of maxima and minima. We have generally omitted specifying this condition, since no practical function ever does behave in such a manner. Such behaviour is exclusively confined to functions invented by mathematicians for the sake of causing trouble."

[It is surely meant that the function does not have infinitely many extrema in any bounded interval; otherwise, how about sines and cosines?]

But how can the author be so sure? And what happens if he is wrong, and an engineer tries to apply what he says in a situation where it is not valid?

Even POINCARÉ expressed a somewhat similar sentiment, although he was smart enough to recognize that bizarre examples serve a legitimate mathematical purpose:

> "Logic sometimes makes monsters. For half a century we have seen a mass of bizarre functions which appear to be forced to resemble as little as possible honest functions which serve some purpose. More of continuity, or less of continuity, more derivatives, and so forth. Indeed from the point of view of logic, these strange functions are the most general; on the other hand those which one meets without searching for them, and which follow simple laws appear as a particular case which does not amount to more than a small corner.
>
> In former times when one invented a new function it was for a practical purpose; today one invents them purposely to show up defects in the reasoning of our fathers and one will deduce from them only that."
>
> *H. Poincaré*[33]

> "The whole problem with the world is that fools and fanatics are always so certain of themselves, and wiser people so full of doubts."
>
> *Bertrand Russell*

[33] cf. [Kli72, p. 973].

I.6.1. Semantics: What is Obvious?

" 'Obvious' is the most dangerous word in mathematics."

E. T. Bell[34]

"[S]ince people have tried to prove obvious propositions, they have found that many of them are false."

Bertrand Russell[35]

"Being a mathematician means that you don't take 'obvious' things for granted but try to reason. Very often you'll be surprised that the most obvious answer is actually wrong."

Evgeny Evgenievich[36]

"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so."

Mark Twain[37]

"[S]cience is always in need of those who are willing, and able, to say 'Hold on a minute, isn't there a problem with ... ?' "

Peter R. Law[38]

"The easiest person to fool is yourself."

Richard Feynman[39]

"I never met a man who thought his thinking was faulty."

Michel Montaigne[40]

It is a surprisingly delicate matter, which can confuse not only students but even professional mathematicians, to decide what in mathematics is "obvious" and can simply be asserted, and what must be carefully spelled out. Often the difficulty is a semantic one, stemming from misuse, or at least careless use, of language: language can be subtly suggestive in a misleading way. We will discuss a few such examples related to the foundations of analysis.

[34] [Bel87, p. 16].
[35] [**?**].
[36] [Fre13, p. 16].
[37] This is one of many quotes often attributed to Twain for which there is no documentation. The first published version of this statement appears to have been recorded by Josh Billings (Henry Wheeler Shaw) in 1874, but it was probably not original there; cf. https://quoteinvestigator.com/2018/11/18/know-trouble/?fbclid=IwAR3X8C7a7zR2ZdG7HC7t58CiN1dnqirWO3qOAYVOrICKwVQwcSsiKL6Cp2M. Another probably apocryphal quote attributed to Twain: "I didn't attend the funeral, but I sent a nice letter saying I approved of it."
[38] MathSciNet review of [Pen05].
[39] [**?**, p. 313].
[40] J. Kacirk, *Forgotten English*, p. 86.

Roping the Earth

I.6.1.1. We begin with a very simple example where many people's intuition is badly wrong. Let us suppose the earth is a perfect sphere, and that it is physically possible to stretch a rope along its surface completely around the equator (such a rope would be roughly 40,000 km long). Now suppose we instead raise the rope to a uniform height of 1m above the surface. How much additional rope will be needed?

The intuition of a lot of people is that many extra kilometers of rope will be necessary, and that the exact amount depends on the circumference of the earth. But this is drastically wrong, as some simple algebra shows. Let r be the radius of the earth. The length of rope needed at the surface is the circumference of a circle of radius r, or $2\pi r$. If we raise the rope to a uniform height h above the surface, the rope length needed will be the circumference of a circle of radius $r + h$, $2\pi(r + h) = 2\pi r + 2\pi h$, so the additional rope needed will have length $2\pi h$, or in the specific problem $2\pi \approx 6.28$ m. Note that the answer does not depend on r!

Actually, there are some subtleties in this problem (besides the physical ones) which we have hidden or glossed over; see I.6.1.5..

Average Speed

I.6.1.2. A man drives to a city 100 km away to deliver something, then immediately drives back. On the way to the city, he drives at a constant speed of 100 km per hour, but on the return trip he is slowed by heavy traffic and can only drive at a constant speed of 50 km per hour. What is his average speed for the trip?

Many people would immediately say 75 km per hour, which is the average of 50 and 100. But this is wrong. It took him one hour to get to the city and two hours to get back, so the trip took three hours in all. He drove 200 km in 3 hours, an average speed of $66\frac{2}{3}$ km per hour.

Why is the knee-jerk answer wrong? Average speed is averaged over *time*, not distance. He did drive the same distance at 100 as at 50, but he drove 50 for twice as long as he drove 100. The correct answer is no harder to calculate than the wrong one, but requires the problem to be thought about in the right way.

In finding the right way to think about something, it is often useful to think about an extreme case of the situation. For this problem, suppose a man tries to get to a town 40 km away. He starts driving 100 km per hour. After 12 minutes he is halfway there, but his car breaks down and he must walk the rest of the way at 4 km per hour, which takes five hours. Is his average speed really 52 km per hour (the average of 4 and 100)?

Here is a variation. A man can row a boat 6 km per hour in still water. He plans to row the boat to a point 12 km upstream in a river with a small current, then back. He figures the current will help him on the return trip as much as it will slow him down upstream, so his average speed will be 6 km per hour and the trip should take 4 hours, irrespective of the current. But this is wrong. For example, if the current is 2 km per hour, he will row upstream at a ground speed of 4 km per hour, so the trip up will take 3 hours, and a ground speed of 8 km per hour back, taking 1.5 hours, a total of 4.5 hours. And if the current is 5 km per hour, it will take him 12 hours just to get upstream! In fact, if the current is x km per hour, it will take him

$$L(x) = \frac{12}{6 - x} + \frac{12}{6 + x}$$

hours. And if $x \geq 6$, he will not be able to get upstream at all. In this version, the right way to think of the problem is to vary the current x, and the extreme case to consider is what happens as x approaches 6. This viewpoint makes it obvious that the length of time cannot be independent of x.

Areas in the Plane

I.6.1.3. The next example concerns areas of sets in the plane. Consider the following two statements:

STATEMENT 1. If R is a (solid) rectangle with length x and width y, then the area $A(R)$ is xy.

STATEMENT 2. If a rectangle R is subdivided into a finite number $R_1, \ldots, R_n$ of nonoverlapping subrectangles (nonoverlapping means two of the subrectangles can have only boundary points in common), then

$$A(R) = \sum_{k=1}^{n} A(R_k) \ .$$

(See Figure I.3; the area of the black rectangle is the sum of the areas of the green rectangles.)

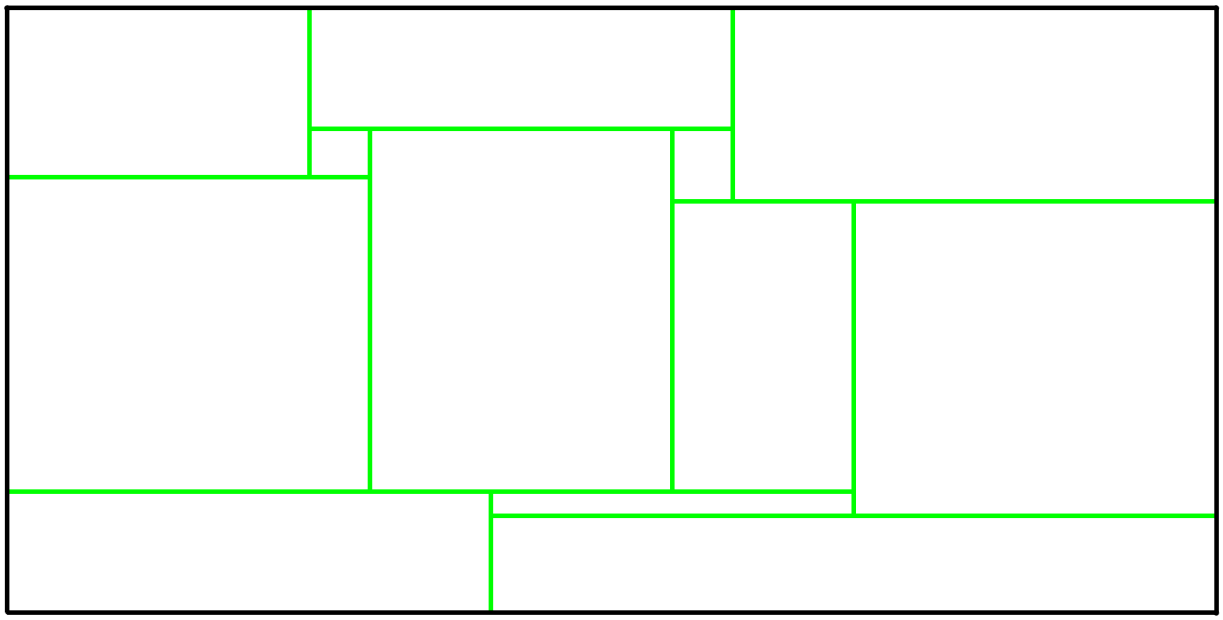


Figure I.3: Subdivision of a Rectangle

Let us set Statement 1 aside for the moment, and consider Statement 2. The question we want to discuss is: Is Statement 2 obvious?

Many people would say that Statement 2 is "geometrically obvious." And not just people untrained in mathematics: for example, on p. 7 (the first page of actual text) of the otherwise carefully written book [SG77], it is asserted without comment that Statement 2, and even its higher-dimensional version, is true, and this assertion is then used as the basis for the whole development of Riemann integration in Euclidean space. The same assertion is made in [KF75, p. 255] (where it is called "obvious") and in [Sak64]; essentially the same statement (including the "obvious" assertion) occurs in [Bog07, p. 180]. (Is this an Eastern European thing? Note, however, that [Nat61] does not have this problem – see p. 64-65 of Volume 2.)

The real issue is whether Statement 1 is a definition or a proposition. If it is a proposition, then what is the definition of the area of a rectangle, or other plane figure? Are we using some intuitive "definition" of area, along with its accompanying list of "natural" properties, among which would be some general version of Statement 2? What justification do we have that there is a notion at all of area having the geometric properties we would expect? (There is, but it takes some work to show it, even for polygons; cf. [Har00, §23]. For more complicated regions, even a disk, it is still more difficult, requiring limits.) See I.3.1.10. for further discussion. We can bypass these difficulties by taking Statement 1 as the *definition* of the area of a rectangle; but then we are not justified in assuming that "area" so defined has the expected geometric properties. In particular, Statement 2 is then not obvious, and requires proof. The proof, while not difficult when the right "trick" is seen, is nontrivial: see IX.3.1.6.. Note that it cannot simply be proved by induction on n: the cases $n = 3$ and $n = 4$ can be easily reduced to the case $n = 2$, which is obvious, but even the case $n = 5$ is not entirely obvious and cannot be simply reduced to applications of the smaller cases, since a rectangle can be divided into five subrectangles in such a way that the union of no two, three, or four subrectangles is a rectangle.

Some references such as [Wyl64] try to define area axiomatically; Statements 1 and 2 are then among the axioms, or at least theorems, of the theory. But this approach does not eliminate the problem: the difficulty then simply becomes showing that the axioms are consistent, a question not really discussed in [Wyl64]. For a baby form of this, consider the following statement:

Statement 3: There is a notion of numerical area A defined for rectangles in the plane satisfying

> A is monotone, i.e. if $R_1 \subseteq R_2$, then $A(R_1) \leq A(R_2)$.
>
> Congruent rectangles have the same area.
>
> A satisfies Statement 2.
>
> The area of the unit square is 1.

The proof of IX.3.1.6. shows that Statement 1, taken as the definition of area, implies Statement 3 (the rest of Statement 3 is immediate). Conversely, Statement 3 easily implies Statement 1 for rectangles with rational sides (a weaker version of Statement 2 for grid subdivisions suffices), and thus for general rectangles by exhaustion; thus Statement 1 and Statement 3 are equivalent. However, it is questionable whether either statement is "obvious." And even if we take Statement 1 as a definition, we only immediately get area for rectangles and it takes considerable work to get it for more general regions, along with its "natural" properties.

The semantic problem is use of the word "area" without specification of its meaning; it is suggestive to the reader of an intuitive interpretation which was perhaps reinforced by informal discussions of the subject in school, even though the informal or intuitive interpretation may not be appropriate in careful mathematics.

I.6.1.4. Anyone who thinks it is obvious that when a figure is cut up and rearranged, the area stays the same, should try the following example (photo from the website http://www.moillusions.com/2006/03/impossible-triangle-illusion-no2.html; the first such example was apparently found by the American puzzlemaster Sam Loyd):

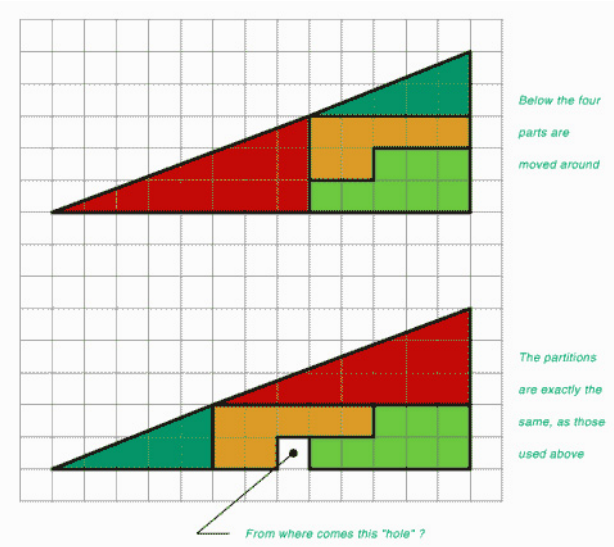


Figure I.4: Where did the extra square come from?

A triangle of area $\frac{65}{2}$ is cut up and rearranged into a figure of area $\frac{63}{2}$ (or is it?) For a great visual version, see https://www.youtube.com/watch?v=3PszMaZ5Ipk.

A much more sophisticated case (which is a real counterexample, not just an illusion) is the Banach-Tarski Paradox (III.12.): a ball in $\mathbb{R}^3$ can be dissected into five pieces which can be rearranged to form two balls of the same size as the original. The paradox is resolved by the fact that the pieces are *nonmeasurable*: they are so bizarre that no "volume" can be assigned to them. (In fact, area in $\mathbb{R}^2$, when defined, *is* preserved under arbitrary finite dissection and rearrangement, but this is quite nontrivial and nonobvious and, as the Banach-Tarski Paradox shows, does not hold for either volume or surface area in $\mathbb{R}^3$. See e.g. [Wag93].)

Circumference of a Circle and the Number π

I.6.1.5. Area, at least for a rectangle, is easy and elementary to define (or calculate). But here is an example where making a careful definition of an intuitive concept is quite complicated.

How do we define the number π? It is usually defined as the ratio of the circumference of a circle to its diameter (for this to make sense, it must be shown that the ratio is always the same for all circles, which is not really so obvious; in fact, it is false in noneuclidean geometry!) Slightly more precisely:

Definition. The number π is the circumference of a circle of diameter 1.

Several things must be made precise before this can be accepted as a real definition: the terms *circumference*, *circle*, and *diameter* (and I suppose even the number 1) must be defined. To get to the heart of the matter, let us tighten up the definition to

DEFINITION. The number π is half the circumference of the circle

$$\{(x, y) : x^2 + y^2 = 1\} \subseteq \mathbb{R}^2 \ .$$

The real issue is the definition of the term *circumference* (or *perimeter*). Although people have a (fairly) good intuition as to what this term means, and one could be led to say the meaning is "obvious", it is not at all easy to give a precise definition, and the meaning of the word is finessed throughout precalculus mathematics (e.g. to paraphrase MANDELBROT, what is the circumference of Britain?) Using calculus, it can be defined as a special case of arc length, which can itself be defined as the supremum of a set of sums (and in good cases exactly calculated as a definite integral). It is then not entirely obvious that a circle even has a (finite) circumference. This was more or less shown already in ancient times by a limit argument, and the numerical value of π was calculated to increasing accuracy over the centuries. But making good precise sense of the circumference of a circle, and hence the number π, is a quite subtle matter.

I.6.1.6. To elaborate on the difficulties, let us consider the traditional (at least since ARCHIMEDES) way circumference is defined and estimated. Fix a circle. We consider inscribed and circumscribed regular polygons. For a fixed n, the perimeter p_n of all inscribed regular n-gons is the same by symmetry, and similarly for the perimeter P_n of circumscribed regular n-gons, and obviously $p_n < P_n$. It is not obvious that p_n increases monotonically as n increases (although this can be shown), but it is obvious, and easily proved, that $p_n < p_m$ if n divides m. It is less obvious, but not difficult to prove, that $P_m < P_n$ if n divides m. Thus, if n divides m, $p_n < p_m < P_m < P_n$. The circumference p of the circle is generally defined to be the supremum (or limit) of the p_n as $n \to \infty$. It is not hard to show that $P_n - p_n \to 0$ as $n \to \infty$ (cf. I.6.1.21.(v)), so p is also the infimum (or limit) of the P_n. By explicitly calculating p_n and P_n for large n, decent upper and lower estimates for p can be obtained.

Incidentally, some references assert that it is "obvious" that $p < P_n$ for any n. I'm sorry, but it is not at all obvious to me. Intuitive arguments for this fact seem to be based on some type of belief in the continuity of arc length (I.6.1.16.). It does follow (using that $P_m < P_n$ if n divides m) from the lower semicontinuity of arc length (VI.4.1.30.), but I don't think that is intuitively obvious. It is intuitively plausible that $p_n < p$ for all n, based on the principle that "a straight line is the shortest distance between two points," and it is not unreasonable to define p as $\sup p_n$. The above arguments which show that $\inf P_n$ is equal to $\sup p_n$ imply that $\inf P_n = p$ and hence $p < P_n$ for all n (this follows more easily directly from the fact that $p_m < P_n$ if n divides m, which implies that P_n is an upper bound for the p_m for any n).

So what is the difficulty with this argument? For one thing, many seemingly small details must be made precise; all the statements above need to be proved, and along the way some terms need to be carefully defined, which is sometimes not as easy as one might think. In addition, one needs to show:

(i) There exist inscribed and circumscribed regular n-gons. Note that this cannot be done simply by using angle measure since angle measure cannot be carefully defined until concepts from trigonometry are developed, which already requires knowledge about the perimeter of a circle and its subarcs. Instead, either geometric or analytic arguments must be used. (Certain regular polygons inscribed in a circle like a square or hexagon can be constructed by straightedge and compass, and their sides repeatedly bisected to obtain regular polygons with arbitrarily large numbers of sides, but not all regular polygons can be constructed this way, although enough can to make the argument work with additional justification.)

(ii) The perimeter of an inscribed polygon must be defined. In order to do this, it must be said what is meant by a finite set of points on a circle being listed in cyclic order. It is "obvious" geometrically

what this means, but not so simple to explicitly state in words (if angle measure cannot be used). Two (equivalent) ways are: (1) interiors of edges between two pairs of consecutive points never intersect (equivalently, all of the other points lie on the same side of the line through any two consecutive points) or (2) the sum of the distances between consecutive points (the perimeter) is minimum among all possible orderings of the given points. (The second characterization may be preferable, since it is not obvious that a given finite subset of the circle has any ordering satisfying (1).)

(iii) The Completeness Property of $\mathbb{R}$ is needed to insure that the supremum or limit of the p_n exists.

In addition, there is another technical difficulty. The arc length of a circle, as defined in calculus, would be the supremum of the perimeters of all inscribed polygons, regular or not. It is not obvious that the same supremum is obtained for all inscribed polygons as for inscribed regular polygons, i.e. that there could not be a nonregular inscribed polygon whose perimeter is greater than the perimeters of all inscribed regular polygons. A way of showing that the two suprema are the same is to fix a point a on the circle, and show that the vertices of regular inscribed polygons with one vertex at a are dense in the circle; by continuity it suffices to consider only inscribed polygons with vertices in this set, and any finite number of points in this set are among the vertices of a single inscribed regular polygon. The density of this set of vertices, while "geometrically obvious", takes some work to prove. As an alternate approach, it can be shown using a compactness argument (based on the Completeness Property) that for any fixed n, there is an inscribed n-gon of maximum perimeter, and it is fairly easy to show that it must be a regular n-gon. Or the arguments of [Pól90a, Chapter X] can be adapted to give an "elementary" proof of this fact, although a careful and complete argument involves much more than is discussed in [Pól90a] (cf. I.6.1.7.).

All in all, it is a quite involved matter to make this approach completely rigorous.

I.6.1.7. There are a number of closely related arguments in [Pól90a]. Let us confine the discussion to the topics in Chapters VIII and X, where max-min problems and the Isoperimetric Problem are considered. The arguments in [Pól90a] are clever and elementary, and at first reading seem convincing, but there are serious logical gaps in the arguments which the author seemingly tries to obscure by his choice of language. Among these are:

(i) It is often implicitly assumed, or even explicitly asserted without explanation, that the desired maximum or minimum exists. To justify the existence requires at least an application of the Completeness Property of $\mathbb{R}$, and in some cases additional continuity or compactness arguments are needed.

This point of justifying existence of a maximum/minimum is not just a pedantic technicality: there are seemingly similar problems, e.g. the Dirichlet principle (), where it turned out that no max/min exists in general, much to the surprise of mathematicians who assumed the contrary as "obvious."

(ii) In some arguments, it is assumed without discussion that there is a notion of area of regions of the plane bounded by various types of curves, which is additive on "nonoverlapping" regions and invariant under rigid motions, leading to more complicated versions of the difficulties discussed in I.6.1.3..

(iii) In several arguments, use is made of "tangent lines" to various curves without defining what this means, much less discussing whether such lines exist and what properties they must have. When the curve is a circle, tangent lines can be defined and found without any explicit use of calculus (but there is still some nontrivial work to be done for a careful treatment), but for more general curves calculus cannot be avoided.

(iv) Several other similar difficulties are dismissed using a comment like "remembering some elementary geometry . . . " and some of these are theorems that take some work to prove or even to carefully state.

See [Oss78], [BZ88], or [Ber03, p. 26-29] for much more careful and complete discussions of the Isoperimetric Inequality.

Is Arc Length an Intrinsic Concept?

I.6.1.8. Continuing the discussion of I.3.1.10. and I.6.1.3., as alluded to above, there is a basic question whether there is an intrinsic notion of arc length for (say smooth) curves, with some "natural" properties, or whether arc length must be defined and the properties proved. We now believe that it must be defined, but all reasonable ways of defining it give the same numerical answer, so whether it is "intrinsic" is a matter of opinion. See VI.4.1. for a full discussion.

I.6.1.9. There is even an unexpected subtlety in defining the arc length of a line segment, which is the first step in defining arc length for curves (this is on top of the question of whether line segments themselves are well defined; cf. I.3.1.3. and I.6.1.48.). For a modern version of the problem which supposedly confounded and upset the Pythagoreans in ancient Greece (modern scholars of ancient Greece are increasingly skeptical about this, cf. [Fow99]), the length (arc length) of the straight line segment in the plane between (0,0) and (1,1) is $\sqrt{2}$. Thus, if our number system were just the rational numbers $\mathbb{Q}$, this segment would not have a numerical length (cf. I.6.5.8.). As geometry developed in Greece, a distinction was made between "magnitudes" and "numbers" (which for them were at most rational); many line segments did not have a numerical length for them. But it was shown by EUDOXUS, cf. EUCLID'S *Elements*, that (in modern language) the set of "constructible" (by straightedge and compass) magnitudes is the positive part of an Archimedean ordered field in which square roots of positive elements can be taken. This field of "constructible" numbers has a natural embedding as a subfield of the real numbers, which can be described algebraically.

I.6.1.10. There is much greater difficulty in defining surface area for curved surfaces; see () for a discussion. This difficulty was already recognized by ARCHIMEDES. Thus it is somewhat more questionable whether surface area of (smooth) surfaces is an intrinsic concept.

> A baseball player ordered a pizza. When asked whether he wanted it cut in six pieces or eight, he thought a minute and replied, "Better make it six. I can't eat eight."

The Area of a Circle

I.6.1.11. The circumference of a circle of radius r is $2\pi r$, and the area is πr^2. Even a casual observer would notice a similarity between these two formulas, and a calculus student would observe that the circumference is the derivative of the area (this is of course no accident, and is the basis for a quick calculus proof of the area formula). Explaining to a layman why the formulas are similar, and why one follows from the other, is a challenge, though.[41] Since the circumference formula is essentially just true by the definition of π, the task amounts to proving the area formula from the circumference formula.

I.6.1.12. The area formula requires calculus in some form for its proof, since some type of limit is necessarily involved. (As explained earlier, limits are involved in even making a careful *definition* of either area or circumference.) ARCHIMEDES described a beautiful argument which seems to make the area formula "obvious" once the informal limiting operations are made rigorous. This argument has recently been again popularized by S. STROGATZ in his New York Times series (http://opinionator.blogs.nytimes.com/2010/04/04/take-it-to-the-limit/), from which the following pictures are taken.

I.6.1.13. The argument goes like this. Divide the circle (actually disk) into a large number of congruent wedges, say 2^n, and rearrange them into a strip:

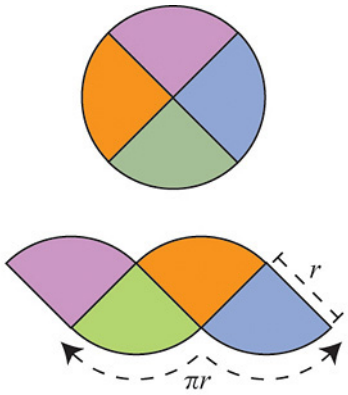


Figure I.5: Rearrangement with 4 wedges

For any n, the wavy strip has the same area as the circle, the top and bottom have arc length πr, and the sides are line segments of length r.

I.6.1.14. As $n \to \infty$, the top and bottom wavy curves flatten out into straight line segments, and the end segments become more and more vertical; thus in the limit the figure becomes a rectangle.

This rectangle has area πr^2, so the area of the circle is also πr^2.

[41] Teachers don't always try. PETR BECKMANN [Bec71, p. 17] says: "Most of us first learned this formula [for the area of a circle] in school with the justification that the teacher said so, take it or leave it, but you better take it and learn it by heart; the formula is, in fact, an example of the brutality with which mathematics is often taught to the innocent."

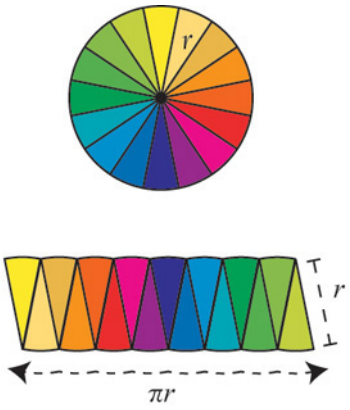


Figure I.6: Rearrangement with 16 wedges

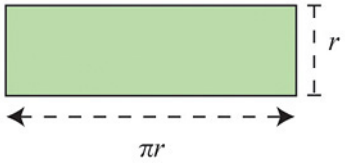


Figure I.7: Limit arrangement

I.6.1.15. This argument is sometimes represented as a "proof" of the area formula. But it is not (yet) since it has some gaps, one of which is quite serious. The obvious gaps are that

(i) The limiting operations must be made precise.

(ii) It must be shown that area is well defined for the types of regions considered, and that it has the subdivision property discussed in I.6.1.3..

(Actually, (ii) may not be obvious to most observers!) However, there are two more gaps which may not be as obvious:

(iii) It must be shown that such a decomposition into wedges is possible, discussed in I.6.1.6.(i).

(iv) It must be shown that both area and arc length are preserved under the limiting operations of (i).

I.6.1.16. The preservation of area under the limiting operation of passing from the wavy strips to the limiting rectangle takes some work to prove rigorously, but is at least true in great generality, and it is possibly justifiable to regard the proof as simply a technical justification for an "obvious" geometric assertion.

However, the arc length assertion is definitely not "obvious." It is not even true in general. Actually, to even formulate it carefully it must be said what the limiting operation is. One reasonable way to state it formally is: regard the bottom curve as the graph of a function f_n, and the limiting curve the graph of a

(constant) function f. Then $f_n \to f$ uniformly. (One could alternately think of them as parametrized curves and consider uniform convergence of the parametrizing functions.) However, arc length is not preserved under such a limiting operation (it is only lower semicontinuous (VI.4.1.30.), not continuous):

I.6.1.17. EXAMPLE. On the interval $[0, \pi]$, define a sequence of functions by

$$f_n(t) = 2^{-n} \sin(2^n t)$$

and a limit function $f = 0$. Then $f_n \to f$ uniformly on $[0, \pi]$, i.e. the curves $y = f_n(x)$ approach the straight line from 0 to π on the x-axis (in the uniform sense). However, the arc lengths of the curves are all the same, as can be seen easily from a geometric argument (the graph of f_{n+1} consists of two pieces, each similar to the graph of f_n scaled by a factor of $\frac{1}{2}$), and this arc length can be computed by calculus () to be

$$\int_0^\pi \sqrt{1 + [f_1'(t)]^2}\, dt = \int_0^\pi \sqrt{1 + \cos^2(2t)}\, dt \approx 3.82$$

which is not the same as the arc length $\pi \approx 3.14$ of the limiting curve.

One can do much worse: set

$$g_n(t) = 2^{-n} \sin(4^n t)$$

on $[0, \pi]$. Then $g_n \to 0$ uniformly, so the graphs again approach the straight line segment from 0 to π, but the arc lengths of the graphs of the g_n go to $+\infty$.

I.6.1.18. A similar example is given in [LN00] ("*le trou normand*"), where it is regarded as a "paradox." Someone with a reasonable level of mathematical sophistication (unlike the authors of that book) would realize that this is not a paradox at all: the problem is simply that if $f_n \to f$ uniformly, then it is generally not true that $f_n' \to f'$ in any ordinary sense, even if the derivatives all exist, which need not be the case. (Actually, under mild assumptions the derivatives do indeed converge to something in the distributional sense, but the limit is a distribution and not a function in general, and this notion is rather sophisticated, cf. XXII.2.1.63..)

In fact, with a little more reflection this disconnect should be expected. The absolute value of the derivative of a function says how *rapidly* the function is changing, not how *much*. A function can change rapidly, i.e. briefly have a large derivative, even if it is almost constant, if the changes are small enough. The derivative can even be large (in absolute value) most of the time, if it oscillates rapidly between positive and negative (a straight line can be uniformly approximated arbitrarily closely by curves that wiggle a lot). Another way of saying this is that the derivative is an unbounded operator with respect to the supremum norm ().

I.6.1.19. Here is perhaps the simplest example of this phenomenon (I am indebted to S. STROGATZ for pointing it out). The diagonal of the unit square in $\mathbb{R}^2$ can be arbitrarily approximated uniformly by a "staircase" curve consisting of short alternating horizontal and vertical segments. Such a staircase curve clearly has arc length 2. So why doesn't the diagonal also have length 2? A circle, say of radius 1, can also be uniformly approximated by a staircase curve in each quadrant, which has a total length of 8. So is $\pi = 4$?

I.6.1.20. The nontrivial extra feature in the case of ARCHIMEDES' argument (which incidentally was not his standard proof, which was a rigorous argument by double *reductio ad absurdum*, cf. ()) is that if f_n is the function whose graph is the lower wavy curve at the n'th stage, then not only does $f_n \to 0$ uniformly, but the derivatives of the f_n (actually the one-sided derivatives since f_n is not differentiable everywhere) also converge uniformly to 0, which is sufficient to conclude continuity of the arc lengths. To make the argument into a rigorous proof, some version of this additional argument is necessary.

I.6.1.21. There is an alternate way to phrase the argument which reduces (but does not eliminate) the technicalities, once the previous arguments about circumference have been done, necessary in any event for a full logical treatment. Inscribe a regular polygon with n sides in a circle of radius r, and connect each vertex to the center of the circle by a line segment (see Figure I.6, from http://www.amsi.org.au/teacher_modules/the_circle.html).

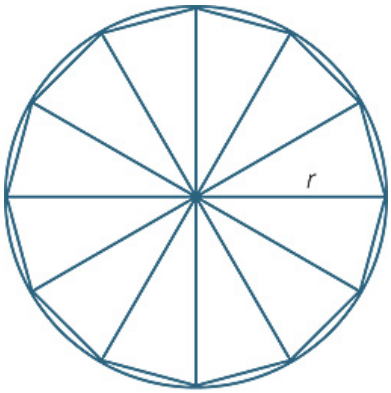


Figure I.8: Inscribed regular polygon with $n = 12$, divided into triangles.

This divides the polygon into n congruent isosceles triangles of short side b_n and height h_n. Each of these triangles thus has area $\frac{1}{2}b_n h_n$. We have that $nb_n = p_n$, the perimeter of the polygon, so the area of the polygon is

$$A_n = n \cdot \frac{1}{2} b_n h_n = \frac{1}{2} p_n h_n \ .$$

As $n \to \infty$, $p_n \to 2\pi r$ and $h_n \to r$, so the area A of the circle is

$$A = \lim_{n\to\infty} A_n = \lim_{n\to\infty} \frac{1}{2} p_n h_n = \frac{1}{2} \cdot 2\pi r \cdot r = \pi r^2 \ .$$

For this argument, we need to verify:

(i) Such a decomposition into triangles is possible, discussed in I.6.1.6.(i).

(ii) The area of the polygon is indeed the sum of the areas of the triangles, i.e. area is additive.

(iii) The area of a triangle with base b and height h is $\frac{1}{2}bh$, which has an elementary proof by dissection once the area of a rectangle is known (in some approaches such as in [Har00], areas of triangles are the primitive data and no proof of this is necessary).

(iv) $p_n \to 2\pi r$, the circumference argument already done.

(v) $h_n \to r$, which takes some work to do carefully but is relatively easy. Indeed, by the Pythagorean Theorem,

$$0 < r - h_n = \sqrt{\left(\frac{p_{2n}}{2n}\right)^2 - \left(\frac{p_n}{2n}\right)^2} = \sqrt{\frac{(p_{2n}+p_n)(p_{2n}-p_n)}{4n^2}} < \sqrt{\frac{4\pi r(2\pi r - p_n)}{4n^2}} < \frac{\sqrt{4\pi r(2\pi r)}}{2n} = \frac{\pi r\sqrt{2}}{n}$$

which can be made arbitrarily small by choosing n large enough.

(vi) $A_n \to A$, one of the same difficulties as in ARCHIMEDES' argument, but easier since it is essentially a monotone argument, traditionally called *exhaustion* (cf. XVII.1.2.7.).

(vii) The elementary limit fact (IV.1.3.21.(ii)) that a limit of a product is the product of the limits.

I.6.1.22. If n is even, the triangles can be rearranged into a true parallelogram of length $p_n/2$ and slant height r sitting inside the wavy parallelogram of Figure I.6. As $n \to \infty$, by (iv) the length of the base approaches πr, so the arc length difficulty in ARCHIMEDES' argument is avoided (or, really, converted into the previous circumference argument). But the other difficulties remain.

Equal Division

I.6.1.23. On the subject of area, here is a problem which is described in some popular books as having an "obvious" solution. Let E be a (bounded) region in the plane, and L a line in the plane. Then there is a line parallel to L which cuts E into two regions of equal area.

The "obvious" solution is to begin with a line parallel to L with E entirely on one side. Then move this line "continuously" toward and past E until it lies entirely on the other side of E. Since the area on the back side of the line begins at 0 and increases continuously to the area $A(E)$ of E, there must have been some point where this area is exactly $\frac{1}{2}A(E)$.

I.6.1.24. What is the difficulty? Some of these references gloss over this question entirely, suggesting there is none at all. But some acknowledge that the result depends on the Intermediate Value Theorem, which is a nontrivial (but perhaps geometrically "obvious") fact. And there is another difficulty which never seems to be mentioned: if $A(t)$ is the area of the part E_t of E behind the line moved a distance t, why is $A(t)$ a continuous function of t? It is, but this fact is essentially the upward and downward continuity of Lebesgue measure in the plane (), again a nontrivial fact, although if E is bounded, continuity of $A(t)$ can be proved more easily by bounding $E_t \setminus E_s$ in a rectangle of fixed length and small width if s is slightly less than t. These difficulties are, of course, in addition to the question of whether $A(E)$ and the $A(t)$ are even well defined and whether $A(E) = A(E_t) + A(E \setminus E_t)$ for all t, discussed earlier.

I.6.1.25. Some popular books go on to discuss the next version (the two-dimensional version of the *ham sandwich theorem* ()): if D and E are (bounded) regions of the plane, then there is a line which simultaneously cuts both D and E into regions of equal area. The "obvious" argument goes as follows: by the first part of the argument, for each θ there is a line $L(\theta)$ making an angle θ with the horizontal which cuts E into two equal parts. If θ is varied from 0 to π, the line $L(\theta)$ passes across D as before, and the division is reversed for $\theta = \pi$ from the division at $\theta = 0$, so there must be a θ for which the division of D is equal.

There is an additional difficulty in this version, rarely if ever identified in popular references: the line $L(\theta)$ is not necessarily unique, e.g. if E is not connected, so an $L(\theta)$ must be chosen for each θ. It is not obvious, and is rather tricky to prove even in the case where the $L(\theta)$ are unique, that the lines $L(\theta)$ can be chosen to vary "continuously" with theta (as θ varies, the line does not simply rotate, but the point of intersection usually changes too). However, it is easier to show (although still a little tricky to prove rigorously) that if D and E are bounded, the area of the part of D on the one side of $L(\theta)$ can be chosen to be a continuous function of θ.

Sequences and Pattern Recognition

I.6.1.26. What is the next term in the sequence $(0, 1, 2, \dots)$? Most people would immediately answer "3". The "simplest" formula for the n'th term x_n giving these values for $n \leq 3$ would be $x_n = n - 1$, so this is the "obvious" answer. However, there are many other possible formulas, and the "correct" answer depends on guessing what the poser of the problem had in mind. For example, if r is any real number $\neq 3$, it is a simple exercise (solving a system of four linear equations in four unknowns; cf. V.3.11.17. for a more efficient way) to find a cubic polynomial f such that $f(1) = 0$, $f(2) = 1$, $f(3) = 2$, and $f(4) = r$.

I.6.1.27. There are other possible rules: for example, x_n could consist of $n-1$ followed by $n-1$ factorial signs (I am indebted to VAUGHAN JONES for this example.) Then

$$x_1 = 0$$

$$x_2 = 1! = 1$$

$$x_3 = 2!![= (2!)!] = 2$$

$$x_4 = 3!!! = 720! \approx 2.6 \times 10^{1746}.$$

I.6.1.28. Here is another example. It appears different since it is not stated in purely mathematical terms, but in fact it is essentially identical. Suppose at 6 AM the temperature is 0°, at 9 AM it is 10°, and at noon it is 20°. What will the temperature be at 3 PM? 6 PM? 9 PM? Does anyone believe the temperature will be 50° at 9 PM? I suppose it's not impossible, but it's certainly not likely!

Actually, the only difference between this problem and the previous one is that there is *more* information: how do you know in the first problem that the terms of the sequence aren't the temperatures (measured in 10° units) at three-hour intervals starting at 6 AM a certain day and location? Or maybe the n'th term is the number of hits a certain baseball player gets in his first n at bats of the season, or the number of Heads in the first n tosses of a coin, or .... How do we know there is any rule or pattern at all - why couldn't it just be a random sequence?

In the second problem meteorologists could predict with considerable accuracy what these temperatures will be, but they would need more information than is stated in the problem. However, in none of the problems can anyone figure out definitively what the next term will be without more information than is given in the problem (and in the coin tossing case no one could determine the next term even with more information).

I.6.1.29. The situation seems to be similar to a television comedy sketch which once appeared on *Saturday Night Live*, about a game show. The final question was: "I'm thinking of a color. What is it?"

I.6.1.30. But is it really similar? Being able to recognize a pattern in a sequence is an important mathematical skill. The above example is too easy to properly illustrate the point. A better example is the following sequence:

$$(2, 5, 10, 17, \dots)$$

It is a bit more challenging, although not difficult, to come up with a pattern here. In fact, there are several natural ways to generate a sequence beginning with these terms, and it is not obvious what rule is "simplest" in this case.

I.6.1.31. This sequence appeared on a TerraNova standardized test for high school students. Unfortunately, it was phrased as a multiple choice question, asking for the next term; three of the choices for the answer were 24, 26, and 28.

The solution, as described in the *Reno Gazette-Journal*, is that the difference between successive terms increases by 2 at each step. The "correct" answer is thus 26.

I.6.1.32. Equivalently, this rule can be phrased as an algebraic formula giving the terms rather than a way of obtaining each new term from the last one. Use the function $f(x) = x^2 + 1$; then the n'th term of the sequence is the value of $f(x)$ for $x = n$. This formula gives the same fifth term, 26, as the published rule (in fact, all the terms of the sequences are the same.)

I.6.1.33. One can give a completely different rule. Let the n'th term be the sum of the first n prime numbers. We then obtain the sequence $(2, 5, 10, 17, 28, 41, 58, \dots)$, so the fifth term is one of the "incorrect" answers.

I.6.1.34. One can, of course, play the polynomial game here too. For example, let

$$f(x) = x^4 - 10x^3 + 36x^2 - 50x + 25.$$

We then have $f(1) = 2$, $f(2) = 5$, $f(3) = 10$, $f(4) = 17$, and $f(5) = 50$. For another example,

$$g(x) = -\frac{1}{12}x^4 + \frac{5}{6}x^3 - \frac{23}{12}x^2 + \frac{25}{6}x - 1$$

has $g(1) = 2$, $g(2) = 5$, $g(3) = 10$, $g(4) = 17$, and $g(5) = 24$, another of the "incorrect" answers from the problem.

I.6.1.35. Most mathematicians are very critical of problems like this one. The objection is not to the problem itself, but the way it is phrased: as written, it is a psychology problem, not a mathematics problem; it is almost like asking, "In what country are the Alps located?" A much preferable way to phrase the problem is: "Describe a rule or pattern giving a sequence of numbers whose first four terms are 2,5,10,17, and find the fifth term in the sequence according to your rule." Unfortunately, there does not seem to be any reasonable way to phrase this type of problem as a multiple-choice question; the only acceptable multiple-choice version would be one with a choice (the correct answer!) "Cannot be determined from the information given."

Questions like this are really very much like questions where letters must be unscrambled to form a word, if only one answer is "right" – see figure I.9, and maybe even like figure I.10.[42]

I.6.1.36. The widespread acceptance of the legitimacy of this type of question is symptomatic of a general human problem: there is often more than one explanation for a phenomenon, but an individual or community often settles on one explanation as the "right" one and rejects any alternatives or even the possibility of an alternative, even sometimes in light of new evidence. Scientific theories which contradict,

[42]The company which produced the instructional software we use in calculus classes was bought out by another company. To use the software, I needed an account with the new company. When I tried to create one, I found that one had already been set up automatically, without anyone telling me. To log in, I needed the password which no one had told me either. When I clicked on "forgot password," I was required to give the answer to my "Secret Question." It would not tell me what the question was, presumably because it was secret.

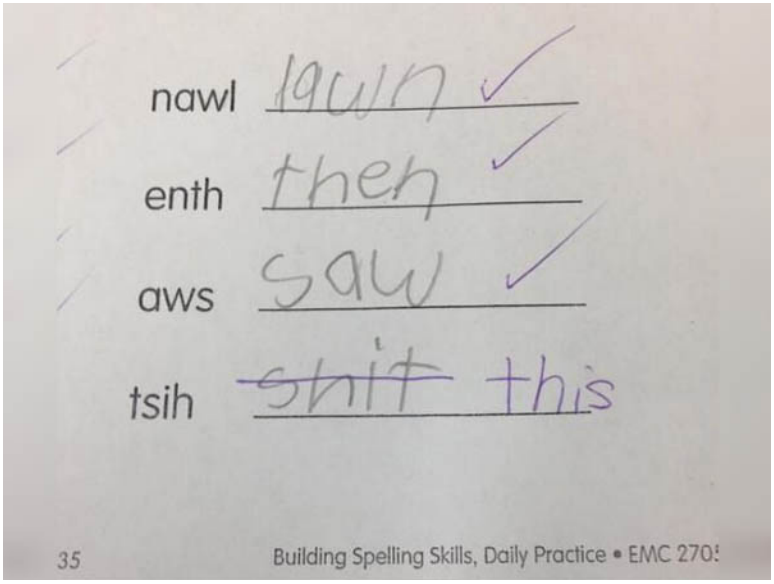


Figure I.9: Unscramble the letters.

or appear to contradict, established religious beliefs are a notorious example. In the sequence problem case, people untrained in mathematics may have the excuse that they think such ambiguity cannot happen in mathematics and that any mathematically phrased problem has a definitive answer; mathematicians know better!

> "There is a lesson in plausible reasoning to be learned from these problems [on polyominoes]. Given certain basic data, we labor long and hard to fit them into a pattern. Having succeeded, we believe the pattern to be the only one that 'fits the facts'; indeed, that the data are merely manifestations of the beautiful, comprehensive whole. Such reasoning has been used repeatedly in religion, in politics, and even in science. The pentominoes illustrate that many different patterns may be possible from the same 'data,' all equally valid, and the nature of the pattern we end up with is determined more by the shape we are looking for than by the data at hand. It is also possible that for certain data, no pattern of the type we are conditioned to seek may exist."
>
> *Solomon W. Golomb*[43]

This may even be inevitable:

> "Something like a paradigm is prerequisite to perception itself. What a man sees depends both upon what he looks at and also upon what his previous visual-conceptual experience has taught him to see. In the absence of such training there can only be, in William James's phrase, 'a bloomin' buzzin' confusion.' "
>
> *Thomas Kuhn*[44]

[43] [Gol54].
[44] *The Structure of Scientific Revolutions.*

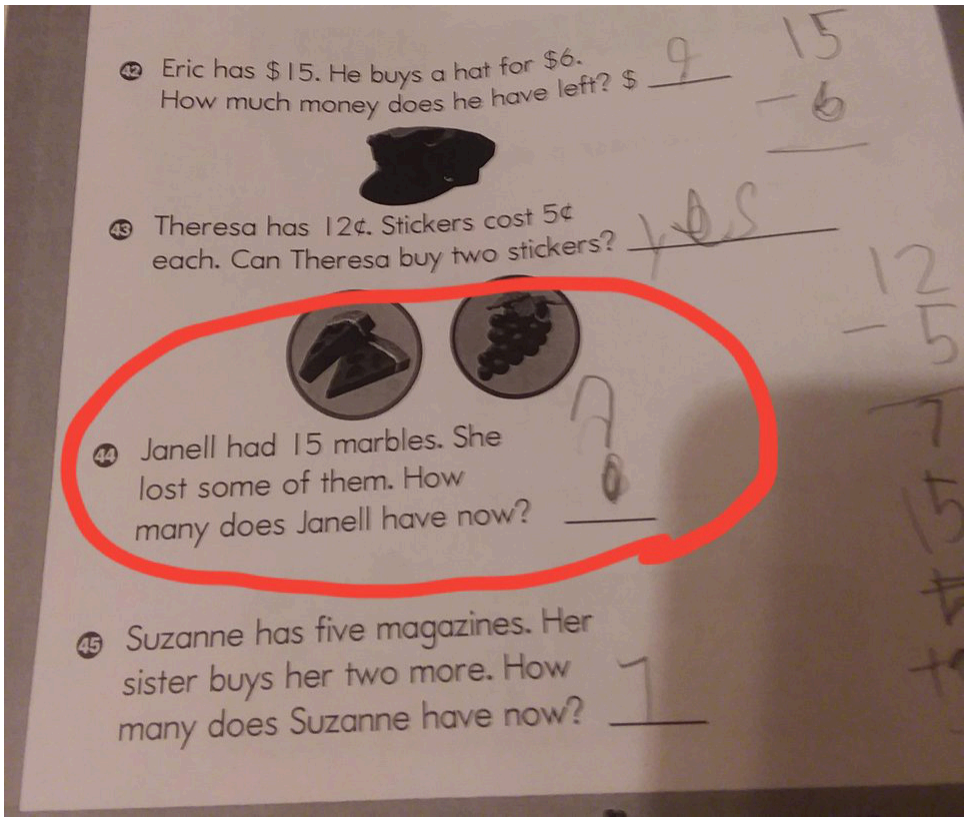


Figure I.10: From a third-grade mathematics workbook.

"We do not see things as they are, we see them as we are."

Anaïs Nin

I.6.1.37. The question remains, however, whether it makes sense in general to try to make precise the notion of the "simplest" rule for a given initial sequence segment. This notion would appear to be highly subjective and mathematically imprecise; however, we should not reject out of hand the possibility that it could have some meaning and that the ability to recognize simple patterns is indicative of general mathematical ability. Thus the proper formulation and use of this type of question on standardized tests is a matter for legitimate debate and study.

I.6.1.38. One possible "simplest" formula for a sequence in which n terms are given is the unique polynomial of degree $\leq n-1$ giving these values at $1, \ldots, n$ (cf. V.11.1.1.), especially if (as in the above cases) the degree of this polynomial is smaller than $n-1$. However, this type of definition excludes simple and natural recursive rules which can be found for certain sequences. A good example is the famous Fibonacci

sequence, which arises naturally in many applied situations, notably in biology. The first few terms are $(1, 1, 2, 3, 5, 8, \dots)$, and each term, beginning with the third, is the sum of the two previous terms.

I.6.1.39. Here is another example, taken from [CG96, p. 76-79], where the "simplest" formula is not a polynomial. What is the next term in the sequence $(1, 2, 4, 8, 16, \dots)$? The "obvious" answer is 32, which is correct if each term is twice the preceding term (i.e. the n'th term is 2^{n-1}). But here is another natural (but somewhat more complicated) sequence beginning with the same terms. To get the n'th term x_n, place n points on a circle and connect each pair of points with a line segment (chord), and count the number of regions the disk bounded by the circle is divided into. It is not too hard to see that the answer does not depend on how the points are arranged on the circle, as long as no three of the chords pass through the same interior point (so, for example, the points could not be evenly spaced around the circle if n is even and more than 4). It is easy to check that $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, $x_4 = 8$, and $x_5 = 16$. But $x_6 = 31$, not 32, and the difference between x_n and 2^{n-1} becomes more dramatic as n increases (it is easily seen that 2^{n-1} is far too large if n is, say, 20 or 30). In fact, it can be shown that

$$x_n = \binom{n-1}{0} + \binom{n-1}{1} + \binom{n-1}{2} + \binom{n-1}{3} + \binom{n-1}{4}$$

(where $\binom{m}{k}$ is the *binomial coefficient* (V.12.5.1.), the number of ways of choosing k elements from a set of m elements), which is a fourth-degree polynomial

$$x_n = \frac{1}{24}(n^4 - 6n^3 + 23n^2 - 18n + 24)$$

This is the unique polynomial of degree ≤ 4 giving the specified terms for $n \leq 5$, and for the circle division rule gives x_n for all n.

I.6.1.40. Here is yet another example. Consider the sequence $(1, 4, 6, 13, \dots)$. These are the first terms of the *busy beaver function* Σ, where $\Sigma(n)$ is the maximum number of 1's an n-state Turing machine with alphabet $\{0, 1\}$ can write on a blank tape (all 0's) and then halt (there are only finitely many such Turing machines, so the maximum exists). Σ is an uncomputable function since there is no algorithm for determining whether a Turing machine halts, and in fact grows faster than any computable function. $\Sigma(5)$ was conjectured to be 4098, and this was recently (apparently) proved by a consortium of amateur mathematicians, some of whom are anonymous. No values of $\Sigma(n)$ are known for $n > 5$. $\Sigma(6)$ is spectacularly larger: it is now known to be at least $2 \uparrow\uparrow (2 \uparrow\uparrow (2 \uparrow\uparrow 10))$ (II.8.6.16.). In general, $\Sigma(2k) > 3 \uparrow^{k-2} 3$, and is asymptotically *much* larger. (There are other related functions sometimes called busy beaver functions, notably the function denoted $S(n)$ or $BB(n)$ whose value is the number of state changes an n-state Turing machine can make before halting, which is larger than $\Sigma(n)$ but not dramatically so. See [Aar20] for a survey.)

Not only is the whole function Σ uncomputable[45], but there are even specific n's for which $\Sigma(n)$ is in principle uncomputable, say in ZF (the smallest n known to have this property is 745, since a Turing machine of this size which just writes 1's can be constructed which halts if and only if ZF set theory is inconsistent; but it is generally believed there are much smaller such n, perhaps as small as 20).

[45]There is no universally accepted definition of a computable function, but we can take this to mean Turing computable. A function $f : \mathbb{N} \to \mathbb{N}$ is *Turing computable* if there is a Turing machine with alphabet $\{0, 1\}$ such that, for any $n \in \mathbb{N}$, if the machine is given an input tape with n 1's it outputs $f(n)$ 1's on the tape and then halts. This is equivalent to several other natural notions of computability for functions. Computability of $f(n)$ for a fixed n is more difficult to define precisely, but means there is a definite finite procedure (perhaps dependent on n) for calculating the numerical value from the definition of the function, or a finite proof that a specific number m (specified by a particular algorithmic or inductive numerical notation, e.g. II.8.9.4.) is equal to $f(n)$.

I.6.1.41. Even a computable function can grow enormously rapidly. The sequence of I.6.1.27. is an example. A more dramatic example, with important logical consequences, is the *Goodstein function*.

Begin with a natural number m. For $n \in \mathbb{N}$, define $g(m,n)$ inductively as follows. Set $g(m,1) = m$. If $g(m,n-1)$ has been defined and is nonzero, to get $g(m,n)$ take the complete normal form of $g(m,n-1)$ with base n (II.8.9.11.), replace n by $n+1$ wherever it occurs, and subtract 1. If $g(m,n) = 0$ for some n, terminate the process (all succeeding terms are also 0). The sequence $g(m,1), g(m,2), g(m,3), \dots$ is called the *Goodstein sequence* with starting point m. For example, for $m = 3$,

$$g(3,1) = 3$$

$$g(3,2) = 3 + 1 - 1 = 3$$

$$g(3,3) = 4 - 1 = 3$$

$$g(3,4) = 3 - 1 = 2$$

$$g(3,5) = 2 - 1 = 1$$

$$g(3,6) = 1 - 1 = 0$$

so the Goodstein sequence with starting point 3 terminates after 6 steps (has length 6).

The example with $m = 3$ is misleadingly simple. The sequence with $m = 4$ seems to increase without bound; for example, $g(4,96) = 11,327$. For larger m the sequence increases much more rapidly.

But it turns out that the Goodstein sequence terminates for every m. This can be proved in ZF (but not in first-order Peano arithmetic!) There is a slick and beautiful ZF proof using ordinals: for every m and n an ordinal $\delta(m,n)$ is defined with the properties that $g(m,n) \leq \delta(m,n)$ for all m,n and $\delta(m,n+1) < \delta(m,n)$ for all m,n for which $\delta(m,n) > 0$. For any m the sequence $(\delta(m,1), \delta(m,2), \dots)$ is a strictly decreasing sequence of ordinals and thus must reach 0 in a finite number of steps since the ordinals are well ordered. For details see [Pot04, 13.1].

So define the *Goodstein function* $G : \mathbb{N} \to \mathbb{N}$ by setting $G(m)$ to be the smallest n for which $g(m,n) = 0$. This is a computable function: any Turing machine which can do arithmetic and a few other simple tasks can compute it. But it grows enormously rapidly: we have $G(1) = 2$, $G(2) = 4$, $G(3) = 6$, but

$$G(4) = 3 \cdot 2^{402,653,211} - 2 \approx 6.895080803 \times 10^{121,210,694}$$

$$G(5) > 10^{10^{10^{19728}}}$$

and $G(12)$ is larger than Graham's number (). Exact formulas for $G(n)$ are known. The fact that it grows so rapidly is the basis for the fact that the termination of the Goodstein sequence for all m cannot be proved in first-order Peano arithmetic. Thus the sentence

$$\forall m \exists n (g(m,n) = 0)$$

is a statement which is true in $\mathbb{N}$ but cannot be proved in first-order Peano arithmetic. (The proof of Gödel's Incompleteness Theorem also gives such a sentence, but it is much more complicated and requires a choice of coding.)

Even though G grows very rapidly, we have $\Sigma(n) \gg G(n)$ for even modestly large n.

So what is the next term in the sequence $(2, 4, 6, \dots)$?

I.6.1.42. Readers interested in exploring mathematical sequences should check out the website

http://oeis.org

This site lets the user search for naturally-occurring sequences beginning with any specified terms. For example, the search currently gives 98 mathematical sequences beginning with (or including as a segment) 2,5,10,17. For example, if x_n is the length of the longest uncrossed knight's path on an $(n+2) \times (n+2)$ chessboard, then the first few terms are $(2, 5, 10, 17, 24, 35, \dots)$.

Poorly Worded Problems

> "A man left Albuquerque heading east at 55 miles per hour. At the same time, another man left Nashville heading west at 60 miles per hour. After some time the two men met. When they met, which man was closer to Nashville?"
>
> *S. Gerofsky*[46]

I.6.1.43. Poorly worded "practical" problems are more generally a bane of mathematics education. Students are frequently subjected to problems which are confusingly worded and/or leave crucial assumptions unspecified. The difficulties are not always as obvious as the one in Figure I.10 (my solution to that problem is that it depends on the Axiom of Choice).

I.6.1.44. One common kind of problem assumes without mention a linear or inverse-linear relationship between quantities which may be invalid or only partially valid. A common version is to say: "One person can do a job in x hours. How long does it take n people to do the job?" It depends, of course, on the job! For example:

> "It takes one man with a shovel 8 hours to dig a trench of a certain size. How long does it take for 4 men with shovels to dig the same trench?"

It probably does indeed take about two hours (but this is unclear if rest breaks are needed). But can 32 men with shovels dig the trench in 15 minutes? Can 480 men with shovels dig the trench in one minute? And:

> If one driver can drive a truck from Reno to San Francisco in four hours, how fast can two drivers drive the truck from Reno to San Francisco?
>
> If five men can win a basketball tournament in three days, how fast can one man win the tournament?
>
> If a man with an annual income of $25,000 can afford to eat one steak dinner a week, how many steak dinners a week can a man with a $1,000,000 annual income afford to eat?
>
> "You can only drink 30 or 40 glasses of beer a day, no matter how rich you are."
>
> *Adolphus Busch*

[46] https://www.collectionscanada.gc.ca/obj/s4/f2/dsk1/tape7/PQDD_0027/NQ51864.pdf.

I.6.1.45. A problem cited in [Str13] as commonly misleading people is:

> Suppose three men can paint three fences in three hours. How long would it take one man to paint one fence?

It is claimed that many people give the knee-jerk "incorrect" answer of one hour, and that the "correct" answer is three hours. This is indeed correct if the painters work independently (plus some other assumptions, such as that all paint at the same rate and that the fences are the same size; some of these are acknowledged in [Str13]). But there could be many reasons why this answer would be wrong. For example, maybe Joe paints the base color, then Bill paints the trim, and then Ed paints the lettering, and these have to be done consecutively, one fence at a time. If this is the case and no drying time is needed, the correct answer might very well be one hour. If drying time is required between coats, the correct answer could be anything between one and three hours. In fact, under almost any assumptions the answer would be between one and three hours; but it could even be less than one hour if the three hours required rest breaks or if the three fences are far apart with significant transit time between them, or it could be more than three hours if only one man actually paints and the other two supply him.

See [**?**] for more serious, sophisticated, and interesting examples. The "too-many-cooks-spoil-the-broth" phenomenon, also known as "Brooks' Law," is discussed in [**?**]. Scheduling problems can be inherently very difficult mathematically.

I.6.1.46. Here are a couple of other badly worded problems which recently appeared on a worksheet used in a university remedial math class:

> "Shawna reduced the size of a rectangle to a height of 2 inches. What is the new width if the rectangle was 12 inches high and 24 inches wide?"

Did she keep the proportions the same?

> "Which is cheaper, a 12.5 ounce bag of chips for $3.69 or a 3 ounce bag for $0.99?"

The large bag may be cheaper *per ounce*, but the small bag is obviously cheaper. And buying things at Costco isn't necessarily economical for everyone.

> How much dirt is in a rectangular hole with dimensions 1m × 2m × 3m?

> "Word problem stories are poor quality fiction at best."

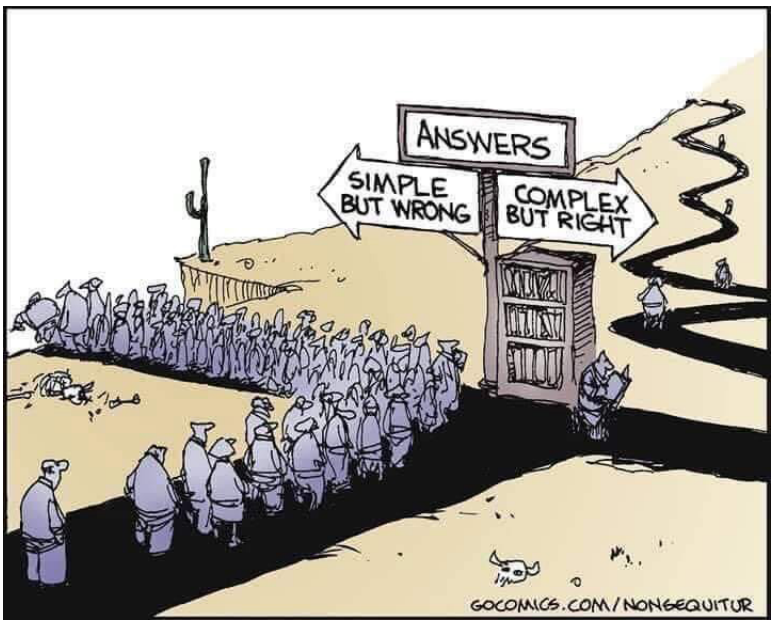
S. Gerofsky[47]

[47] https://www.collectionscanada.gc.ca/obj/s4/f2/dsk1/tape7/PQDD_0027/NQ51864.pdf.

I.6.1.47. If one first-rate physicist can solve a problem in a month, how long will it take for two second-rate physicists to solve it?

"If we can't hire one first-rate physicist, we'll just hire two second-rate physicists."

Ronald Reagan[48]



In fact, the right answer is not necessarily more complex than the wrong one; it just may be harder to find and take more thought and investigation, as well as an open mind.

[48] When told that no reputable physicist would agree to work on his "Star Wars" missile project.



The cartoonist may be a racist, but this one is funny.

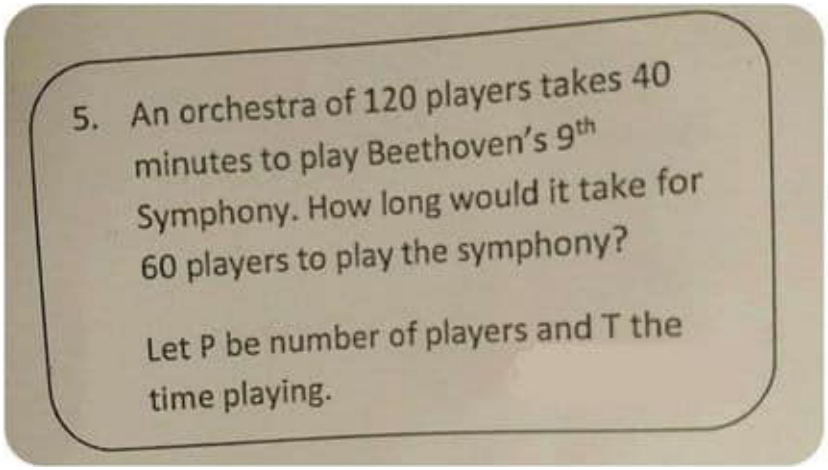


Figure I.11: One last word problem.

Euclid's Parallel Postulate

I.6.1.48. Next, we consider an issue which vexed mathematicians for centuries: EUCLID'S Parallel Postulate. I believe one of the important reasons it took so long to understand this situation was that mathematicians were misled and constrained in their thoughts by subtle implications of the terms EUCLID used in the formulation. (EUCLID himself was misled: in some of his proofs he used additional assertions which must have seemed "obvious" to him, but which do not follow from his axioms.)

I.6.1.49. One of EUCLID'S postulates (or axioms) for geometry, called the Parallel Postulate, states that "if L is a line and p a point not on L, there is one and only one line through p parallel to L." (Actually EUCLID stated this postulate in a rather different and more complicated way; the version we have stated is often called *Playfair's Axiom*, although it was known and used at least as far back as PROCLUS. See [Wil, I.1] for a discussion of the origin and demise of the distinction between axioms and postulates.) In EUCLID'S time, and for a very long time afterwards, axioms were supposed to be "self-evident truths," and the Parallel Postulate is indeed true in ordinary ($\mathbb{R}^2$) geometry. Mathematicians long wondered not whether the Parallel Postulate was "true", but whether it could be proved from Euclid's other axioms (since the actual statement was sufficiently complicated as not to be "self-evident.")

I.6.1.50. Although over the centuries various mathematicians assumed the Parallel Postulate was false, and even proved what are now regarded as theorems of noneuclidean geometry, it was always done with the goal of obtaining a contradiction; it apparently never occurred to mathematicians until the nineteenth century (with the possible exception of J. LAMBERT in the eighteenth) that there really could be "geometries" satisfying the rest of EUCLID'S axioms but not the Parallel Postulate, or, in modern language, models for Euclidean geometry with the Parallel Postulate replaced by its negation. Once the possibility was contemplated, it did not take long for such geometries to be found, in fact independently by several mathematicians: GAUSS, BOLYAI, and LOBACHEVSKII for hyperbolic geometry, and RIEMANN for elliptic (some of EUCLID'S other postulates arguably have to be modified a little in this case – see below); an actual model of hyperbolic geometry was not constructed until later in the century, by BELTRAMI. It follows immediately that the Parallel Postulate is independent of the other axioms of Euclidean geometry, i.e. cannot be proved (or disproved!) from them. (In fact, we know now, via model theory, that the converse is also true: if a statement such as the Parallel Postulate is independent of the other axioms, then there is a model for the other axioms in which the statement is false, although this was far from clear in the nineteenth century.)

I.6.1.51. While the mathematical atmosphere in the nineteenth century was much more conducive to the thought processes leading to these examples than in earlier times, I cannot help but think that the language used by EUCLID had a stifling effect on creative thought about geometry. EUCLID used the terms "point" and "line" as essentially undefined objects (he did give "definitions" which would not be considered true definitions today, e.g. "defining" a point to be "that which has no part" and a line to be "that which has length but no breadth"), but the terms used were strongly suggestive of a particular interpretation in which the Parallel Postulate is rather "obviously" true. It finally took some very smart and creative thinkers to realize that the terms "point" and "line" could be interpreted to mean something other than what EUCLID evidently had in mind, e.g. what we would now call geodesics on surfaces could play the role of lines. This thinking has evolved further (pioneered by POINCARÉ, and culminating in HILBERT'S treatment of the Foundations of Geometry), to the degree that we now realize that "point" and "line" can refer to any kind of mathematical objects at all, so long as the axioms (with suitable interpretation) hold for them.

I.6.1.52. I wonder what would have happened if EUCLID had used some non-suggestive terms (such as "bxptq" and "zkmr") instead of "point" and "line." One of two things would probably have happened: either models of noneuclidean geometry would have been discovered sooner, or (more probably) no one would have continued to study EUCLID at all since his work would not have seemed to be "about" anything! Suggestive terminology has undeniable advantages so long as the pitfalls are understood.

I.6.1.53. Incidentally, I am pretty tired of reading uninformed nonsense about how the development of noneuclidean geometry somehow divorced geometry from physical reality and precipitated a "crisis" in mathematics, especially in its foundations. Such an argument is more worthy of ALFRED E. NEUMAN than JOHN VON NEUMANN, and is based on a very simplistic view of "reality." Although for a brief time the new ideas were evidently unsettling to some mathematicians, I do not think it was a prevalent attitude among mathematicians to be upset by the developments and fear that mathematics might be a house of cards which was about to collapse; in fact, it seems that most good mathematicians thought it was an exciting time in mathematics when the frontiers of knowledge and understanding were expanding rapidly. (There was a controversy in the German mathematical community in the 1920s called the "foundational crisis," but this was really a dispute over which parts of mathematics, if any, could be considered "true" or "meaningful" and not whether mathematics held together logically.) It is certainly true that the geometric discoveries, along with the roughly simultaneous "arithmetization" of analysis, led to a profound revolution in the nature and conception of mathematics. But even before EINSTEIN came along to definitively discredit the idea that the geometry of physical space is Euclidean (POINCARÉ, HILBERT, and MINKOWSKI played important roles too), it could be seen that RIEMANN'S spherical (elliptic) geometry was in fact the geometry of the surface of the earth, physical geometry of the most basic sort (recall that the very word *geometry* comes from Greek meaning "measuring the earth"). Hyperbolic geometry has less immediate direct application to the familiar world, but it cannot be reasonably argued that it has no connection with physical reality. Indeed, general relativity could not have been developed without geometric ideas which were outgrowths of the discovery of noneuclidean geometry.

> "The value of non-Euclidean geometry lies in its ability to liberate us from preconceived ideas in preparation for the time when exploration of physical laws might demand some geometry other than the Euclidean."
>
> *B. Riemann*[49]

Anyone who thinks noneuclidean geometry is divorced from reality should read [Pen05] and [MTW73]. See [Gra08] for a thorough and intelligent study of the "modernist" revolution in mathematics without resort to the "crisis" characterization.

> "[T]he development of mathematics may seem to diverge from what it had been set up to achieve, namely to simply reflect physical behaviour. Yet, in many instances, this drive for mathematical consistency and elegance takes us to mathematical structures and concepts which turn out to mirror the physical world in a much deeper and broad-ranging way than those that we started with."
>
> *Roger Penrose*[50]

[49] cf. [Gre93, p. 371].
[50] [Pen05, p. 60].

Euclid's Postulates

I.6.1.54. Here are EUCLID'S five postulates, according to [Euc02][51] (other sources give slightly different wording):

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any center and distance.
4. That all right angles are equal to one another.
5. That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

Postulate 5 is EUCLID'S version of the Parallel Postulate (note that EUCLID neglected to specify that the two lines not both pass through the same point of the line falling on them! Maybe this is implicit in the "interior angle" term which is actually not defined.) It is conventional wisdom, stated in many references, that Playfair's Axiom is equivalent to EUCLID'S Postulate 5, in the presence of Postulates 1–4; however, as we shall see, this equivalence requires that some things be read into Postulates 1–3 which are (arguably) not really there. These additional assertions evidently seemed "obvious" to everyone from EUCLID'S time until the nineteenth century (M. PASCH seems to have been the first to notice) and apparently escaped the notice of everyone who tried to deduce Postulate 5 or its believed equivalent Playfair's Axiom from the other postulates (some seem to escape the notice of even some contemporary writers!)

Does Spherical Geometry Satisfy Euclid's Postulates?

A good case study to illustrate the principles is spherical geometry. Whether spherical geometry satisfies EUCLID'S postulates is a matter of interpretation which also illustrates how unwarranted consequences can be read into statements.

I.6.1.55. In spherical geometry, a sphere S is fixed. "Points" of the geometry are just points of S. "Lines" ("straight lines") of the geometry are great circles. "Line segments" ("finite straight lines") are (usually minor) arcs of great circles. "Circles" are ordinary circles on S, not necessarily great circles; we will also call singleton sets circles. The center of a circle has its usual interpretation (a circle actually has two antipodal centers!); the center of a singleton circle is the point itself (or the antipodal point). "Angles," particularly "right angles," are interpreted in the usual way using tangent lines.

Note that it does not actually contradict the postulates that certain circles are also "straight lines" (or single points), or that circles with different centers can coincide – it only contradicts our intuition about what these terms should "mean." (Note that there is an ambiguity in EUCLID'S definitions between "circle" and "disk"; in spherical geometry, distinct disks can have the same boundary circle.)

It must also be noted that spherical geometry is not at all new – the basic facts were already known in EUCLID'S time. But it was not until the nineteenth century that it was realized that his postulates have an interpretation in spherical geometry.

[51] Translation by SIR THOMAS L. HEATH, K.C.B., K.C.V.O., F.R.S., Sc.D. Camb., Hon. D.Sc. Oxford, Honorary Fellow (Sometime Fellow) of Trinity College, Cambridge; with these qualifications, it must be right.

I.6.1.56. The first observation is that spherical geometry actually satisfies Postulate 5 as EUCLID stated it. Since it obviously does not satisfy Playfair's Axiom (according to EUCLID'S definition, "parallel" means "nonintersecting," so there are no parallel lines in spherical geometry), either spherical geometry fails to satisfy all of the other four postulates or Playfair's Axiom is not really equivalent to Postulate 5.

I.6.1.57. Next look at Postulate 4. What this postulate evidently means is that there is an assumed notion of congruence. An easy way to accomplish this (although not the most general way) is to assume there is a set (group) of sufficiently many "rigid motions" available to translate and rotate figures (specifically, which is transitive on points and lines). There is such a group ($O(3)$ or $SO(3)$) for the sphere.

I.6.1.58. The suspect postulates in spherical geometry are 1–3. First consider 1. If p and q are distinct points on a sphere, there is clearly a line through p and q, so if Postulate 1 is read strictly as EUCLID stated it, it is satisfied in spherical geometry. However, most readers, beginning with EUCLID himself, include into Postulate 1 an implicit requirement that the line be unique, which fails on the sphere if p and q are antipodal points. So whether spherical geometry satisfies Postulate 1 is a matter of interpretation. (In the projective plane, the quotient of the sphere by the antipodal map, the line through two given points is unique, although the line *segment* between two points is not necessarily unique. EUCLID does not clearly distinguish in his postulates between lines and line segments; indeed, he appears to actually only consider line segments.)

I.6.1.59. Now look at Postulate 2, which leads to similar but more interesting interpretation questions. The statement is somewhat vague, and can be interpreted in several reasonable ways. It might mean simply that every line segment is contained in a line (which may even be equal to the line segment). But the word *produce* might implicitly mean that the line is larger than the segment (or just that a line segment can be extended to a larger line segment). We might expect that it means that the line segment between, say, p and q can be "extended beyond q," whatever that means from the Postulates, and similarly beyond p. It might then mean that a line segment from p to q could always be continued to a point r giving a line segment from p to r containing q, with the line segment from q to r the magnitude of a fixed reference segment, and similarly beyond p. With this interpretation, "lines" in spherical geometry, i.e. great circles, satisfy the postulate. But someone thinking about ordinary Euclidean geometry reading the postulate would probably interpret it as meaning that the extension of the line segment consists entirely of distinct points, i.e. the extension beyond q does not intersect the extension beyond p, which fails on the sphere. EUCLID himself made and used this assumption, sometimes called the "infinitude of lines," without mentioning it in his postulates. The "infinitude of lines" interpretation is necessary if one wants to have a notion of one point being "between" two others on a line, which is important in Euclidean geometry; there is no such notion in spherical geometry. "Betweenness" is not mentioned in EUCLID'S postulates, but is a fundamental concept in HILBERT'S more careful axioms for Euclidean geometry. Actually, "betweenness" is implicit in the notion of line segment with endpoints; in the presence of Postulates 1 and 2, along with a weak uniqueness statement that three distinct points can lie on at most one line (which does not follow from EUCLID'S postulates as stated), the "infinitude of lines" is equivalent to the statement that if three distinct points are on a line, exactly one is "between" the other two.

The "infinitude of lines" is distinct from the question of whether the extension in Postulate 2 is unique, i.e. whether two different lines can have a segment in common, another assumption made by EUCLID not included, but arguably implicit, in the postulate; extensions are unique in spherical geometry. (The uniqueness question becomes trickier if "line" means "line segment," as it apparently did to EUCLID: two line segments can obviously have a smaller segment in common.) I think it is somewhat harder to argue that the "infinitude of

lines" is implicit in Postulate 2 than that extensions are unique. Thus whether spherical geometry satisfies Postulate 2 is again a matter of interpretation.

There is another interpretation ambiguity: what does the word "continuously" mean in this postulate? Any reasonable interpretation would seem to be satisfied in spherical geometry, although some authors have suggested that "continuously" refers to the infinitude of lines, which seems doubtful to me.

I.6.1.60. Postulate 3 appears to be false in spherical geometry; there is a limit to the radius of circles on a sphere. But this conclusion depends on a modern interpretation of the phrase "any distance." EUCLID did not use the notion of numerical distance between points; in fact, in ancient Greece a clean distinction was made between numbers (natural numbers), which denoted quantity, and "magnitudes," which denoted size (they also considered "ratios" of numbers, and of magnitudes, some of which were "irrational"). To EUCLID, a "distance" would have meant the "magnitude" of a line segment between two points. Thus he arguably would have agreed that Postulate 3 could have been equivalently phrased:

3′. Given points p and q, a circle can be drawn with center p passing through q.

In fact, this is the form of Postulate 3 always used in his proofs (and in straightedge-and-compass constructions); cf. [Gre93]. With this interpretation, spherical geometry does indeed satisfy Postulate 3. There is a slight problem if p and q are antipodal points on the sphere; but this problem evaporates if we allow a circle to be a single point, or if we pass to the projective plane, although this could cause (slight) difficulties with Postulate 5 since a line does not divide the projective plane.

I.6.1.61. Note also that the notion of "magnitude" of a line segment, as well as Postulate 4, only make sense if there is an assumed notion of congruence. As previously noted, there is such a notion for the sphere (and for the projective plane).

I.6.1.62. Thus it can be argued, although perhaps somewhat weakly, that spherical geometry satisfies all of EUCLID'S postulates (it can be argued somewhat less weakly that the projective plane satisfies all the postulates), and also that the equivalence of Playfair's Axiom with Postulate 5 depends on reading extra assertions into the other postulates (specifically, uniqueness in Postulate 1; cf. [**?**]).

Other Interpretations of Euclid's Postulates

I.6.1.63. One could also consider the set P of points in $\mathbb{R}^2$ with rational coordinates (*rational points*). Lines in P will consist of all rational points on an ordinary line through two rational points, and circles in P will consist of rational points on an ordinary circle with rational center passing through a rational point. (There is a dense set of rational points on any such line or circle.) Then Postulates 1, 2, and 5 hold in P, including uniqueness and the "infinitude of lines," and Postulate 3 holds in the form of 3′, with no bound on the radii of circles. The (slight) difficulty this time is with Postulate 4, since the group of rigid motions is too small to allow comparison of all pairs of right angles; but the group of similarities is large enough to allow such comparison in a well-defined way, so Postulate 4 can be said to be satisfied in P. (Comparison of magnitudes in P is also not possible in general for the same reason.) But in this geometry EUCLID'S first proposition, construction of an equilateral triangle with two given vertices, fails since the circle with center p passing through q and the circle with center q passing through p do not intersect even though they "overlap." EUCLID assumed in his proof that such circles intersect, an assumption which does not follow from his postulates. (The assumption also fails in spherical geometry if p and q are too far apart.) In fact, P contains no equilateral triangles.

I.6.1.64. Consider also the following simple system. The space consists of exactly four points. Lines are defined to be sets containing exactly two points. A circle is a set of exactly three points; the fourth point is called the center of the circle. A right angle is a set of three points with one of them designated as the vertex. Two lines are parallel if they are disjoint. The angle between two nonparallel lines is defined to be the right angle consisting of the union of the lines with the intersection as vertex. Euclid's postulates 1, 2, and 3′ are satisfied (except possibly for the word "continuously" in Postulate 2, whatever it is supposed to mean), including uniqueness in 1 and 2 (and in 3′); the "infinitude of lines" is also trivially satisfied since the extension of a line segment is just the segment itself. If the group of rigid motions is taken to be the group of all permutations of the space, Postulate 4 is also satisfied. And Postulate 5 is vacuously satisfied since the angle between any two nonparallel lines is always a right angle. Playfair's Axiom also holds in this space. This example and the previous one show that Euclid's postulates do not uniquely determine a geometry.

I.6.1.65. Note that it is also consistent with the postulates that there be no points, and any number of lines (including none), or that there be exactly one point and any number of lines, or that there be any number of points and exactly one line containing all of them. In these cases, some or all of the postulates will hold vacuously. These situations could be ruled out by assuming that there exist three points not all on the same line, another assumption Euclid made in many places.

The Simplest Noneuclidean Geometry?

I.6.1.66. There was a counterexample staring everyone in the face who tried to deduce Euclid's Parallel Postulate from the other postulates, if they had been open-minded enough to realize it! (To be fair, it would be unreasonable to expect them to have recognized it before the abstraction of the nineteenth or even twentieth century.) Just take $\mathbb{R}^3$ with the usual points, lines, and line segments, and interpret a "circle" to be a sphere. Euclid's first four postulates hold, even with the extra uniqueness provisions added. But if we use Euclid's definition that "parallel" means not intersecting, Playfair's Axiom fails (skew lines are "parallel"). There is a semantic problem with Euclid's fifth postulate because it refers to a "side" of a line, but this can be bypassed simply by declaring that a line has only one side (i.e. the side of the line consists of all points not on the line); with this interpretation, the fifth postulate also fails.

This example would surely have been regarded as cheating (even if abstract interpretations were accepted allowing a sphere to be called a "circle"), since Euclid's postulates were clearly designed to define plane geometry. But there is nothing in any of Euclid's postulates or common notions which would force two-dimensionality or rule out examples of this sort (he could have easily done so with an additional postulate). From modern hindsight, this example would immediately show that neither Euclid's fifth postulate nor Playfair's Axiom can be proved from the other assumptions Euclid spelled out.

Was Euclid More Prescient Than We Thought?

I.6.1.67. If Euclid's definitions, imperfect as they are, are closely examined, it could be argued that they include implicit additional postulates which rule out examples such as I.6.1.63.–I.6.1.66., and even force the space considered to be at least what we would call a surface or 2-manifold. He defines a "line" (meaning what we would call a "curve"; he goes on to define a "straight line" to be a line satisfying a condition I find incomprehensible) to be "that which has length but no breadth." Length and breadth are not defined, but one could argue that this definition means that lines must be topologically one-dimensional, an interpretation reinforced by the word "continuously" in Postulate 2 which is otherwise mysterious, and must have a well-defined arc length. In particular, lines (curves) must be isometric to intervals in $\mathbb{R}$, anticipating modern

principles of completeness and measure which were far in the future in ancient Greece.

EUCLID'S definition of a circle is "a plane figure contained by one line such that all the straight lines falling upon it from one point among those lying within the figure equal one another." This actually defines what we would call a disk, but if we take the term "circle" to mean the boundary of such a figure, it must be a line or curve in the above sense, and must implicitly have an interior and exterior as well as a center. Note though that this definition, like some of his others, is circular (pun intended) in that it refers to a "plane figure" although the whole theory is intended to establish what a "plane figure" is. There is also an inconsistency in his definitions: he says the extremities of a line (curve) are points (suggesting that he only considers line segments), but a circle has no extremities. (Granted, he does not actually say how many extremities or endpoints a curve has.)

I.6.1.68. Thus, although EUCLID'S postulates and axioms themselves fall far short of adequacy to describe plane geometry, when the additional postulates arguably implicit in his definitions (and postulates) are added, as most interpretations over the centuries have done, they come remarkably close.

I.6.1.69. **Exercise:** Consider the cylinder, flat torus (), hexagonal torus (), Möbius strip (), and Klein bottle () as quotients of $\mathbb{R}^2$ with the induced Riemannian structure. Which of EUCLID'S postulates, including additional ones implicit in his definitions and postulates, are satisfied in these examples?

> "Proclus shared the human problem of distinguishing between the obvious and that which only appears to be obvious."
>
> *George E. Martin*[52]

> "One of the pitfalls of working with a deductive system is too great familiarity with the subject matter of the system. It is this pitfall that accounts for most of the blemishes in Euclid's *Elements*."
>
> *Howard Eves*[53]

I.6.1.70. **Exercise:** Consider the following proof from the *Elements* (Book 1, Proposition 16) that an exterior angle of a triangle is always greater than either of the other interior angles. This proposition implies that for any line L and point p not on L, there is at least one line through p parallel to L (how?). Assume Postulates 1, 2, 3′, and 4 as written, and interpret Postulate 2 to mean that a line segment from p to q can always be continued to a point r giving a line segment from p to r containing q, with the line segment from q to r the magnitude of a fixed reference segment, and similarly beyond p. (We do not need to assume or use Postulate 5.) Assume it has been previously proved that any line segment can be bisected, that opposite angles of intersecting lines are equal, and SAS congruence holds (these could be taken as additional postulates if desired).

Given triangle $\triangle ABC$, extend AB to a point D beyond B. We wish to prove that $\angle CBD$ is greater than angles $\angle ACB$ and $\angle CAB$. Let G be the midpoint of BC, and extend AG to a point H so that $GH = AG$. Then triangles $\triangle ACG$ and $\triangle HBG$ are congruent by SAS, so $\angle ACB = \angle CBH < \angle CBD$. Also, BH is parallel to AC, so $\angle CAB = \angle HBD < \angle CBD$. See Figure I.12.

[52] [?, p. 138].

[53] An Introduction to the History of Mathematics, 1953, p. 126.

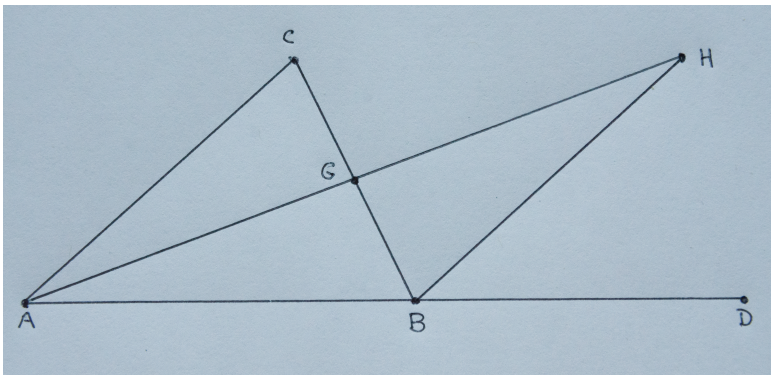


Figure I.12: Euclid's Proposition 16 Diagram.

Note that all the assumptions are true in spherical geometry (along with Postulate 5), but the conclusion does not hold. What is wrong with the proof? [How do you know the segment GH does not intersect the line through A and B? How do you even know $H \neq A$?] Which additional implicit assumptions are needed to make the proof work?

I.6.1.71. My purpose in discussing these matters is not to find fault with EUCLID, but simply to illustrate how easy it is to jump to conclusions not supported rigorously. (In [**?**] it is stated that only "little people" like to point out flaws in EUCLID; then the author himself describes some of the flaws over the next several pages!) Although we may find flaws in it, EUCLID'S *Elements* was a remarkable intellectual achievement (not just by EUCLID, but also by his predecessors who developed much of the theory) which is a leading candidate for the greatest mathematical work of all time, and remains the most widely read, studied, and influential non-religious book ever written on any subject. No mathematical work, with the arguable exception of ARCHIMEDES' writings, rivaled it in importance at least until NEWTON'S *Principia* two thousand years later (and to call *Principia* a mathematical work might be a bit of a stretch), and the *Elements* was more "modern" and rigorous than anything else written before about the nineteenth century.

> "If mathematics does not teach us how to think correctly, it at least teaches us how easy it is to think incorrectly."
>
> *Harald Bohr*[54]

> "The genius of Euclid is insufficiently celebrated. When he sat down to write his *Elements of Geometry*, he probably did not know that he was going to create a work that would last for centuries and would forever change how the human race thought, not just about geometry but about truth and how to get at it. He probably did not realize that he was one of the chief figures in a revolution that, in a mere three hundred years or so, made our world possible by elevating reason to its proper place. Probably he was thinking only that it would be a good idea to gather up all the pieces of mathematics that were scattered around (some of them, he may have thought,

[54] [Ped89, p. ix].

shockingly badly written) and put them in one place, neatly and logically organized. After all, Euclid worked in a library and may have had some of the habits of mind that librarians are either born with or quickly absorb from the atmosphere of libraries. But whatever his motives, Euclid deserves to be remembered.

Euclid's book has been a model of deductive reasoning ever since it appeared. Deductive reasoning is highly prized as a method of getting at truth, since from truth only truth can follow by deduction. Science is based on it, and even lawyers and politicians attempt to use it at times. It is such a part of our civilization that we cannot imagine a real civilization without it. Deductive reasoning has made us what we are. That is why we insist that high-school students be exposed to a version of Euclid: it is not that any of them will ever have to prove a theorem about triangles, it is just our way of acknowledging our debt to deductive reasoning."

Underwood Dudley[55]

What is Space?

The questions about what "space" or a "line" really *is*, and whether they are really collections of points, are rather subtle quite beyond the issues debated in earlier centuries.

I.6.1.72. There are several distinct notions of space, which must not be confused: intuitive space, mathematical space (of which there are many kinds, geometric or Euclidean space being only one of them), and physical space, which according to theories of modern physics may not be much like any of these: it may fail to be flat (relativity), three-dimensional (relativity/string theory), or continuous or homogeneous at small scales (quantum mechanics). Despite the misconceptions described in (), humans do indeed seem to regard intuitive space as being "continuous" in some sense, perhaps even more "continuous" than the real numbers.

I.6.1.73. Although geometric space can be pretty naturally coordinatized using numbers, we must distinguish geometric space from intuitive space. One major conceptual problem is that we can have linearly ordered sets which are "larger" than $\mathbb{R}$. This problem can be dismissed as just another overly literal interpretation of mathematicians' use of the term "linear ordering" as a synonym for "total ordering." However, in this case I think such criticism would be off base, and that there is a real conceptual difficulty. I do indeed visualize linearly ordered sets as points on a line, even linearly ordered sets which cannot be embedded in $\mathbb{R}$ such as the nonstandard real numbers or the first uncountable ordinal (although I know not to read too much into this visualization), and I imagine many other mathematicians do likewise. This conception raises a serious set-theoretic question: if an intuitive line is so rich in points that any totally ordered set can be embedded in it, can we really regard a line as a set of points? There is an apparent contradiction: it is not difficult to construct totally ordered sets in which the intervals have arbitrary cardinality. In fact, HAUSDORFF concocted a totally ordered "set" (class) in which every interval contains a "subset" order-isomorphic to the class of all ordinals. (See, for example, [Ruc82, p. 88-89] for a description.)

[55] [Dud92, p. 78].

"Obvious" is Subjective

I.6.1.74. Finally, we turn to a somewhat different issue: it is a highly subjective matter what details of an argument an author or lecturer feels need to be explained and justified, and which ones can be safely considered "obvious." Opinions about this can vary widely from person to person, and even from time to time with the same person, and are not always justifiable. Rather than discuss this issue further, let us just recall a standard joke which illustrates it nicely:

> A mathematician was giving a lecture, and at one point he said, "It is obvious that ..." He then paused in mid-sentence, a concerned look came over his face, and he paced back and forth, scratching his head and stroking his chin, deep in thought. After about five minutes, his face suddenly brightened and he exclaimed, "Ah, yes! It is obvious."

If an instructor ever tells you something is obvious and you aren't so sure, or it is not obvious to you, don't be reluctant or embarrassed to ask for an explanation. It may turn out that it is not so obvious to the instructor either!

> "Mathematics is often erroneously referred to as the science of common sense."
>
> *E. Kasner and J. Newman*

Does Familiarity Breed Contempt?

> "I thought at first you had done something clever. But [after Holmes's explanation] I see that there was nothing in it after all."
>
> *Jabez Wilson (to Sherlock Holmes)*[56]

I.6.1.75. It is a pervasive human characteristic to be more in awe of things we do not understand. Once we see a rational explanation of something, the wonder of it tends to fade and it becomes more mundane or commonplace. We, like Jabez Wilson, may even be fooled into thinking it is obvious or easy, and we often underappreciate the effort and/or genius required to find the phenomenon or its explanation in the first place.

Mathematicians are not immune from this tendency. Consider, for example, the wonderful (possibly apocryphal) description by William Lilly[57] of the first encounter between Henry Briggs, one of the principal developers of the theory of logarithms, and John Napier, their inventor, at Napier's home in 1615:

> "At first, when the Lord Napier, or Marchiston, made publick his Logarithms, Mr. Briggs, then reader of the astronomy lecture at Gresham-College in London, was so surprized with admiration of them, that he could have no quietness in himself, until he had seen that noble person the Lord Marchiston, whose only invention they were: he acquaints John Marr herewith, who went into Scotland before Mr. Briggs, purposely to be there when these two so learned persons should meet. Mr. Briggs appoints a certain day when to meet at Edinburgh: but failing thereof, the Lord Napier was doubtful he would not come. It happened one day as John Marr and the Lord

[56] *The Red-Headed League*, Arthur Conan Doyle.
[57] *History of his Life and Times*,?

> Napier were speaking of Mr. Briggs; 'Ah, John,' saith Marchiston, 'Mr. Briggs will not now come:' at the very instant one knocks at the gate; John Marr hasted down, and it proved Mr. Briggs, to his great contentment. He brings Mr. Briggs up into my Lord's chamber, where almost one quarter of an hour was spent, each beholding the other almost with admiration, before one word was spoke: at last Mr. Briggs began.
>
> 'My Lord, I have undertaken this long journey purposely to see your person, and to know by what engine of wit or ingenuity you came first to think of this most excellent help unto astronomy, viz. the Logarithms; but, my Lord, being by you found out, I wonder no body else found it out before, when, now known, it is so easy.' "

Mathematicians tend to become particularly blasé about parts of elementary mathematics as they attain higher levels of knowledge and sophistication in the subject (perhaps including some of the mathematics we discuss in this book). The fact that in the not-too-distant past some of this mathematics represented the pinnacle of human thought can be overlooked in its familiarity to us.

> "A mathematician's first steps into unknown territory constitute the first phase of a familiar cycle.
>
> After the darkness comes a faint, faint glimmer of light, just enough to make you think that something is there, almost within reach, waiting to be discovered. ... Then, after the faint, faint glimmer, if all goes well, you unravel the thread – and suddenly it's broad daylight! You're full of confidence, you want to tell anyone who will listen about what you've found.
>
> And then, after day has broken, after the sun has climbed high into the sky, a phase of depression inevitably follows. You lose all faith in the importance of what you've achieved. *Any idiot could have done what you've done, go find yourself a more worthwhile problem and make something of your life.* Thus the cycle of mathematical research ... " [Emphasis original]
>
> *C. Villani*[58]

I.6.1.76. There is also a human tendency to need some sort of explanation of mysterious phenomena. This is the origin of most religion; many people are satisfied and comforted with religious explanations for things with no apparent rational explanation. However, I have observed that mathematicians often have quite the opposite instinct: they demand a rational explanation, and do not properly appreciate something, especially a mathematical result, until they understand it. (I do not mean to suggest that mathematics and religion are incompatible; indeed, many mathematicians, including great ones, have been deeply religious. But among modern mathematicians, the percentage is low: a study [LW98] suggests less than 15% of current mathematicians believe in God. I am among the majority here.) In fact, the drive to understand phenomena and revel in the understanding might tend to be one characteristic of people with a predilection towards mathematics, although I do not want to attempt a universal generalization here; this tendency may more be characteristic in distinguishing people who are intellectually capable and energetic, whatever their academic interests. In any event, a healthy curiosity about the world, both concrete and abstract, seems to be an almost necessary attribute of anyone who is successful in mathematics.

[58] [Vil16, p. 243].

I.6.2. Set Theory and Logic

One of the main goals of many (but not all) mathematicians in the early 20th century was to carefully axiomatize all of mathematics, beginning with set theory. The hope was that a complete list of all theorems of mathematics could be generated, and there would be a mechanical process by which the truth or falsity of any mathematical statement could be determined. This hope turned out to be impossible, but at least set theory has been put on a generally accepted axiomatic footing.

*****Rewrite*****: Through the early 20th century, and to a small extent even today, a debate raged between two schools of mathematics, the "intuitionists" and the "formalists". The current consensus is that the truth lies somewhere between the two camps. On the one hand, the intuitionists' belief that Cantor's theory of the transfinite, and the subsequent formalizations of set theory and mathematical logic, are unsound and have no place in real mathematics has been thoroughly discredited; it is now almost universally accepted that these theories have a legitimate and even central place in the world of mathematics. On the other hand, the intuitionists' main point seems valid: many (perhaps most) mathematical concepts appear to have an intrinsic meaning within the human intellect quite beyond just being a collection of symbols.

Gödel's Incompleteness Theorem (1931) showed that it is not possible to completely carry out the program described in the first paragraph of this section. Perhaps it is just as well: while it is certainly a worthwhile endeavor to try to put the foundations of mathematics on a sound logical footing, attempting to reduce all of mathematics to mechanical symbol manipulation would rob mathematics of much of its richness, beauty, and usefulness. As Ian Stewart puts it in his book, *Nature's Numbers* [**?**]:

> "Textbooks of mathematical logic say that a proof is a sequence of statements, each of which either follows from previous statements in the sequence or from agreed axioms – unproved but explicitly stated assumptions that in effect define the area of mathematics being studied. This is about as informative as describing a novel as a sequence of sentences, each of which either sets up an agreed context or follows credibly from previous sentences. Both definitions miss the essential point: that both a proof and a novel must tell an interesting story. They do capture a secondary point, that the story must be convincing, and they also describe the overall format to be used, but a good story line is the most important feature of all."

Here are other trenchant observations on the role of proof in mathematics:

> "If a mathematical statement is false, there will be no proofs, but if it is true, there will be an endless variety of proofs, not just one! Proofs are not impersonal, they express the personality of their creator/discoverer just as much as literary efforts do. If something important is true, there will be *many* reasons that it is true, many proofs of that fact. ... And each proof will emphasize different aspects of the problem, each proof will lead in different directions. Each one will have different corollaries, different generalizations ... Mathematical facts are not isolated, they are woven into a vast spider's web of interconnections."
>
> *Gregory Chaitin*[59]

> "Appreciating a theorem in mathematics is rather like watching an episode of *Columbo*: the line of reasoning by which the detective solves the mystery is more important than the identity of the murderer."
>
> *C. Villani*[60]

[59] [Cha05, p. 23-24].
[60] [Vil16, p. 32].

I.6.3. Proofs vs. Formal Proofs (or, How Mathematics Really Works)

> "A proof is a proof. What kind of a proof? It's a proof. A proof is a proof. And when you have a good proof, it's because it's proven."
>
> *Jean Chrétien, Prime Minister of Canada*[61]

I.6.3.1. The standard logical foundation of mathematics is that everything is defined within set theory, beginning with a set of axioms, and that all arguments and proofs are done using the standard rules of logic. Every mathematical statement can be phrased in this way, and every theorem or proposition needs to have a formal proof from the axioms in which every statement follows from the previous ones by an application of one of the rules of formal logic. Free use can be made of previously proved statements (whose proofs technically then become pieces of the proof of the new statement, in the nature of a "subroutine"). In this way, a large edifice is gradually created.

In practice, it is impossibly complicated and tedious to translate all but the simplest mathematical statements into foundational form and write out formal proofs, and most working mathematicians rarely if ever contemplate doing this with their mathematics. In fact, one can say that much of mathematics consists of developing efficient notation and terminology to allow rigorous analysis of complex mathematical ideas without reducing to the foundational level.

> "By relieving the brain of all unnecessary work, a good notation sets it free to concentrate on more advanced problems, and, in effect, increases the mental power of the race."
>
> *A. N. Whitehead*[62]

I.6.3.2. Mathematicians understand the term "proof" to mean something different from "formal proof". Mathematical proofs are in essence outlines of formal proofs, and the usual (although not universal) unwritten rule is that the proof should be written in sufficient detail that a competent (and perhaps masochistic) reader, with sufficient time and effort, could translate it into a formal proof. As in the case of formal proofs, free use can be made of other results whose proofs have already been accepted. (It is a gray area whether the author of a new proof is expected to carefully check the validity of the proofs of other results used.) This standard is, of course, subjective, and also depends in part on the audience; a proof written by a specialist to be read by other specialists can be written more concisely than would be possible if intended for nonspecialists or students. The standard for acceptance of a proof of a new result is generally a consensus of experts that the proof is correct and complete enough to meet the formal translation criterion.

> "The essence of a proof is that it compels belief."
>
> *Pierre de Fermat*[63]

[61] Speaking on the need for evidence of weaponry in Iraq, *CBC News*, Sept. 5, 2002; cf. J. Borwein and D. Bailey, *Mathematics by Experiment: Plausible Reasoning in the 21st Century*, A K Peters, 2004, p. 175.

[62] In P. Davis and R. Hersh, *The Mathematical Experience*, Boston: Birkhäuser, 1981.

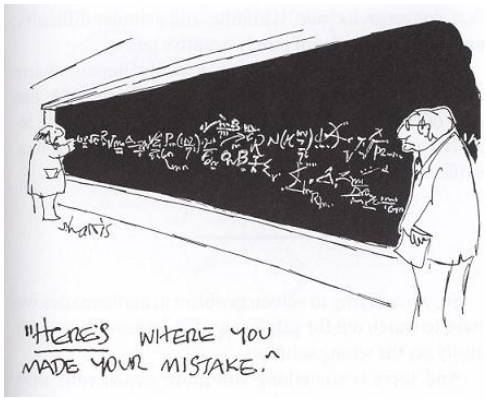
[63] [Wei07, p. 50].

I.6.3.3. It should be noted, however, that mathematicians sometimes use and accept arguments of a type for which there does not seem to be any possibility of a formal translation, so the question of what a mathematical proof *really* is is more complicated than we have suggested here. J. R. Lucas [Luc00] regards a proof as a dialog (the view of a mathematical argument as a dialog goes back at least to Plato; a notable recent proponent was Imre Lakatos [Lak76]. Indeed, the very term "argument" suggests a dialog.) Lucas then considers a formal proof to be the extreme special case when the dialog is held with a "recalcitrant fool," and adds:

> "An argument is not really cogent unless it works even on morons and skeptics."

(Even cogency may not suffice in such a situation!) In the more common case where the discussion is with an intelligent and reasonable fellow truth-seeker, an argument of a different type or level of detail may suffice. See [**?**] and [**?**] for further discussion of the nature of mathematical proofs.

I.6.3.4. There are two types of flaws a purported proof can have. The first is actual errors. An *error* in an argument is a step which is demonstrably false. One error in one step invalidates the entire argument.



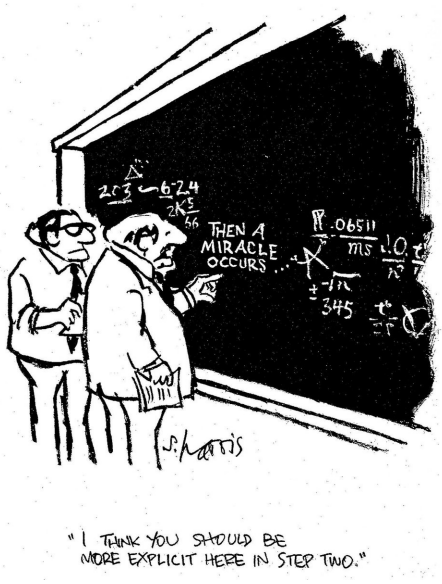
An argument with an error can be *repaired* in either of two ways:

(1) By finding a correct logical path around the error, a valid justification for the conclusions drawn from the erroneous assertion.

(2) By adding extra hypotheses to the statement of the result, in effect restricting to a special case where the erroneous statement is valid.

The first type of repair is preferable when it is possible. Both types of repair must sometimes be used together. (Of course, some errors cannot be repaired!)

I.6.3.5. A more common flaw in "proofs" offered by mathematicians is a gap. A *gap* is a step in the argument which the reader is unable to bridge. Gaps are often in the eye of the beholder: what constitutes a gap for one reader may not be a gap for the author or other readers.



> "I never come across one of Laplace's '*Thus it plainly appears*' without feeling sure that I have hours of hard work before me to fill up the chasm and find out and show *how* it plainly appears."
>
> *N. Bowditch*[64]

The most serious gaps are ones which, once identified, cannot be quickly bridged by either the author or other readers. The gap then must be repaired before the argument can be accepted as a valid proof. A gap can be repaired in either of the two ways described above, or by simply finding an argument which bridges the gap directly (the word "simply" must be taken with a grain of salt; it is often not simple at all to do this!) And some gaps turn out to actually be errors, which cannot always be repaired.

Purported formal proofs can also contain errors or gaps, but these are rather easy to identify. In fact, a computer can be programmed to check the correctness of a formal proof. There is no such definite procedure for mathematical proofs.

I.6.3.6. For a proof to be accepted as valid, there must be general agreement among specialists that it contains no errors or gaps. The process of accepting a proof is thus a somewhat subjective one which is not entirely rigorous; it can happen (but, fortunately, is very rare) that an error or gap is noticed in an argument long after it has been generally accepted as valid. In defense of mathematicians, it must be said that the process almost always works well, and acceptance of a mathematical result is much less dependent on the reputation or force of personality of the author than is the situation in most other academic subjects. And no interesting or useful result can stand the test of time for very long unless it is fully and correctly

[64] Quoted in Cajori, *Teaching and History of Mathematics in the U.S.* (Washington, 1896); cf. [Mor93, 985]. According to W. W. R. Ball [Mor93, 986], "Biot, who assisted Laplace in revising [the *Méchanique Céleste*] for the press, says that Laplace himself was frequently unable to recover the details in the chain of reasoning, and if satisfied that the conclusions were correct, he was content to insert the constantly recurring formula, '*Il est àisé a voir.*' "

established. Standards of rigor have evolved over time: many arguments accepted as proofs as late as the early nineteenth century would not pass muster today, although the statements of the results were rarely wrong (but sometimes exact technical hypotheses were not carefully spelled out, or even understood), and the arguments can usually be filled out to be acceptable as proofs today. Mathematicians as a group are very careful and critical in testing the validity of proofs, especially since a proved statement becomes a "building block" in the "edifice of mathematics" (), and a flawed block could cause a whole section of the building to collapse. Besides, once something in mathematics is really proved[65], it stays proved forever; it can never be invalidated by new theories or points of view. Proofs found by ancient Greeks, when written carefully according to modern standards, are just as good today as they were 2500 years ago!

> "I sometimes think that we worry ourselves overmuch about the enduring consequences of our errors. They may work a little confusion for a time. In the end, they will be modified or corrected or their teachings ignored. The future takes care of such things. In the endless process of testing and retesting, there is a constant rejection of the dross, and a constant retention of whatever is pure and sound and fine."
>
> *Benjamin N. Cardozo*[66]

> "We must not be surprised, however, that many minds apply themselves to other studies [than arithmetic and geometry], or to philosophy. Indeed everyone allows himself more freely the right to make his guess if the matter be dark than if it be clear, and it is much easier to have on any question some vague ideas than to arrive at the truth itself on the simplest of all."
>
> *Descartes*[67]

> "It has long been a complaint against mathematicians that they are hard to convince: but it is a far greater disqualification both for philosophy, and for the affairs of life, to be too easily convinced; to have too low a standard of proof. The only sound intellects are those which, in the first instance, set their standards of proof high. Practice in concrete affairs soon teaches them to make the necessary abatement: but they retain the consciousness, without which there is no sound practical reasoning, that in accepting inferior evidence because there is no better to be had, they do not by that acceptance raise it to completeness."
>
> *J. S. Mill*[68]

[65] I do not just mean here that a proof has been accepted; mathematicians' opinions about what constitutes a proof can be fallible! For contrary views about the sentiment expressed here, see J. Crowe, Ten Misconceptions about Mathematics, in [AK88], p. 267-271.

[66] *The Nature of the Judicial Process*, p. 179.

[67] *Rules for the Direction of the Mind: Torrey's Philosophy of Descartes* (New York, 1892), p. 63; cf. [Mor93, 1628].

[68] *An Examination of Sir William Hamilton's Philosophy* (London, 1878), p. 611; cf. [Mor93, 811].

I.6.4. How to Understand a Proof

I believe I am fairly typical of mathematicians in my approach to understanding a proof I try to read. After reading through any previous explanatory material to be sure I understand the setup, context, notation, and terminology, I first read the statement of the result and try to think of my own proof. I usually fail, especially if it is a research-level result in my field or a result outside my specialty, and I often do not spend a great deal of time trying. If I do succeed, or if I think I have succeeded, I read through the written argument to see if it is the same as mine. Sometimes it is, and I then understand the proof because I have internalized it. Sometimes it is a different argument, and after seeing the general form of the author's argument I put the written proof down and again try to come up with my own argument along those lines (and I also rethink my first argument to see if it is really correct, especially if it is simpler than the written proof.) Sometimes the written argument is basically the same as mine, but treats a technical point which I have overlooked. Occasionally I find that the author has overlooked a technical point which I noticed, and if so I have found a gap or error in the written proof.

It frequently happens that I can identify an overall strategy for proving the result, but there is a crucial step in the proof that I cannot figure out, or I can reduce the statement of the result to another apparently simpler or more elementary statement which I cannot see how to prove. Sometimes, especially if it is a known problem within my specialty, there may even be a well-known strategy or reduction of the problem that I am aware of. In these cases, I home in on the author's treatment of the step I regard as the essential difficulty, or look to see if the author has taken a new or different approach entirely.

In the more common case where I have not been able to get very far in finding my own proof, I look at the written proof and try to identify its overall outline, and then using that I again look for my own proof. There are usually several iterations of this process (each of which can take anywhere from a few minutes to several hours). After a couple of rounds I frequently have most of the argument down, but I am still puzzled by one part (or more) of the argument; I then concentrate the process on the part(s) of the argument I am having difficulty with. At the end of the process I (hopefully) really understand the proof, and the test I use to make sure is whether I can explain the whole argument to someone else with reference to notes only for technical items like formulas or calculations.

> "At least I have got a grip of the essential facts of the case. I shall enumerate them to you, for nothing clears up a case so much as stating it to another person."
>
> *Sherlock Holmes*[69]

> "If you ever hear yourself saying, 'I think I understand this,' that means you don't."
>
> *Richard Feynman*

With experience, you will develop a sense of knowing when you really understand a piece of mathematics and when you don't (although this sense is not infallible!)

[69] *Silver Blaze*, Arthur Conan Doyle.

I.6.5. Proof by Contradiction

> "When you have excluded the impossible, whatever remains, however improbable, must be the truth."
>
> *Sherlock Holmes*[70]

I.6.5.7. One of the commonly used schemes of proof is proof by contradiction, sometimes called *reductio ad absurdum*. If one is trying to prove a statement P, it is assumed that P is false and from this a contradiction or absurdity is obtained; thus P cannot be false and is therefore true. This technique is a variant of proving an implication $P \Rightarrow Q$ by assuming the conclusion Q is false and proving that at least one hypothesis cannot hold, i.e. that P is also false (that is, proving the implication $P \Rightarrow Q$ by proving the contrapositive $(not\ Q) \Rightarrow (not\ P)$); however, a proof by contradiction does not necessarily involve a proof of a contrapositive, at least explicitly.

A good example of a proof by contradiction is the famous ancient proof that $\sqrt{2}$ is irrational, which dates from about 500 BCE. The proof is often attributed to PYTHAGORAS, but is likely due to one of his followers (possibly HIPPASUS; some attribute this proof to ARISTOTLE). We will write the proof in modern algebraic notation, which did not exist in ancient Greece.

I.6.5.8. THEOREM. $\sqrt{2}$ is irrational: there is no rational number r such that $r^2 = 2$.

PROOF: Suppose there is such an r. Then $r = \frac{a}{b}$ with $a, b \in \mathbb{N}$ (the *natural numbers*, or positive integers). By repeatedly dividing out common factors of 2, we may assume a and b are not both even. We have

$$\frac{a^2}{b^2} = 2$$

$$a^2 = 2b^2$$

so a^2, and hence a, is even. Then $a = 2c$ for some $c \in \mathbb{N}$, so

$$a^2 = 4c^2 = 2b^2$$

$$b^2 = 2c^2$$

and b^2, hence also b, is even, contradicting our assumption that a and b are not both even. The actual contradiction is to the assumption that r exists, i.e. that a and b exist, since if they do the reduction to the case where they are not both even follows from basic properties of $\mathbb{N}$ (cf. III.1.7.4.). 👍

See III.1.12.6. and III.1.14.7. for other elementary proofs by contradiction of the irrationality of $\sqrt{2}$.

[70] *Adventure of the Beryl Coronet*, Arthur Conan Doyle.

I.6.5.9. It should be noted that intuitionist mathematicians of the Brouwer school do not accept the Principle of the Excluded Middle, i.e. that (*P or not P*) is a tautology for every statement P, and hence generally reject the validity of proof by contradiction (they do not entirely reject *reductio ad absurdum*, however: for example, the above proof of the irrationality of $\sqrt{2}$ is acceptable to most intuitionists). But almost all mathematicians embrace proof by contradiction as a legitimate method of proof, and agree with the sentiment in the famous statement of G. H. HARDY[71]:

> "*Reductio ad absurdum*, which Euclid loved so much, is one of a mathematician's finest weapons. It is a far finer gambit than any chess play: a chess player may offer the sacrifice of a pawn or even a piece, but a mathematician offers the game."

I.6.5.10. A careful distinction must be made between proof by contradiction and an *impossibility proof*, a proof that a certain mathematical object does not exist. (Theorem I.6.5.8. is an impossibility theorem: "There does not exist a rational number r with $r^2 = 2$.") Proof by contradiction is a technique of proof which can be used to prove any sort of mathematical statement. Impossibility results (of which there are many) are commonly proved by contradiction, but not necessarily so: for example, the standard abstract algebra proof that a general fifth-degree polynomial is not solvable by radicals is usually expressed as a direct argument, not a proof by contradiction. Fermat's Last Theorem is another example (although the proof there is very deep). See I.6.6.10. for more on impossibility results.

I.6.5.11. Most, if not all, proofs by contradiction can in principle be converted into direct proofs, but it may be very clumsy and unnatural to do so in practice. It is also true that almost any direct mathematical statement can be converted into an impossibility statement, but also often in a clumsy and unnatural way.

I.6.5.12. My experience is that beginning students tend to overuse proof by contradiction. It is usually good advice to look for a direct proof first and only use proof by contradiction as a last resort, unless one happens to notice an obvious and simple contradiction argument. As a rule, the only types of mathematical statements where it is worthwhile to begin by looking for a proof by contradiction are impossibility statements, or a statement of the form $\forall x(P(x))$, which is an impossibility statement in disguise ("There does not exist an x for which $P(x)$ is false.").

Using a Cannon to Kill a Fly

I.6.5.13. As a (slightly) related matter, when proving something one should try not to use more powerful results than necessary. This is important for at least three reasons:

(i) Since a main purpose of a proof is to explain why the statement is true, one much prefers proofs where all steps can be easily understood, not where a crucial step depends on a deep or difficult theorem.

(ii) The overall philosophy of mathematics is that it is gradually built up from elementary results to more advanced ones, and using advanced results to prove elementary ones interferes with the architecture of the edifice.

(iii) When using advanced results whose proofs have not been carefully examined to conclude elementary consequences, there is a serious danger of circularity.

[71] *A Mathematician's Apology*, London, Cambridge University Press, 1941.

Of course, some apparently elementary statements (e.g. Fermat's Last Theorem, the Four Color Theorem) turn out to not be elementary at all, and advanced arguments may be necessary.

I.6.5.14. Exercise. (a) Give a proof that $\sqrt[3]{2}$ is irrational along the lines of I.6.5.8..
(b) Critique the following proof which appeared in an online text for an Introduction to Proofs class:
PROOF: Suppose $\sqrt[3]{2} = \frac{a}{b}$ for natural numbers a and b. Then

$$\frac{a^3}{b^3} = 2$$

$$a^3 = 2b^3 = b^3 + b^3$$

contradicting Fermat's Last Theorem. 👍

(Granted, the special case of FLT with $n = 3$ is not nearly as deep as the general case, but it's still not easy, much harder than for $n = 4$.)

You Have to Start Somewhere

> Chairman of a meeting of the Society of Logicians: "Before we put the motion: 'That the motion be now put,' should we not first put the motion: 'That the motion: "That the motion be now put" be now put'?"
>
> *Punch Magazine*[72]

[72] [Gar01, p. 315].

I.6.6. Proofs and Counterexamples: Mathematics as Creative Activity

> "All truths are easy to understand once they are discovered; the point is to discover them."
>
> *Galileo*

I.6.6.1. As important as it is to be able to understand proofs written by other mathematicians, the true essence of mathematics is the ability to solve problems. Mathematical problems range from very specific to completely general and open-ended, and from straightforward to intractable (the two scales, by the way, are largely uncorrelated; very specific problems can be, or seem, intractable). Solutions to mathematical problems can range from simple numerical calculations or straightforward applications of standard mathematical results up to remarkably intricate, clever, and ingenious creative new arguments which can spawn entire new areas of mathematics. The best solutions to even simple problems are ones which are stimulating and thought-provoking beyond the problem at hand.

> "In mathematics resolving of important specific questions and development of an abstract general theory are processes as inseparable as inhaling and exhaling."
>
> *V. Zorich*[73]

> "A naive non-mathematician ... understandably concludes [from formal presentations] that in mathematics, axioms come first. ... But anyone who has done mathematics knows what comes first – a problem. Mathematics is a vast network of interconnected problems and solutions. ... Sometimes a solution is a set of axioms! ... When a piece of mathematics gets big and complicated, we may want to systematize and organize it, for esthetics and for convenience. The way we do that is to axiomatize it. Thus a new type of problem (or 'meta-problem') arises:
>
> > 'Given some specific mathematical subject, to find an attractive set of axioms from which the facts of the subject can conveniently be derived.'
>
> Any proposed axiom set is a proposed solution to this problem. The solution will not be unique. ... In developing and understanding a subject, axioms come late. Then in the formal presentations, they come early."
>
> *Reuben Hersh*[74]

> "Mathematical statements, mathematical theorems, are answers to questions. ... The mystery of how mathematics grows is in part caused by looking at mathematics as answers without questions. That mistake is made only by people who have had no contact with mathematical life. It's the questions that drive mathematics. Solving problems and making up new ones is the essence of mathematical life."
>
> *Reuben Hersh*[75]

[73] [Zor09, p. V]
[74] *What is Mathematics, Really?*, p. 6.
[75] *What is Mathematics, Really?*, p. 18.

I.6.6.2. One of the most common kinds of mathematical problems is: Is statement P true? True, of course, means *always* true: the conclusion holds in every situation in which the hypotheses are satisfied. A solution to a problem of this sort consists of a proof or a disproof of the statement. Note that a statement cannot be proved by verifying that it is true in one or more examples unless *every* example in which the hypotheses hold is checked (which is indeed sometimes feasible in specific problems); at best, verifying the statement in some examples provides evidence that the statement may be true in general. However, one example where the hypotheses are satisfied but the conclusion does not hold *disproves* the statement (proves that the statement is false). Such an example is called a *counterexample* for the statement. It is possible in theory, and occasionally in practice, to disprove a statement without finding a specific counterexample; but a "smoking gun" counterexample is usually the simplest, and often the most satisfying, way to disprove a statement.

> "The mathematician as the naturalist, in testing some consequence of a conjectural general law by a new observation, addresses a question to Nature: 'I suspect that this law is true. Is it true?' If the consequence is clearly refuted, the law cannot be true. If the consequence is clearly verified, there is some indication that the law may be true. Nature may answer Yes or No, but it whispers one answer and thunders the other, its Yes is provisional, its No is definitive."
>
> *G. Polya*[76]

Finding a Solution

> "It will seem not a little paradoxical to ascribe a great importance to observations even in that part of the mathematical sciences which is usually called Pure Mathematics, since the current opinion is that observations are restricted to physical objects that make impression on the senses. As we must refer the numbers to the pure intellect alone, we can hardly understand how observations and quasi-experiments can be of use in investigating the nature of the numbers. Yet, in fact, as I shall show here with very good reasons, the properties of the numbers known today have been mostly discovered by observation, and discovered long before their truth has been confirmed by rigid demonstrations. There are even many properties of the numbers with which we are well acquainted, but which we are not yet able to prove; only observations have led us to their knowledge. Hence we see that in the theory of numbers, which is still very imperfect, we can place our highest hopes in observations; they will lead us continually to new properties which we shall endeavor to prove afterwards. The kind of knowledge which is supported only by observations and is not yet proved must be carefully distinguished from the truth; it is gained by induction, as we usually say. Yet we have seen cases in which mere induction led to error. Therefore, we should take great care not to accept as true such properties of the numbers which we have discovered by observation and which are supported by induction alone. Indeed, we should use such a discovery as an opportunity to investigate more exactly the properties discovered and to prove or disprove them; in both cases we may learn something useful."
>
> *L. Euler*[77]

[76] [Pól90a, p. 10].
[77] *Opera Omnia*, ser. 1, vol. 2, p. 459; cf. [Pól90a, p. 3].

"I have the result, but I do not yet know how to get it."

C. F. Gauss[78]

I.6.6.3. In order to solve the problem of whether statement P is true or not, one must usually first form an opinion about the answer, which when firmly held and/or communicated to others is called a *conjecture*. The next step is to try to verify the conjecture by either looking for a proof (if the conjecture is that P is true) or a counterexample (if the conjecture is that P is false). The search often involves doing both things to some extent: obstacles encountered in trying to prove P may suggest a possible counterexample, and on the other hand inability to construct a counterexample can sometimes suggest a method of proof of P.

I.6.6.4. Although mathematics is (for the most part) not an experimental science, there is a good analogy with experimental sciences: in both mathematics and experimental sciences, people form a conjecture (called a "theory" in sciences) based at least in part on evidence from observations or experiments. Then either the creator of the conjecture (theory) or someone else "tests" the conjecture (theory) with "experiments." A well-designed set of experiments, or candidates for counterexamples in mathematics, provide the most critical tests for consequences or predictions of the conjecture (theory): if the conjecture (theory) "passes the test", it can be deemed worthy of being taken seriously. Of course, the biggest difference between mathematics and experimental sciences is that theories can never be proved in science (at least not in the same sense as proofs in mathematics), and acceptance of a theory in science is based only on believability, observational or experimental evidence, accuracy of prediction, and/or lack of experimental refutation; such acceptance is always subject to revision in light of new evidence.

I.6.6.5. One need not, and often should not, stop when the problem at hand is solved. It is a much more significant accomplishment to come up with a general scheme for solving a class of related problems, so that new problems as they arise can also be solved more or less mechanically without a similar investment of time and effort. This is the origin of new subdisciplines in mathematics. (One should not be under the illusion, however, that all of mathematics can eventually be reduced to a set of mechanical processes or algorithms; there will always be room for creative genius in mathematics.)

> "In general the position as regards all such new calculi is this – That one cannot accomplish by them anything that could not be accomplished without them. However, the advantage is, that, provided such a calculus corresponds to the inmost nature of frequent needs, anyone who masters it thoroughly is able – without the unconscious inspiration of genius which no one can command – to solve the respective problems, yea, to solve them mechanically in complicated cases in which, without such aid, even genius becomes powerless. Such is the case with the invention of general algebra, with the differential calculus, and in a more limited region Lagrange's calculus of variations, with my calculus of congruences, and with Möbius's calculus. Such conceptions unite, as it were, into an organic whole countless problems which would otherwise remain isolated and require for their separate solution more or less application of inventive genius."

C. F. Gauss[79]

[78] Isaac Asimov and J. A. Shulman, ed., *Isaac Asimov's Book of Science and Nature Quotations*, Weidenfeld and Nicolson, London, 1998, p. 115.

[79] *Werke*, Bd. 8, p. 298; cf. R. Moritz, *Memorabilia Mathematica*, 1215.

"How thoroughly it is ingrained in mathematical science that every real advance goes hand in hand with the invention of sharper tools and simpler methods which, at the same time, assist in understanding earlier theories and in casting aside some more complicated developments."

D. Hilbert

"... [M]ath isn't the art of answering mathematical questions, it is the art of asking the right questions, the questions that give you insight, the ones that lead you in interesting directions, the ones that connect with lots of other interesting questions – the ones with beautiful answers!"

Gregory Chaitin[80]

"The outcome of any serious research can only be to make two questions grow where only one grew before."

Thorstein Veblen[81]

Is There a Solution?

I.6.6.6. One question which must be faced, at least in passing, in any mathematical problem is: does the problem have a solution at all? Indeed, in some mathematical problems it is not entirely clear what one would exactly mean by a "solution," and part of formulating such a problem well is deciding what this should mean.

I.6.6.7. Even if the problem is well-formed and it is obvious what a "solution" would be, it is often not obvious whether there is one. In fact, in many problems the most difficult part of the problem is establishing that there is a solution: if a solution is known to exist, it can sometimes be found relatively easily from the properties it must have. And sometimes knowing the mere existence of a solution can itself have important consequences. Thus considerable mechanisms have been developed in mathematics to establish the existence of solutions to certain kinds of problems:

The notions of *completeness* and *compactness* can be thought of as properties insuring the existence of solutions to problems involving limits, maxima, and minima (although not all such problems!)

The Axiom of Choice is a set-theoretic assumption insuring the existence of certain kinds of sets and functions.

I.6.6.8. On the other hand, it can sometimes be shown that a certain problem has no solution. This can also have important consequences quite beyond simply letting mathematicians avoid wasting time and effort looking for one. Proving that a problem has no solution is a type of impossibility result (). Actually, though, establishing nonexistence of a solution is often not the end of the story: such results sometimes lead to important expanded theories in which there is a solution. Here are some examples where the absence of a solution led to important advances:

[80] [Cha05, p. 24].
[81] *The Place of Science in Modern Civilization and Other Essays.*

(i) The equation $x^2 = 2$ has no solution in the set of rational numbers (). This is one of the examples which eventually led to the study of irrational numbers and formulation of the theory of real numbers, which underlies all of analysis.

(ii) The equation $x^2 = -1$ has no solution in the set of real numbers. The complex numbers were developed specifically to provide a solution to this equation; they then turned out to be important and useful beyond the wildest dreams of the mathematicians who first considered them.

(iii) The "Dirac delta function" is a function δ on the real numbers $\mathbb{R}$ such that $\delta(x) = 0$ for all $x \neq 0$ and such that $\int_{-\infty}^{+\infty} \delta(x)\, dx = 1$. Such a "function" would have many uses in both pure and applied mathematics. But it is easily proved that there is no such function. This example led to the theory of generalized functions and distributions (), where δ exists as a generalized function and, with care, can be used in many of the applications.

> "Remember that when the usual road does not lead to success, one should not be content with this determination of impossibility, but should bestir oneself to find a new and more promising route. Mathematical thought, as such, has no end. If someone says to you that mathematical reasoning cannot be carried beyond a certain point, you may be sure that the really interesting problem begins precisely there."
>
> *F. Klein*[82]

I.6.6.9. One must be careful about assuming that simply stated problems either have relatively simple solutions or none at all. For example, does the equation

$$x^2 - 1141y^2 = 1$$

have any solutions in positive integers? An exhaustive search through reasonably, semi-reasonably, and even not-so-reasonably sized pairs of positive integers would not turn up a solution, so one could be led to believe there is no solution. However, it turns out that there is a solution, in fact infinitely many, the smallest of which is

$$x = \ 1{,}036{,}782{,}394{,}157{,}223{,}963{,}237{,}125{,}215 \ \approx 10^{27}$$

$$y = \ 30{,}693{,}385{,}322{,}765{,}657{,}197{,}397{,}208 \ \approx 3 \times 10^{25} \ .$$

This is an example of Pell's equation (); see e.g. [Sta78] for a discussion.

Another example is ARCHIMEDES' cattle problem (see [Var98] for a formulation and nice discussion). This is a problem which can be expressed in words without use of mathematical symbols, yet the smallest solution is approximately 7.76×10^{206544}.

> "At any moment there is only a fine layer between the 'trivial' and the impossible. Mathematical discoveries are made in this layer."
>
> *A. Kolmogorov*[83]

[82] [Kle04, p. 143]. Note that this was written before GÖDEL; nonetheless KLEIN'S point is well taken.
[83] [**?**, p. 83].

Unsolved vs. Unsolvable

I.6.6.10. It is worth a digression to discuss an endemic problem in mathematics which continues to plague mathematicians: misunderstanding the difference between *unsolved* and *unsolvable*.

Many mathematical problems are unsolved; some mathematical problems are unsolvable. The classes are quite distinct. (There is a nonempty overlap: some unsolved problems will undoubtedly turn out to be unsolvable, although "most" will not.) What many nonmathematicians do not understand is that many mathematical problems, including some famous ones, have been *proved* to be unsolvable; thus they cannot be described as "unsolved." In fact, impossibility theorems are precisely solutions to "unsolvable" problems. Probably the most famous such problem is trisecting an angle, or more precisely giving a general procedure for trisecting a given angle with straightedge and compass, one of the "big three" problems from ancient Greece. The other two ancient Greek problems, almost as famous, are "squaring the circle" (constructing with straightedge and compass a square with the same area as a given circle, cf. [Bla97]) and "doubling the cube" (constructing with straightedge and compass a cube whose volume is twice the volume of a given cube). These problems were unsolved until the nineteenth century, when some abstract algebra now regarded as elementary (covered in undergraduate abstract algebra courses) was developed which proved that the problems are unsolvable, i.e. there can be no such procedure (the transcendence of π, or something similar, is also needed for the circle squaring problem, cf. III.9.3.2.).

Mathematicians and mathematics departments regularly receive purported "solutions" to these problems, especially angle trisection. These arguments are often complicated and clever, but of course must necessarily contain a mistake. Mathematicians are understandably extremely reluctant to make the effort to find the inevitable mistakes, and thus tend to ignore or summarily dismiss such communications, leading to bad public relations and charges that mathematicians are closed-minded. Some mathematics departments have actually assigned a faculty member to respond to them for public relations purposes. (Physicists and engineers face something similar with proposed perpetual-motion machines.) The people who submit such attempted solutions are generally not crackpots (although there are some of those – see [Dud92] and Figure I.13), but intelligent people who know enough mathematics to be able to formulate the problems and discuss them but not enough to really understand them. In mathematics, a little knowledge can be a dangerous thing.

The most widespread source of confusion is that the problems are often slightly misstated so they sound unsolved, such as: "No one has ever found a procedure to trisect an angle with straightedge and compass." This is a true statement. But it does *not* mean the problem is unsolved! It is also a true statement that "No one has ever found a rational number x such that $x^2 = 2$." And on a simpler level: "No one has ever found an integer x such that $2x = 5$." No intelligent person would waste time trying to find such an integer, as it is elementary to see there is none. The problem is solved, proved to be unsolvable. Understanding why the trisection problem has no solution is not elementary enough that a person without the requisite knowledge of field theory could be expected to come up with it (even Euler could not), but to a modern mathematician it is quite as impossible to trisect angles as it is to solve the $2x = 5$ problem.

Here are some examples, in roughly increased order of sophistication. These are unsolvable problems, not unsolved problems (the last two were solved in the twentieth century):

There is no integer x with $2x = 5$.

There is no rational number x with $x^2 = 2$.

There is no procedure to trisect angles with straightedge and compass.

There is no procedure to double a cube with straightedge and compass.

There is no procedure to square a circle with straightedge and compass.

There is no algebraic formula for the roots of a general fifth-degree polynomial.

There is no elementary formula for $\int e^{-x^2}\,dx$ (XII.12.3.18.). [But this does *not* mean there is no antiderivative!]

The Axiom of Choice and the Continuum Hypothesis are neither provable nor disprovable in ZF set theory.

There are no natural numbers x, y, z, n with $n > 2$ for which $x^n + y^n = z^n$.

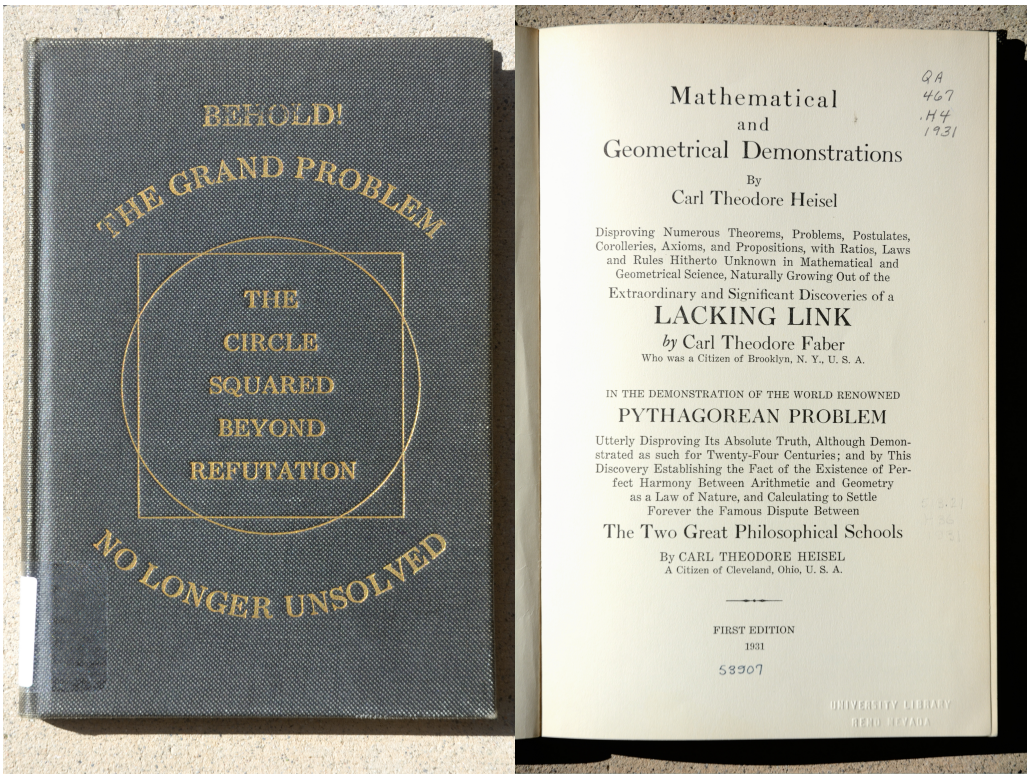


Figure I.13: A famous crackpot publication. The author claims that $\pi = \left(\frac{4}{3}\right)^4 = 3\frac{13}{81} \approx 3.16$. Heisel claims to be named after Faber – were they relatives? Circle-squaring disease is called *morbus cyclometricus* in [**?**].

The assertion that $\pi = \left(\frac{4}{3}\right)^4 = 4\left(\frac{8}{9}\right)^2$ was obtained in ancient Egypt (Rhind Papyrus, circa 1650 BCE); cf. e.g. [Bec71, p. 22–25] or [Bla97]. Things could be worse: there are passages in the Bible (I Kings 7:23 and II Chronicles 4:2) which can be interpreted to say that $\pi = 3$, so mathematicians could be up against religious dogma as much as evolutionary biologists are. (I have an opinion why this has not happened, which I think I will not commit to print.) Note that mathematics is not immune from religious dogma: the theory that a line is composed of a collection of indivisible points came under withering attack by the Jesuits in the sixteenth century as contradicting the Eucharistic doctrine of transubstantiation (cf. I.3.1.2., [She18]).

Mathematics by Amateurs

I.6.6.11. New and significant mathematics can indeed be done by "amateurs," people who have little formal mathematical training and/or are unknown in the mathematical community. This was quite common in earlier times, when the state of mathematical knowledge was less extensive and the distinction between amateur and professional mathematicians was hazy. The most famous example in the twentieth century was RAMANUJAN. Even today research papers by undergraduates and unknown mathematicians sometimes appear in leading journals (indeed, every research mathematician must begin with a first paper!) For the most part, professional mathematicians are quite open-minded about the possibility of worthwhile research produced by unknowns, and judge mathematical work primarily on its intrinsic quality rather than the fame or reputation of the author.

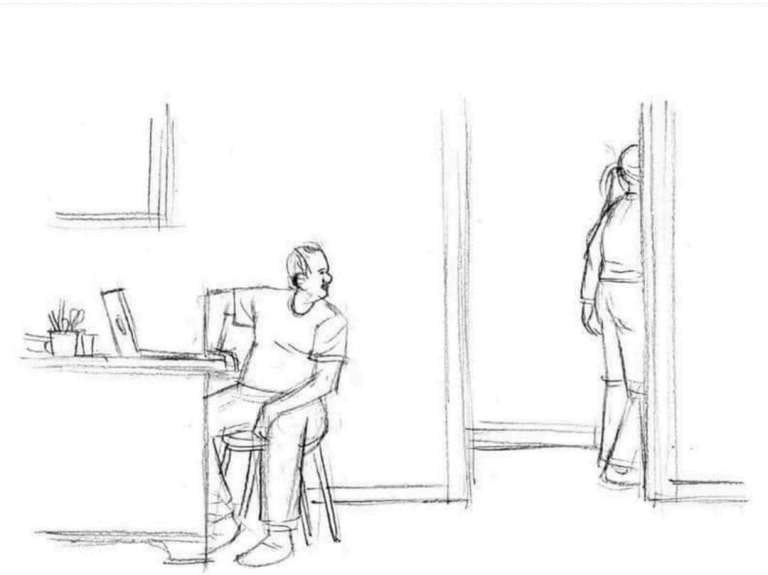
I.6.6.12. However, a very important qualification must be added. Mathematics is not produced by dilettantes. Mathematical research is serious business and requires a lot of deep, creative thought. It is especially unlikely that an amateur can solve a famous problem (the problems amateurs typically try). No one should think they can contribute to mathematical knowledge by casual thinking about famous problems. In addition, any problem known to amateurs has been thought about and worked on by many very smart and knowledgable professionals, sometimes for decades or centuries, so the possibility that an amateur could come up with a crucial new idea is remote.

> "It would be rash to say that nothing remains for discovery or improvement even in elementary mathematics, but it may be safely asserted that the ground has been so long and so thoroughly explored as to hold out little hope of profitable return for a casual adventurer."
>
> *Isaac Todhunter*[84]

I.6.6.13. KLEIN talked in the early twentieth century about "Fermat's Last Theorem," which was eventually proved in 1995 (using sophisticated ideas no amateur could have produced):

> "It will be clear ... how difficult the winning of this prize [the Wolfskehl prize for proving Fermat's Last Theorem] must seem to the mathematician, who understands the situation and who knows what efforts have been made by Kummer and his successors to prove the theorem. But the *great public* thinks otherwise. Since the summer of 1907 ... we have received a prodigious heap of alleged proofs. People of all walks of life ... have shared in these contributions. The common thing about them all is that they have *no idea of the serious mathematical nature of the problem.* Moreover, they have made no attempt to inform themselves regarding it, but have trusted to finding the solution by a sudden flash of thought, with the inevitable result that their work is nonsense. One can see what absurdities are brought forth if one reads the numerous critical discussions of such proofs ... It is amusing to read these wholesale slaughterings, sad as it is that they are necessary." [Emphasis original]
>
> *F. Klein*[85]

[84] *Private Study of Mathematics*, p. 73; cf.[Mor93, #617].
[85] [Kle04, p. 48-49].



"Honey, come look! I've found some information all the world's top scientists and doctors missed."

Determination and Hard Work

> "It would be a mistake to think that solving problems is a purely 'intellectual affair'; determination and emotions play an important role. Lukewarm determination and sleepy consent to do a little something may be enough for a routine problem in the classroom. But, to solve a serious scientific problem, will power is needed that can outlast years of toil and bitter disappointment."
>
> *G. Polya*[86]

I.6.6.14. One psychologically difficult characteristic of the process of solving mathematical problems is that one can work hard for a long time without apparent progress, and even disappointing regress when mistakes are discovered in previous work. But eventually (not always, but often enough), everything suddenly comes together in an "aha!" moment. This moment can even occur while not actively thinking about the problem; after a lot of thought and concentration on a problem the mind continues to mull it over subconsciously even during other activities. When the moment occurs, we can be fooled into thinking that all the previous effort without progress was wasted and could have been avoided if we had just been smarter about the problem, but the fact is that we would have been extremely unlikely to come up with the right idea without all the preceding thought and effort.

[86] *How to Solve It*, p. 93.

"Once you solve a problem, you tend to remember only the short path that got you from A to B. You forget all the dead ends. It's a bit of a shame. It gives the wrong impression that people who are good at mathematics only choose the right steps. But there is a lot of trial and error and really terribly embarrassing ideas. Sometimes there is a 'eureka' moment, but it's more of a hitting-your-head moment: 'Of course, why was I so dense?' "

Terry Tao[87]

"Feeling stupid for not having understood something before just shows that you are now cleverer than you were then."

Eugenia Cheng[88]

EDISON'S famous statement that "genius is 1% inspiration and 99% perspiration" may be an exaggeration, but contains a good deal of truth. And there is nothing more uplifting for a mathematician than that moment of understanding, which makes (or should make) all the previous work worthwhile.

"Every mathematician worthy of the name has experienced ... the state of lucid exaltation in which one thought succeeds another as if miraculously ... this feeling may last for hours at a time, even for days. Once you have experienced it, you are eager to repeat it but unable to do it at will, unless perhaps by dogged work ... "

André Weil[89]

This does not just happen in mathematics: for example, athletes often describe being "in the zone" and experiencing something similar. But long hours of previous training are necessary preparation.

I.6.6.15. One song I can really relate to is JOHN DENVER'S "Looking for Space." Although he was not a mathematician, and probably knew little about mathematics or the process of doing it, he beautifully captured the emotional roller-coaster mathematicians can deal with in their work:

"On the road of experience, I'm trying to find my own way.
Sometimes I wish that I could fly away
When I think that I'm moving, suddenly things stand still
I'm afraid 'cause I think they always will

And I'm looking for space
And to find out who I am
And I'm looking to know and understand
It's a sweet, sweet dream
Sometimes I'm almost there
Sometimes I fly like an eagle
And sometimes I'm deep in despair

[87] *Princeton Alumni Weekly*, 11/13/19, p. 32.
[88] [Che15, p. 30].
[89] *The Apprenticeship of a Mathematician.*

All alone in the universe, sometimes that's how it seems
I get lost in the sadness and the screams
Then I look in the center, suddenly everything's clear
I find myself in the sunshine and my dreams

And I'm looking for space
And to find out who I am
And I'm looking to know and understand
It's a sweet, sweet dream
Sometimes I'm almost there
Sometimes I fly like an eagle
And sometimes I'm deep in despair

On the road of experience, join in the living day
If there's an answer, it's just that it's just that way
When you're looking for space
And to find out who you are
When you're looking to try and reach the stars
It's a sweet, sweet sweet dream
Sometimes I'm almost there
Sometimes I fly like an eagle
But sometimes I'm deep in despair
Sometimes I fly like an eagle, like an eagle
I go flying flying."

The trick a mathematician must learn is to avoid despair during the dry periods, and persevere to savor the moments of insight and enlightenment.

> "Perhaps I can best describe my experience of doing mathematics in terms of a journey through a dark unexplored mansion. You enter the first room of the mansion and it's completely dark. You stumble around bumping into the furniture, but gradually you learn where each piece of furniture is. Finally, after six months or so, you find the light switch, you turn it on, and suddenly it's all illuminated. You can see exactly where you were. Then you move into the next room and spend another six months in the dark. So each of these breakthroughs, while sometimes they're momentary, sometimes over a period of a day or two, they are the culmination of, and couldn't exist without, the many months of stumbling around in the dark that precede them."
>
> *Andrew Wiles*[90]

[90] Quoted in interview for the PBS TV Nova program, 'The Proof'.

"Doing research in mathematics is frustrating and if being frustrated is something you cannot get used to, then mathematics may not be an ideal occupation for you. Most of the time one is stuck, and if this is not the case for you, then either you are exceptionally talented or you are tackling problems that you knew how to solve before you started. There is room for some work of the latter kind, and it can be of a high quality, but most of the big breakthroughs are earned the hard way, with many false steps and long periods of little progress, or even negative progress."

Peter Sarnak[91]

"Thought is only a flash between two long nights, but this flash is everything."

Henri Poincaré[92]

"Mathematical discoveries, small or great, are never born of spontaneous generation. They always presuppose a soil seeded with preliminary knowledge and well prepared by labour, both conscious and subconscious."

Henri Poincaré[93]

"Well, some mathematics problems look simple, and you try them for a year or so, and then you try them for a hundred years, and it turns out that they're extremely hard to solve."

Andrew Wiles[94]

"Unfortunately, mathematics is often taught as a mere matter of learning instructions and following them. You may have the habit of not trying anything in mathematics until you already 'know what to do,' until you feel that you 'understand' the problem. If this is your habit, try to overcome it as quickly as possible, for it will be the main obstacle to your progress."

Kenneth O. May[95]

You don't really know what you can do until you try!

"*How does one discover how to prove a theorem?* ... You must proceed as any person does who is confronted with a problem for which there are no set rules. ... half the battle will be won if you accept the fact that you are not supposed to know in advance how to prove theorems. *The problem is not to know, but to discover.*"

Kenneth O. May[96]

"In research, if you know what you are doing, then you shouldn't be doing it."

R. Hamming[97]

[91] *The Princeton Companion to Mathematics*, p. 1009.
[92] In J. R. Newman (ed.) *The World of Mathematics*, New York: Simon and Schuster, 1956.
[93] *Science and Hypothesis.*
[94] http://www.brainyquote.com/quotes/keywords/mathematics.html
[95] *Fundamental Mathematics*, preliminary edition, 1955, p. v.
[96] *Fundamental Mathematics*, preliminary edition, 1955, p. 107-108.
[97] [Ham86, p. 704].

I.6.7. Scholarship vs. Practice

For every mathematician, from student to leading research mathematician, there is a constant tradeoff between studying the work of others and working on one's own problems. Both are necessary for growth and development as a mathematician, although the relative emphasis can vary widely: some mathematicians are primarily scholars of mathematics, with vast knowledge of the work done almost entirely by others, while some of the most brilliant mathematicians have surprisingly sparse knowledge of others' work, preferring to solve most problems arising in their work themselves as they come (this is definitely the exception rather than the rule among top mathematicians, however). For most of us, study of the work of others provides us not only with inspiration, but also a vast supply of machinery we can use in our own work.

According to P. Sarnak [Vil16, p. 115-116], Paul Cohen did not believe mathematics progresses incrementally, but rather by sudden great leaps, and therefore he strongly discouraged using, or even reading, related previous work on one's problems. This is roughly the "Moore method" () carried to its logical extreme. Although much good can be said about the Moore method, and there are even some positives about Cohen's attitude (assuming it is accurately portrayed; indeed, Cohen's own most famous result used his careful analysis of previous work of A. Fraenkel), and although Cohen was unquestionably an outstanding mathematician, I think this attitude is almost completely wrong: some of the biggest advances in mathematics do indeed come out of the blue, but most do not, and it is generally a very bad idea for a research mathematician to be ignorant (willfully or otherwise) of other work in the area. For one famous example, how can it be explained that calculus was simultaneously developed by two people in the latter 17th century? It is not adequate to simply say that two of the greatest mathematical minds of all time coincidentally did this at the same time; this is arguably a true statement, but it is also true that the mathematical terrain and atmosphere at the time was ripe for the development. Calculus did not suddenly arise out of nowhere, and most of the ideas and techniques of calculus already existed in some form; the main contribution of Newton and Leibniz (which I do not at all minimize) can be accurately described as largely "putting the pieces together." In fact, if Cohen was right, worthwhile collaborative research in mathematics would be impossible.

Attempting to do mathematical research as a "lone wolf" has at least two related drawbacks: (1) You often end up "reinventing the wheel" due to inadequate familiarity with other work, although sometimes the new solutions are better than the previous ones and make an important contribution to the subject; (2) Especially nowadays when so much mathematical research is being done in such varied areas, unless you are an unusually great mathematician there is a good chance that someone at least as good as you are has already thought extensively about the problem or a closely related one, and can impart at least some words of wisdom, if not useful techniques or partial results.

A (rather imperfect) analogy with medicine is perhaps useful. The primary goal of medicine is to treat patients. To do this properly, a practicing doctor must not only have a rigorous medical school training, but must regularly keep up with medical journals and other research reports. On the other hand, there is a place, even a necessity, for some doctors to spend most of their time doing laboratory research, or instruction at medical schools and hospitals; however, almost all research or teaching physicians "keep their hand in practice" by treating at least a few patients.

I.6.8. Is Mathematics Created or Discovered?

One of the most basic controversies in the philosophy of mathematics is the question of whether working mathematicians actually create new mathematics which did not exist before, or whether they simply discover mathematical truths which already existed abstractly, somewhere (or whether it even makes sense to ask

these questions.) Belief in the *a priori* existence of mathematical "objects" is referred to as *realism* (the term *Platonism* is also often used, but Platonism is technically only one form of realism).

One popular version of the question is: if there is an intelligent race of aliens on another planet, would they develop mathematics, and if so would it bear any resemblance to the mathematics developed so far by humans? If mathematics is indeed universal, then it should be essentially the same everywhere in the universe.

> "Beethoven and Shakespeare did unique work, and we would not have *Fidelio* or *King Lear* if they had not lived. However, symmetric functions would have been discovered by others if Newton had not lived."
>
> *Richard Askey*[98]

My own view leans more towards the realist philosophy. Mathematical results are not arbitrary; the (provable) truth of a mathematical statement does not depend on the opinions or desires of the mathematicians who study it. Two mathematicians studying the same problem, making the same assumptions, cannot arrive at contradictory solutions (provided that the assumptions they make are not self-contradictory).

> "[N]umbers (and all mathematical ideas, for that matter) have lives of their own. We can't control them. Even though they exist in our minds, once we decide what we mean by them we have no say in how they behave. They obey certain laws and have certain properties, personalities, and ways of combining with one another, and there's nothing we can do about it except watch and try to understand."
>
> *S. Strogatz*[99]

There is a perceptive analogy in [**?**]: an artist can mix together any colors of paint in any proportions they choose, but once this is done they have no control over what mixed color results – it is what it is. (Through experience they learn what mixtures give what they want.)

The importance of a new (or newly discovered) piece of mathematics is a much more subjective matter, and there can very well be differences of opinion among mathematicians; however, as time goes on there almost always seems to develop a remarkable consensus of opinion about what is significant and what is the "right" approach to an area of mathematics. This seems to provide evidence that there is some pre-existing mathematics either outside or deep within the human intellect. However, it seems likely that the exact form of the mathematics considered in the human mind or communicated between humans is dependent on specific human characteristics as well as on human history and culture (this is obviously true at least on the superficial level of notation and terminology, but probably goes much deeper). The "continent of mathematics" analogy () thus seems appropriate.

Realism as a philosophy of mathematics has definitely been unfashionable among philosophers. However, among mathematicians it is probably the predominant view. (Philosophers tend to be dismissive of the opinions of mathematicians on this issue as unreliable, unsophisticated, and/or uninformed, not entirely without justification.) Mathematicians generally have a strong feeling that the abstract objects they work with are "real," although this could of course be an illusion. This feeling is reflected in the language mathematicians use to describe advances in mathematics, almost invariably referring to "discovering" or

[98] R. Askey, *How can mathematicians and mathematical historians help each other?* [AK88, p. 202]. This is just one version of a line often used with many variations.

[99] [Str13, p. 4].

"finding" new results or proofs. The term "invent" is rarely used, and when it is (e.g. "Newton and/or Leibniz invented calculus") it usually refers to the development of a new subdiscipline or point of view in mathematics, which can be viewed as a way of facilitating analysis of the structure of part of the "continent of mathematics" rather than creation of new mathematics *per se*.

> "[M]ath always involves both invention *and* discovery: we invent the concepts but discover their consequences. ... in mathematics our freedom lies in the questions we ask – and in how we pursue them – but not in the answers awaiting us."
>
> *S. Strogatz*[100]

> "One way of thinking about axioms is that you're not saying they *are* true, you're just creating a world in which they are true, and studying it to see what will happen."
>
> *E. Cheng*[101]

I.6.9. Analogies for the Structure of Mathematics

There are at least three analogies which have been used to describe the subject of mathematics. All are useful, if of course vast oversimplifications; the differences between them reflect different philosophies of the nature of mathematics (we will not attempt to discuss such philosophies here; see I.7.1.).

The Edifice of Mathematics

This picture or analogy was popularized by Descartes (who introduced the term "edifice"), but it may well not have originated there. It is probably the most widely-used picture of mathematics, although it primarily reflects a formalist or logicist philosophy:

> "The discovery that all mathematics follows inevitably from a small collection of fundamental laws is one which immeasurably enhances the intellectual beauty of the whole; to those who have been oppressed by the fragmentary and incomplete nature of most existing chains of deduction this discovery comes with all the overwhelming force of a revelation; like a palace emerging from the autumn mist as the traveller ascends an Italian hill-side, the stately storeys of the mathematical edifice appear in their due order and proportion, with a new perfection in every part."
>
> *Bertrand Russell*[102]

If this view of mathematics is accepted, even partially, one must eventually consider what the building stands on and whether it has a solid foundation. Until the nineteenth century, construction and examination of the edifice of mathematics took place almost entirely "above the surface," building upward from such surface features as numbers, points, and lines. Mathematicians for the most part did not concern themselves with, or even recognize or acknowledge, the problem (from the "edifice" point of view) that the building might stand on shifting sand instead of solid rock.

[100] [Str13, p. 4].
[101] [**?**, p. 92].
[102] [**?**].

In the latter part of the nineteenth century, continuing into the twentieth and to this day, mathematicians started exploring the part of the edifice lying below the surface. The development of set theory seemed to provide at least a basement, if not a foundation, for the structure. Logic itself seemed to lie either in a subbasement or an entire substructure of the building. But as investigations went further, it became less clear, rather than more, that there is a solid and well-formed foundation underlying it all.

The Tree of Mathematics

The analogy of mathematics as a tree instead of a building has also been popular for a long time in various forms. The tree is a living, growing organism instead of a structure built with blocks of stone. As it grows, more and more branches of the tree appear and develop. Sometimes parts of these branches become so intertwined that it is nearly impossible to trace where the different strands come from.

The tree of mathematics not only grows upwards above the surface of the ground, but a root system also develops and grows, giving the tree increased (but not total) stability. If the stability of the tree is threatened in a certain direction, there is an impetus to extend the root system in that direction to restabilize the tree.

One problem with this view of mathematics as a tree, of course, is that it is unclear what role human mathematicians have in the growth and development of the tree; the idea that the tree grows on its own, with only fertile soil, water, and sunlight needed, is rather contrary to the picture of mathematics as a human activity. On the other hand, mathematics could be regarded as having a life of its own:

> "Upon superficial observation mathematics appears to be a fruit of many thousands of scarcely related individuals scattered through the continents, centuries and milleniums. But the internal logic of its development looks much more like the work of a single intellect that is developing his thought continuously and systematically, using as a tool only the variety of human personalities. As in an orchestra performing a symphony by some composer, a theme is passing from one instrument to another, and when a performer has to finish his part, another one is continuing it as if playing from music."
>
> *I. R. Shafarevich*[103]

> "Math, I would argue, is the most impressive and longest-running collective enterprise in human history."
>
> *Alex Bellos*[104]

The view of mathematics as a tree makes some of the foundational questions less compelling: what is usually called "foundations of mathematics" is simply the root structure of the tree. Indeed, it can be argued from this point of view that the whole idea of a "foundation" of mathematics is not necessary or appropriate, so long as we have confidence that the tree is growing in some kind of solid ground.

[103] On some tendencies of the development of mathematics; cf. [Bog07], v. 2, p. 439.
[104] [**?**, p. xii].

The Continent of Mathematics

> "I seem to have been only like a boy playing on the seashore, and diverting myself now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me."
>
> *Isaac Newton*

> "The known is finite, the unknown infinite; intellectually we stand on an islet in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land."
>
> *T. H. Huxley*[105]

An analogy I particularly like, which I first saw in Ian Stewart's book, *Nature's Numbers* [**?**] (cf. also [Bel87, p. 8]), compares mathematics to the landscape of a mostly unexplored continent. To paraphrase and expand on Stewart, when man first arrived on this remote continent, he could see very little through the impenetrable jungle beyond the beach. The continent has many resources, but is covered by thick vegetation and crisscrossed by mountains and gorges which are difficult and dangerous to cross. In time, some areas of the continent are explored, and settlements and roads built, and gradually more and more of the continent becomes known and accessible. The towns and roads impose a certain social order on the explored region, and some places become established as important centers of the society (perhaps with universities having mathematics departments.)

Here and there are high peaks which are mostly obscured by fog. On clear days, however, expert climbers can reach the summit and get panoramic views of large areas of the continent. They may be able to see undiscovered nearby verdant valleys which could be inhabited, or passes or natural bridges across mountains or gorges previously thought impassable. They may also be able to see that some areas thought to be far apart and unconnected are only actually separated by a narrow ridge and can easily be connected by a short road. (We assume this society never developed air or space travel.)

> "Creating a new theory is not like destroying an old barn and erecting a skyscraper in its place. It is rather like climbing a mountain, gaining new and wider views, discovering unexpected connections between our starting point and its rich environment. But the point from which we started out still exists and can be seen, although it appears smaller and forms a tiny part of our broad view gained by the mastery of the obstacles on our adventurous way up."
>
> *Albert Einstein*

Mathematicians are the architects of the creation of a society on the continent of mathematics (which is actually a "virtual continent", not physically altered by the society and capable of supporting simultaneous independent settlement by other cultures). Some mathematicians explore new regions of the continent; others build roads and bridges for the public to use; others design and build the cities; others invent ways of using the resources. Some operate farms or mines in recently explored regions where the climate or topography allows new crops to be grown or new minerals found. [Farms are probably a better analogy than mines, since mathematical resources, once discovered, cannot be depleted; but mines are apt in the sense that digging

[105] "On the Reception of the Origin of Species," in Francis Darwin, ed., *The The Life and Letters of Charles Darwin*, vol. 1.

below the surface in mathematics can sometimes yield new and unexpected treasures.] Some logicians write the laws governing the society, and others study the scientific principles underlying the formation of the continent. The top mathematicians are the expert climbers who can go up the peaks to get the big picture and see undiscovered regions and connections, although these can also sometimes be discovered at ground level. Most of the mathematicians may not be able to reach the summits without cable cars and guided tours, but there is much worthwhile to be done at all levels of the society.

In this view, the topography of the continent of mathematics is pre-existing and is discovered, not created, by the mathematicians. However, the society created by the mathematicians is not preordained: choices are made about where to build the cities and roads, what kind of political and social order to establish, and which resources to use and how; these choices depend on the intellectual ability and cultural background of those building the society, as well as seemingly random events such as what parts of the continent happen to be discovered or explored at a given time, or whether a sufficiently skilled climber comes along during the right weather conditions to get to the top of a critically located peak to catch the view. It could well happen that a different group of settlers (such as aliens from another planet) would establish a quite different society on the same continent.

> "The true mathematician is always a good deal of an artist, an architect, yes, of a poet. Beyond the real world, though perceptibly connected with it, mathematicians have intellectually created an ideal world, which they attempt to develop into the most perfect of all worlds, and which is being explored in every direction. None has the faintest conception of this world, except he who knows it."
>
> *A. Pringsheim*[106]

But as E. T. Bell [Bel87, p. 8] observes, exploration of the continent of mathematics is an imperfect process:

> "If we marvel at the patience and the courage of the pioneers, we must also marvel at their persistent blindness in missing the easier ways through the wilderness and over the mountains. What human perversity made them turn east and perish in the desert, when by going west they could have marched straight through to ease and plenty? This is our question today. If there is any continuity in the progress of mathematics, the men [sic] of a hundred years hence will be asking the same question about us."

Of course, the continent view, or any view, of mathematics is not entirely adequate:

> "Mathematics is not a book confined within a cover and bound between brazen clasps, whose contents it needs only patience to ransack; it is not a mine, whose treasures may take long to reduce into possession, but which fill only a limited number of veins and lodes; it is not a soil, whose fertility can be exhausted by the yield of successive harvests; it is not a continent or an ocean, whose area can be mapped out and its contour defined; it is as limitless as that space which it finds too narrow for its aspirations; its possibilities are as infinite as the worlds which are forever crowding in and multiplying upon the astronomer's gaze; it is as incapable of being reduced to definitions of permanent validity, as the consciousness of life."
>
> *J. Sylvester*[107]

[106] *Jahresbericht der Deutschen Mathematiker Vereinigung*, Bd. 32, p. 381; cf. R. Moritz, *Memorabilia Mathematica*, 1108.
[107] cf. [**?**, p. 124].

I.6.10. Pure vs. Applied Mathematics

Everyone recognizes that mathematics has enormous and vital applications to the "real world." Eugene Wigner describes this as "the unreasonable effectiveness of mathematics." (In my view, this thinking is a little backwards: it is not self-evident to me that there should be physical laws at all, but if there are, and the apparent orderly behavior of the universe suggests there are, they must necessarily, like everything else, operate in accordance with mathematical principles.) But mathematicians and nonmathematicians tend to view the applicability of mathematics in rather different ways. Nonmathematicians often view mathematics as just a tool to be used, a discipline whose importance is primarily in support of scientific and technical subjects (and increasingly other academic areas too); while they recognize its importance, they also understand its limitations in this respect. The following statement by an engineer is fairly typical of this point of view:

> "When the difficulty of a problem lies only in finding out what follows from certain fixed premises, mathematical methods furnish invaluable wings for flying over intermediate obstructions; but whenever the chief difficulty of a problem lies in the multiplicity and dubiousness of the premises themselves, and in reconciling them with each other, there is no safe course but to remain continuously on the solid ground of concrete fact."
>
> *A. M. Wellington*[108]

Mathematicians, on the other hand, generally regard mathematics as an intellectual pursuit with a life of its own independent of applications; mathematics is no more simply a support subject for science than physics is just a support subject for engineering or chemistry a support subject for medicine. Mathematics as done by mathematicians bears the same relation to mathematics as used by scientists as literature bears to business communications.

> "[C]onsider mathematics as some kind of giant computer with a large number of terminals on its periphery, representing fields of application. A practising scientist is like the terminal user. He is primarily interested in the output and will know something about what the computer can do for him, but he is not involved in what goes on inside the heart of the computer. In the early days of computers, users and designers were frequently the same people, but with their rapid growth and sophistication this is now the exception rather than the rule. Similarly it is the increasing sophistication of mathematics which has led to the large gap between 'users' and 'designers'."
>
> *M. Atiyah*[109]

The interest of many mathematicians in applications is limited to being pleased that nonmathematicians find mathematics important enough that it is worth encouraging and supporting (e.g. via university mathematics departments and NSF grants), but being scornful of occasional attempts by nonmathematicians or administrators to direct resources to "useful" parts of mathematics. Indeed, most mathematicians would say that no one can predict what parts of mathematics will turn out to be most useful, and anyway utility outside mathematics should play only a minor role, if any at all, in the day-to-day work of a mathematician.

These attitudes are largely shared by both "pure" and "applied" mathematicians. There is really not much systematic difference in the way pure and applied mathematicians work or in their approach to mathematics,

[108] *The Economic Theory of Railway Location*, 1887, p. viii.
[109] [**?**].

although there is some difference in the areas of theoretical mathematics they work in most heavily. The main difference, which provides the best definition of "pure mathematics" and "applied mathematics", is simply in the *origin* of the problems they work on: in pure mathematics the problems come from within mathematics, while in applied mathematics the problems come from outside mathematics. (The applied mathematician also sometimes has the additional task of relating the solution of the mathematical problem to the original nonmathematical setting.) Thus almost all mathematicians are "pure mathematicians" at times and "applied mathematicians" at other times, and it is usually rather artificial to label a mathematician as one or the other.

Here are some more expressions of the mathematicians' views on applications:

> "Practical application is found by not looking for it, and one can say that the whole progress of civilization rests on that principle."
>
> *J. Hadamard*[110]

> "Only impractical dreamers spent two thousand years wondering about proving Euclid's parallel postulate, and if they hadn't done so, there would be no spaceships exploring the galaxy today."
>
> *M. Greenberg*[111]

> "Mathematics has the unique character of being a scientific discipline with applications to all other kinds of sciences and to practical life. But mathematics is also an art, the beauty lying in the symmetries, patterns, and intricately deep relationships which enchant the beholder. Discoveries that require the invention of new methods and great ingenuity are indeed to be hailed as important – at least from one point of view. Will these be of any practical use someday? Is it a legitimate question? Indeed, numerous are the examples when theories seemed for centuries to be gratuitous speculations, like the study of prime numbers, but today a mainstay of crucial applications in communications. It is the intrinsic quality of a new result which confers its importance."
>
> *Paulo Ribenboim*[112]

> "The mathematician does not study pure mathematics because it is useful; he studies it because he delights in it and he delights in it because it is beautiful."
>
> *Henri Poincaré*

> "Understanding is a lot like sex. It's got a practical purpose, but that's not why people do it normally."
>
> *Frank Oppenheimer*[113]

Scientists, too, express similar thoughts on an excessive preoccupation with applications:

[110] cf. [Gre93, p. xxix].
[111] [Gre93, p. xxix].
[112] *My Numbers, My Friends: Popular Lectures on Number Theory*, Springer, 2000, p. 327.
[113] K. C. Cole, *The Universe and the Teacup*, 1998.

"If in the range of human endeavor after sound knowledge there is one subject that needs to be practical it is Medicine. Yet in the field of Medicine it has been found that branches such as biology and pathology must be studied for themselves and be developed by themselves with the single aim of increasing knowledge; and it is then that they can be best applied to the conduct of living processes. So also in the pursuit of mathematics, the path of practical utility is too narrow and irregular, not always leading far. The witness of history shows that, in the field of natural philosophy, mathematics will furnish the more effective assistance if, in its systematic development, its course can freely pass beyond the ever-shifting domain of use and application."

A. R. Forsyth[114]

"He who seeks for immediate practical use in the pursuit of science, may be reasonably sure, that he will seek in vain. Complete knowledge and complete understanding of the action of forces of nature and of the mind, is the only thing that science can aim at. The individual investigator must find his reward in the joy of new discoveries, as new victories of thought over resisting matter, in the esthetic beauty which a well-ordered domain of knowledge affords, where all parts are intellectually related, where one thing evolves from another, and all show the marks of the mind's supremacy; he must find his reward in the consciousness of having contributed to the growing capital of knowledge on which depends the supremacy of man over the forces hostile to the spirit."

H. Helmholtz[115]

"In *Plato's* time mathematics was purely a play of the free intellect; the mathematic-mystical reveries of a Pythagoras foreshadowed a far-reaching significance, but such a significance (except in the case of music) was as yet entirely a matter of fancy; yet even in that time mathematics was the prerequisite to all other studies! But today, when mathematics furnishes the *only* language by means of which we may formulate the most comprehensive laws of nature, laws which the ancients scarcely dreamed of, when moreover mathematics is the *only* means by which these laws may be understood,– how few learn today anything of the real essence of our mathematics! ... In the schools of today mathematics serves only as a disciplinary study, a mental gymnastic; that it includes the highest ideal value for the comprehension of the universe, one dares scarcely to think of in view of our present day instruction."

F. Lindemann, 1904[116]

"What is the usefulness of a new-born child?"

Benjamin Franklin[117]

[114] Presidential Address, British Association for the Advancement of Science; *Nature* 56 (1897), p. 377; cf. [Mor93, 1540].
[115] *Vorträge und Reden* (Braunschweig, 1884), Bd. 1, p. 142; cf. [Mor93, 1545].
[116] *Lehren und Lernen in der Mathematik* (München, 1904); cf. [Mor93, 1523].
[117] When the usefulness of a new invention was questioned. J. Parton, *Life and Times of Benjamin Franklin* (1864), Part IV, Chapter 17.

"That it should have to justify itself by its applications, as a tool for making the mundane or improving the ephemeral, is an affront not just to mathematics but to the creature that invented it. What higher calling can there be than searching for useless and beautiful truths? Number theory is as beautiful and no more useless than mastery of the balance beam or the well-thrown forward pass. And our culture expends enormous sums on those exercises without asking what higher ends they serve."

Charles Krauthammer[118]

"Mathematics as cultivated by professionals contains a great deal that has not yet been applied to science. If we may extrapolate from past experience, much of this vast mass of abstractions will some day find a use in the exploration of nature. Roughly the progress of mathematics and its applications is like a game of leapfrog: at one time pure mathematics is far ahead of any evident uses; at another, scientific advances invite new advances in mathematics."

E. T. Bell[119]

I.6.11. Is Mathematics an Art or a Science?

"Mathematics is an art, and as an art chooses beauty and freedom. It is an aid to technology, but is not a part of technology. It is a handmaiden of the arts, but it is not for this reason an art. Mathematics is an art because its mode of discovery and its inner life are like those of the arts."

Marston Morse[120]

I.6.11.1. Nonmathematicians, and especially scientists, tend to think of mathematics as a science; although there are obvious differences in character between mathematics and experimental sciences, the normal demonstrative nature of visible mathematics has much in common with techniques in science. Because of this, and the many fruitful interactions between mathematics and the experimental sciences, mathematics is often organizationally grouped among the sciences in universities.

Many mathematicians are not so sure. Mathematics is a unique discipline with some flavor of science but also a creative aspect which to many more resembles art. Mathematics is not simply a mechanical or purely logical process; imagination plays an essential role, and what distinguishes great mathematicians from ordinary ones is much more a difference in imagination than technical ability. (To be sure, top-quality science also has a strong creative thought component.) And mathematics possesses an intrinsic beauty and elegance which can overwhelm those who delve deeply enough into the subject to experience it.

I.6.11.2. Indeed, mathematics cannot be considered a science at all unless one believes in some form of realism, since science is by definition the systematic study and understanding of some aspect of nature. Thus a logicist, formalist, or intuitionist/constructivist cannot regard mathematics as a science. But even a realist need not entirely subscribe to the notion that mathematics is a science; the Continent of Mathematics picture, for example, portrays mathematicians as far more than just scientists.

[118] *Time* magazine, April 18, 1988.
[119] [Bel87, p. 320].
[120] Kenyon Address [**?**].

"I like to look at mathematics almost more as an art than as a science; for the activity of a mathematician, constantly creating as he is, guided though not controlled by the external world of the senses, bears a resemblance, not fanciful I believe but real, to the activity of an artist, of a painter let us say. Rigorous deductive reasoning on the part of the mathematician may be likened here to technical skill in drawing on the part of the painter. Just as no one can become a good painter without a certain amount of skill, so no one can become a mathematician without the power to reason accurately up to a certain point. Yet these qualities, fundamental though they are, do not make a painter or mathematician worthy of the name, nor indeed are they the most important factors in the case. Other qualities of a far more subtle sort, chief among which in both cases is imagination, go to the making of a good artist or good mathematician.

Maxime Bôcher[121]

"Mathematics, rightly viewed, possesses not only truth, but supreme beauty – a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show."

Bertrand Russell[122]

"When I am working on a problem, I never think about beauty. I think only of how to solve the problem. But when I have finished, if the solution is not beautiful, I know it is wrong."

Buckminster Fuller[123]

"Surely the claim of mathematics to take a place among the liberal arts must now be admitted as fully made good. Whether we look at the advances made in modern geometry, in modern integral calculus, or in modern algebra, in each of these three a free handling of the material employed is now possible, and an almost unlimited scope is left to the regulated play of fancy."

J. J. Sylvester[124]

"The true mathematician is always a good deal of an artist, an architect, yes, of a poet. Beyond the real world, though perceptibly connected with it, mathematicians have intellectually created an ideal world, which they attempt to develop into the most perfect of all worlds, and which is being explored in every direction. None has the faintest conception of this world, except he who knows it."

A. Pringsheim[125]

[121] Fundamental conceptions and methods in mathematics, *Bulletin of the American Mathematical Society*, 9 (1904), p. 133; cf. R. Moritz, *Memorabilia Mathematica*, 1103.
[122] *The study of Mathematics: Philosophical Essays*; cf. R. Moritz, *Memorabilia Mathematica*, 1104.
[123] M. Livio, *The Golden Ratio*, p. 10.
[124] *Collected Mathematical Papers*, v. 3, p. 123; cf. R. Moritz, *Memorabilia Mathematica*, 1106.
[125] *Jahresbericht der Deutschen Mathematiker Vereinigung*, Bd. 32, p. 381; cf. R. Moritz, *Memorabilia Mathematica*, 1108.

"Mathematics has beauties of its own – a symmetry and proportion in its results, a lack of superfluidity, an exact adaptation of means to ends, which is exceedingly remarkable and to be found elsewhere only in the works of the greatest beauty. ... The beauties of mathematics – of simplicity, of symmetry, of completeness – can and should be exemplified even to young children. When this subject is properly and concretely presented, the mental emotion should be that of enjoyment of beauty, not that of repulsion from the ugly and the unpleasant."

J. W. A. Young[126]

"The beautiful has its place in mathematics as elsewhere. The prose of ordinary intercourse and of business correspondence might be held to be the most practical use to which language is put, but we should be poor indeed without the literature of imagination. Mathematics too has its triumphs of the creative imagination, its beautiful theorems, its proofs and processes whose perfection of form has made them classic. He must be a 'practical' man who can see no poetry in mathematics."

W. F. White[127]

"The mathematician is entirely free, within the limits of his imagination, to construct what worlds he pleases. What he is to imagine is a matter for his own caprice; he is not thereby discovering the fundamental principles of the universe nor becoming acquainted with the ideas of God. If he can find, in experience, sets of entities which obey the same logical scheme as his mathematical entities, then he has applied his mathematics to the external world; he has created a branch of science."

J. W. N. Sullivan[128]

"[I]t was not its worldly utility that led Euclid or Archimedes or Georg Cantor to devote so much of their energy and genius to mathematics. These individuals did not feel compelled to justify their work with utilitarian applications any more than Shakespeare had to apologize for writing love sonnets instead of cookbooks or Van Gogh had to apologize for painting canvases instead of billboards."

William Dunham[129]

"Mathematics is not the rigid and rigidity-producing schema that the layman thinks it is; rather, in it we find ourselves at that meeting point of constraint and freedom that is the very essence of human nature."

Hermann Weyl

[126] *The Teaching of Mathematics* (New York, 1907), p. 44; cf. R. Moritz, *Memorabilia Mathematica*, 1110.
[127] *A Scrap-Book of Elementary Mathematics* (Chicago, 1908), p. 208; cf. R. Moritz, *Memorabilia Mathematica*, 1119.
[128] *Aspects of Science*, 1925.
[129] [Dun90, p. vi].

"Mathematics is the most perfect union between exact knowledge and theoretical thought."

E. Curtius[130]

"Proof is the end product of a long interaction between creative imagination and critical reasoning. Without proof the program remains incomplete, but without the imaginative input it never gets started."

M. Atiyah[131]

"It is a relatively good approximation to truth ... that mathematical ideas originate in empirics, although the genealogy is sometimes long and obscure. But, once they are conceived, the subject begins to live a peculiar life of its own and is better compared to a creative one, governed by almost entirely aesthetical motivations."

J. von Neumann[132]

"Historically, mathematics developed out of man's concern with physical problems. And although mathematics is ultimately a construction of the mind alone and exists only as a magnificent collection of ideas, from the earliest times down to the present day it has been stimulated and inspired and enriched by contact with the external world. Its problems are often idealizations of problems first encountered by the physicist or the engineer, or, more recently, by the social scientist. Many of its concepts are abstractions from the common experience of all men. And many areas of mathematics stem originally from the needs of scholars for more powerful analytical tools with which to pursue their investigations of the world around them. Surely, without contact with the external world, mathematics, if it existed at all, would be vastly different from what it is today.

But mathematics repays generously her indebtedness to the other sciences. At the mere suggestion of a new problem, mathematics sets to work developing procedures for its solution, generalizing it, relating it to work already done and to results already known, until finally it gives back to scholars in the original field a well-developed theory for their use.

Oftentimes mathematics outruns completely the demands of the physical problem which may have stimulated it, and careless critics scoff at its 'pure' or 'abstract' or 'useless' character. Such criticism is absurd on two quite different counts. In the first place, all mathematics worthy of the name is pure or abstract or even, in a certain sense, useless, just as poetry and music and painting and sculpture are useless except as they bring satisfaction to those who create and to those who enjoy. It is no more appropriate to criticize mathematics for possessing the attributes of one of the creative arts than it is to criticize the arts themselves, unless it be that perhaps the number of those who find enjoyment in mathematics, though considerable, is less than those who enjoy the more conventional arts. Then in the second place, by a remarkable coincidence which is almost completely responsible for the existence of all present-day science and technology and which has no counterpart in the arts, even the most abstract parts of mathematics have turned out again and again to be highly 'practical,' as science, becoming ever more sophisticated, finds that it needs mathematical tools of greater and greater refinement."

[130] *Berliner Monatsberichte*, 1873, p. 517; cf. [Mor93, 1409].
[131] *The Princeton Companion to Mathematics*, p. 1002.
[132] J. R. Newman, Ed., *The World of Mathematics*, v. 4, p. 2063.

C. R. Wylie, Jr.[133]

"Practical, scientific, aesthetic, and philosophical interests have all shaped mathematics. It would be impossible to separate the contributions and influences of any one of these forces and weigh it against the others, much less assert sweeping claims to its relative importance. On the one hand, pure thought, the response to aesthetic and philosophical interests, has decisively fashioned the character of mathematics and made such unexcelled contributions as Greek geometry and modern non-Euclidean geometry. On the other hand, mathematicians reach their pinnacles of pure thought not by lifting themselves by their bootstraps but by the power of social forces. Were these forces not permitted to revitalize mathematicians, they would soon exhaust themselves; thereafter they could merely sustain their subject in an isolation which might be splendid for a short time but which would soon spell intellectual collapse."

Morris Kline[134]

"Mathematics has in modern times grown to such vast proportions that no one can now hope to have a detailed knowledge of, let us say, even a tenth of the whole field. Moreover, it is still growing, and more rapidly than ever. Hundreds of research papers are published each year developing new processes and announcing theorems previously unknown.

Much of this higher mathematics is very abstract. But it is not therefore valueless, even from the standpoint of applications. Several subjects which initially developed in a theoretical way with no thought of a practical application have been later taken over bodily by some practical science ... Several others have contributed powerful methods to the solution of particular problems. The theoretical mathematics of today may be practical mathematics tomorrow.

But this is only one aspect of the matter. The intellectual values obtainable from a contemplation of the power, elegance, and absolute precision of mathematical reasoning, and of the perfect harmony existing among the various branches, are very great indeed. Abstract mathematics is a work of invention – *a free creation of the human spirit*, as truly a work of art as the Moonlight Sonata or the Sistine Madonna, but on a much vaster scale than the entire library of great symphonies, the whole gallery of famous paintings, or, indeed, the total assemblage of celebrated cathedrals. ... [M]en and women who have time to get an understanding of the more advanced branches find nothing finer and more inspiring than the wonderfully abstruse investigations of pure mathematics – achievements of the reason which far transcend the realms of physical sense, and time and space."

Frank L. Griffin, 1921[135]

"Without mathematics we cannot penetrate deeply into philosophy.
Without philosophy we cannot penetrate deeply into mathematics.
Without both we cannot penetrate deeply into anything."

Leibniz[136]

[133] [Wyl64, p. 5].
[134] [**?**, p. 7].
[135] [Gri21].
[136] [Cha05, p. 3].

I.7. The Foundations of Mathematics

Despite more than a hundred years of intense investigation, the foundations of mathematics remain almost as shaky as ever.

Until the nineteenth century, construction and examination of the edifice of mathematics took place almost entirely "above the surface," building upward from such surface features as numbers, points, and lines. Mathematicians for the most part did not concern themselves with, or even recognize or acknowledge, the problem that the building might stand on shifting sand instead of solid rock.

In the latter part of the nineteenth century, continuing into the twentieth and to this day, mathematicians started exploring the part of the edifice lying below the surface. The development of set theory seemed to provide at least a basement, if not a foundation, for the structure. Logic itself seemed to lie either in a subbasement or an entire substructure of the building. But as investigations went further, it became less clear, rather than more, that there is a solid and well-formed foundation underlying it all.

I.7.1. Philosophies of Mathematics

> "There is no primal language, or world view, or real line, but there are many languages, world views, and real lines, each of which can be understood to a large extent in terms of another but can only be fully understood on its own terms."
>
> *T. W. Körner*[137]

There are, at least among philosophers, four principal philosophies of mathematics, or, more accurately, of the foundations of mathematics: realism or Platonism, intuitionism or constructivism, logicism, and formalism. Each has several variations. While the arcane details are more of interest to philosophers than working mathematicians, it is important for mathematicians to have some understanding of the differences.

Most philosophers admit that none of the four philosophies is adequate by itself as an overall philosophy of mathematics (see, for example, [Goo86b]; my extended review of this article can be found at ()). Some modern-day philosophers are also critical of the historical preoccupation of philosophy with the foundations of mathematics, at the expense of studying its actual practice – see, for example, [Tym86] and [Cor03]). Indeed, each of them in its crudest form can be rather easily demolished as an absurdity. Nonetheless, each has some appealing and apparently accurate aspects, and most mathematicians (insofar as they can articulate any coherent personal philosophy of mathematics at all) seem to believe some combination of the philosophies; philosophers too acknowledge that belief in one of the philosophies does not preclude partial acceptance of the others. In fact, I would say the primary shortcoming of each of the philosophies, at least in their more nuanced versions, is the assertion of exclusivity.

> "The divergent views [on the foundations of mathematics] are not divergent answers to the same questions; they are deeply opposed construals of those questions. What counts as giving 'foundations' for mathematics is entirely different for Poincaré than for Russell, Frege, or Zermelo. This difference is not readily adjudicable through the arguments found in the explicit controversies of the literature."
>
> *Warren Goldfarb*[138]

[137] [Kör04, p. 394].
[138] [Gol88].

We will only give a bare introduction to these philosophies, which will necessarily be quite superficial and will give short shrift to the complex history of the ideas. Readers are referred to books on the philosophy of mathematics for details. I will also express some personal opinions; I do this advisedly, despite the warning from [Pot04, p. v]:

> "... mathematicians have often been tempted, especially in later life, to commit to print philosophical reflections which are either wholly vacuous or hopelessly incoherent."

Realism and Platonism

Realism as a philosophy of mathematics has definitely been unfashionable among philosophers, to the extent that many philosophers talk about "the three schools of the philosophy of mathematics," meaning the other three! However, among mathematicians it is probably the predominant view, and one which I largely subscribe to myself. (Philosophers tend to be dismissive of the opinions of mathematicians on this issue as unreliable, unsophisticated, and/or uninformed, not entirely without justification. Mathematicians are often similarly dismissive of philosophers; cf. I.7.1.30..) Mathematicians generally have a strong feeling that the abstract objects they work with are "real," although this could of course be an illusion. This feeling is reflected in the language mathematicians use to describe advances in mathematics, almost invariably referring to "discovering" or "finding" new results or proofs. The term "invent" is rarely used, and when it is (e.g. "Newton and/or Leibniz invented calculus") it usually refers to the development of a new subdiscipline or point of view in mathematics, which can be viewed as a way of facilitating analysis of the structure of part of the "continent of mathematics" () rather than creation of new mathematics *per se*.

> "Mathematical truth has some sort of necessity about it, which contrasts with the merely contingent beliefs we have about the world of sense experience. There is a hardness about mathematical truth which makes it not only ineluctably true, but profoundly true, because it is immune to the changes and chances of this fleeting world of transient phenomena, and tells us about not what just is, but what must be, the case."
>
> *J. R. Lucas*[139]

> "[W]hile the other sciences search for the rules that God has chosen for this universe, we mathematicians search for the rules that even God has to obey."
>
> *Jean-Pierre Serre*[140]

> "Platonism seems obvious when you are thinking about mathematical truth, but impossible when you are thinking about mathematical knowledge."
>
> *W. D. Hart*[141]

[139] [Luc00, p. 20].
[140] *AMS Notices*, v. 60 no. 4 (April 2013), p. 414.
[141] cf. [Luc00, p. 13].

Existence

I.7.1.3. Some critics of realism or Platonism construe "existence" too narrowly, limiting it to actual observable physical objects. By this restricted use of the term most, if not all, mathematical "objects" do not exist. But then other things we commonly talk about also cannot exist. What about gravity, or more generally "force" as used in physics? Gravity is not an object and cannot itself be observed (at least so far – we can only observe *effects* we attribute to gravity). What about "good" and "evil"? Are all these things just constructs of the human mind, or do they really "exist"?

> "Of course it is happening inside your head, Harry, but why on earth should that mean that it is not real?"
>
> *Albus Dumbledore*[142]

I.7.1.4. Actually, what does "existence" mean for an abstract concept? This is a tough question, with many logical pitfalls. For example, while "the set of all sets" is a logical impossibility due to Russell's Paradox, if we say that the set of all sets does not exist, then how can we talk about it? What is the meaning of the sentence "The set of all sets does not exist."? How can a sentence refer to a nonexistent concept?

I.7.1.5. We even use nonexistent objects or concepts routinely in mathematics. For example, the ancient proof that $\sqrt{2}$ is irrational, as it is usually written today (I.6.5.8.), begins with: "Suppose a and b are natural numbers with $\frac{a}{b} = \sqrt{2}$." The proof goes on to show by contradiction that such a and b cannot exist. So how can we *suppose* they exist when they do not? Can we make a distinction between a *conceivable* concept and a concept that exists? If a concept does not exist, how can it be conceivable?

We often do not know whether mathematical objects exist. For example, no one knows whether "the smallest even natural number larger than two which cannot be written as a sum of two primes" exists or not. If it does not exist, how can I talk about it, as I just did?

I.7.1.6. The question of "existence" of numbers like 1 and 2 seems to me essentially the same as the question of "existence" of hydrogen and helium. Note that the existence of "hydrogen" is a subtly different question than the existence of hydrogen *atoms*. It is indisputable that hydrogen atoms exist in great abundance, as do helium atoms. But one needs a *concept* of "hydrogen" to say that two distinct hydrogen atoms are both hydrogen, and that a helium atom is not hydrogen. If "hydrogen" exists only in the human mind, how can hydrogen and helium atoms act differently chemically, and by what mechanism do all hydrogen atoms act identically? Although the concept "hydrogen" may be a human construct, there is an aggregate of properties or characteristics possessed by some atoms and not others constituting "hydrogenness" independently of any involvement of the human intellect. So it is with "oneness." See I.7.1.12. for more comments.

If "one" and "two" exist as transcendent concepts, then what about the rest of the natural numbers? Do they all exist, and if not, what is the first one which does not?

I.7.1.7. Some things, if they "exist" at all, are certainly pure constructs of the human mind, e.g. Mickey Mouse, unicorns. But the distinction is not always so clear-cut: did the Piltdown Man ever exist, and if so, in what sense?

[142] J. K. Rowling, *Harry Potter and the Deathly Hallows*, Scholastic, 2007.

I.7.1.8. We can, of course, open a whole different can of worms by asking whether the empty set exists. On the one hand, the existence of *anything* would seem not to make sense unless we could contrast it with the existence of *nothing*; and if there does not exist anything, then nothing exists. On the other hand, if "nothing" exists, what kind of a thing could it be? And is the empty set the same as "nothing", and is "zero" the same as either? See [Gar01] for an essay on this subject. Existence of "nothing" or the "void" has actually been the subject of a long and serious debate, cf. [Kap00] and [Sei00].

I.7.1.9. Atiyah [**?**] argues that even the natural numbers are simply constructs of the human mind, and would not necessarily come into the consciousness of an intelligent creature of a different form:

> "But let us imagine that intelligence had resided, not in mankind, but in some vast solitary and isolated jellyfish, deep in the depths of the Pacific. It would have no experience of individual objects, only with the surrounding water. Motion, temperature and pressure would provide its basic sensory data. In such a pure continuum the discrete would not arise and there would be nothing to count."

(Such a creature is certainly conceivable, and may actually exist, perhaps on Enceladus or Europa?) But Atiyah confuses *existence* with *familiarity*. Just because natural numbers might not be within its sense experience and may be discovered late or not at all, it does not mean that they are any less "real." Four-dimensional geometry is not within our sense experience and thus began to be accepted and studied only relatively recently, but from a mathematical standpoint it is just as "real" as three-dimensional geometry. And eventually such a creature would find things to count, the dimensions in which it could move if nothing else; if it is intelligent enough to develop some form of calculus, it would also encounter first derivatives, second derivatives, etc.

Atiyah goes on to raise a more troubling point:

> "Even more fundamentally, in a purely static universe without the notion of time, causality would disappear and with it that of logical implication and of mathematical proof."

We can hardly conceive of the absence of time and causality, and mathematics and logic as they exist for us certainly have causality at their root. But mathematics, broadly enough defined, might well have aspects not even dependent on causality.

Atiyah makes some good points and convincingly argues that mathematics as we know and understand it is a creation of the human mind with origins in human sense experience, but his arguments that mathematical objects and truths do not have some type of independent existence do not hold water.

I.7.1.10. A common but misguided way of dismissing realism which I find particularly obnoxious is to equate it to theology or belief in God. Apparently the rationale is that everything must be created by *someone*, and if mathematics is discovered and not created in the human mind, it must have been created by God. This is simply baloney. One direction of the argument has some validity: realism would evidently be attractive to a religious mathematician, and it is a historical fact that some mathematicians and philosophers (including Plato himself) have been drawn to realism by their religious beliefs. But belief in mathematical realism does not in any way entail belief in a Creator, any more than belief in the existence of physical laws does; a mathematical realist is no more compelled to be religious than an astronomer or biologist.

"[Platonic] 'existence' is really just the objectivity of mathematical truth."

Roger Penrose[143]

"Mathematics is independent of the existence of material objects. In mathematics the word exist can only have one meaning: it signifies exemption from contradiction."

Henri Poincaré[144]

I.7.1.11. An interesting and rather different existence question is discussed in [Gow02]: Does the complex number i exist? This is not the same as asking whether the complex numbers exist: even if we accept that they do, there are *two* complex numbers whose square is -1, and unlike the two square roots of 2 in $\mathbb{R}$ where there is a reasonable way to pick one (the positive one) to call $\sqrt{2}$, there is no logical way to pick one of the square roots of -1 to be called i. But this is really more a question of definition rather than existence (if $\mathbb{C}$ is defined as in III.6.1., specification of i is included in the definition): it is really the fact that there is no naturally-defined (e.g. continuous) square root function defined on $\mathbb{C}$. It also relates to the logical problem of making the common definition of $\mathbb{C}$ as the complex plane: in addition to the logical difficulties discussed in III.6.2.24., there are two equally reasonable ways to identify $\mathbb{C}$ with $\mathbb{R}^2$.

Invention vs. Discovery vs. ???

I.7.1.12. Returning to the subject of chemical elements, some recent developments in chemistry/atomic physics raise interesting questions related to the existence question of mathematical objects, and whether mathematicians "discover" or "invent" new mathematics. It was reported that scientists had, with high probability, created a few atoms of the element with atomic number 118 (now called *oganesson*). Although these atoms lasted for only a minute fraction of a second, the existence of this element was transformed from a theoretical possibility to a physical fact.

I.7.1.13. It does not seem accurate to say that these scientists had either "discovered" or "invented" Element 118. Some news accounts referred to the achievement as a "discovery." This assertion seems rather similar to the common statement in history books that Peary "discovered" the North Pole. While Peary, or a member of his team, was (possibly!) the first person to *reach* the North Pole, he did not *discover* it; it was discovered by Winnie-the-Pooh. (If Peary discovered it, why was he in the Arctic in the first place?) In fact, the North Pole was actually discovered by the first person who realized that the earth is a round ball and rotates on an axis (the North Star has been known and used for navigation since ancient and possibly even prehistoric times, although it was a different star in ancient times due to precession; once one realizes that the North Star rises higher in the sky as one moves north, it is not much of a leap to expect that it will eventually be straight overhead if one goes far enough north, assuming the earth has no edge, although uniqueness of this point on the earth is not so obvious). After all, did Neil Armstrong discover the moon?

I.7.1.14. It also does not seem accurate to say that the scientists "invented" Element 118. While they were evidently the first humans to actually assemble such an atom, the potential existence of Element 118 has long been recognized in increasingly great detail, and even the likely methods of creation of a 118 atom were widely known to specialists. This is more of a semantic question than the "discovery" one: for example,

[143] [Pen05, p. 13].
[144] *Science and Method*, p. 151-152.

is it accurate to say that the WRIGHT brothers "invented" the airplane? Of course, they built the first one that actually flew, and one cannot minimize the importance of that achievement (and they were the first to solve the problem of balance and control, perhaps their most important contribution); but the concept of the airplane goes back to antiquity, and specific designs of airplanes which could have flown with sufficient technical refinements go back at least to LEONARDO DA VINCI (in fact, if the internal combustion engine had been available to DA VINCI, he might very well have built the first flying airplane; it is no coincidence that the first airplane flight came less than 25 years after the first practical internal combustion engine.)

I.7.1.15. Perhaps the best verb for this situation is *produce*: the scientists were the first to produce an element 118 atom. This might be an appropriate term to also describe some advances in mathematics, avoiding the "invent" vs. "discover" question.

I.7.1.16. Scientists have long speculated about even heavier elements. For example, many scientists believe that if the right isotope of an atom with atomic number 126 could be assembled, it would be fairly stable, and they can predict with a fair degree of confidence what its chemical and physical properties would be. So let us ask: does Element 126 exist?

I.7.1.17. Of the three possible answers, Yes, No, and Maybe, the only one which cannot be rationally justified is No. No one can know that there is not a single atom of this element anywhere in the universe. A more interesting question is whether existence of such an atom somewhere is even relevant to the discussion, since it seems there is no possible way for us to ever know whether Element 126 atoms currently exist or existed at one time, say, in the remnants of a star in a galaxy millions of light-years away (according to the theory of relativity, it does not even make sense to ask whether such an atom currently exists.)

I.7.1.18. To answer the existence question, one must really decide what "existence" means in this case. Does it mean having an actual atom, or does it just mean having a very specific and rather well-understood theoretically possible physical object? Without actual atoms at hand, Element 126 is only a theoretical concept, but it seems quite different in character than, say, the postulated extra dimensions of string theory. Even a high-school student who knows about the periodic table can easily speculate about the existence of Element 126 and understand exactly (at least in general terms) what kind of physical object it would be.

I.7.1.19. It would be easy to say that Element 126 thus clearly exists as an abstract concept. After all, otherwise how could scientists even talk about it, much less study its properties? But a note of caution must be given here. Mathematicians have often talked about "the set of all sets," and sometimes still do, even though it is now well known that such a set is a logical impossibility. Someday someone may discover a scientific principle which makes assembly of an Element 126 atom impossible; how would such a discovery affect the existence of Element 126 as an abstract concept?

I.7.1.20. Furthermore, if Element 126 is an object whose existence can be reasonably contemplated, what about Element 1000 (an atom with atomic number 1000)? I don't imagine most chemists or physicists think there is any theoretical possibility that atoms this large could ever be created. What is the largest atomic number which is conceivable as a potential physical object? If 1000 is not big enough, what about Element 10^{100}, which is clearly physically impossible since there are (according to our present knowledge) far less than 10^{100} protons and electrons in the universe.

Logicism

Logicism has the most succinct nutshell description of any of the philosophies of the foundations of mathematics: in Bertrand Russell's words, it is the belief that "Mathematics is logic." Or, as he later said in a more detailed and perhaps more accurate statement:

> "Mathematics and logic, historically speaking, have been entirely distinct studies. ... But both have developed in modern times: logic has become more mathematical and mathematics has become more logical. The consequence is that it has now become wholly impossible to draw a line between the two; in fact the two are one. They differ as boy and man: logic is the youth of mathematics and mathematics is the manhood of logic."
>
> *Bertrand Russell*[145]

Specifically, logicism holds that all mathematical objects and truths, beginning with the natural numbers, are just simple consequences of the laws of logic. All more advanced parts of mathematics are built up from basic logical principles by repeated applications of the same rules of logic.

Russell [**?**, p. 3] gives the following definition of (pure) mathematics:

> "Pure mathematics is the class of all propositions of the form 'p implies q,' where p and q are propositions containing one or more variables, the same in the two propositions, and neither p nor q contains any constants except logical constants. And logical constants are all notions definable in terms of the following: Implication, the relation of a term to a class of which it is a member, the notion of *such that*, the notion of relation, and such further relations as may be involved in the general notion of propositions of the above form. In addition to these, mathematics *uses* a notion which is not a constituent of the propositions which it considers, namely the notion of truth."

Quite apart from the vagueness of parts of this definition, it would surely not be considered an adequate description of mathematics by most modern mathematicians. (It should be noted that this definition was written more than a hundred years ago, and the conception of mathematics and its breadth and depth have changed markedly over this time. Indeed, Russell himself expressed second thoughts about this definition in the preface to the second edition of [**?**] in 1938.)

It is hard to see, for example, how the Axiom of Choice, or even the Axiom of Infinity or Peano's (second-order) axiom P5 (), could be considered as derivable from basic logic. Indeed, Russell admits [**?**, p. viii]:

> "Such existence-theorems, with certain exceptions, are, I should now say, examples of propositions which can be *enunciated* in logical terms, but can only be proved or disproved by empirical evidence."

This assertion seems to seriously undermine the argument that mathematics is simply a part of logic.

[145] *Introduction to Mathematical Philosophy*, 1919.

Intuitionism and Constructivism

The terms *intuitionism* and *constructivism* are often used more or less interchangeably to mean a combination of two rather distinct notions. They are usually thought of together since the primary proponent, L. Brouwer, forcefully argued for the combination. However, although there are obvious reasons why constructivism would be attractive to an intuitionist, there does not appear to be any necessary implication between the two, in either direction.

The thesis of intuitionism is that mathematics consists of the study of "mental mathematical constructions." To an intuitionist, mathematics, or a "mathematical construction", need not be based on formal logic or even be formalizable in the accepted sense of the term. Acceptable modes of mathematical reasoning are not completely specified, but simply consist of all reasoning the human mind considers convincing. As with all other aspects of human thought, the conception and understanding of a piece of mathematics may vary slightly from person to person, and communication of mathematics (or writing it down) is at best an imperfect representation of the mental construction involved. Thus it cannot be said that any piece of mathematics is entirely rigorous. The question of whether mathematical objects have a transcendent or mind-independent existence is not relevant in intuitionism, and perhaps does not even make sense.

I have considerable (but not complete) sympathy with the philosophy of intuitionism when it is expressed in this way, so long as the term "mental mathematical construction" is given a sufficiently liberal interpretation. For example, I see no reason to think that humans cannot make a "mental construction" of an infinite set, or even the set of real numbers. Most intuitionists have a rather restricted view of what a "mental mathematical construction" can be, and thus reject existence of infinite sets as "completed" objects, although they accept "potentially infinite" mathematical objects such as the natural numbers or even the real numbers. Most intuitionists reject the Law of the Excluded Middle, and some even reject the whole idea of negation (cf. [Gri46]–[Gri51], [Hey66, 8.2]), while others embrace the notion of contradiction in a limited form. As with all philosophies of mathematics, there is considerable variation in the specifics of the theories of various proponents of intuitionism.

Mathematicians often make use of arguments which do not have any obvious formalization, and sometimes accept such arguments as proofs. A good example is given in [Lak76,]. Formalizable but nonconstructive arguments are even more common in modern mathematics, although they were unusual as recently as the nineteenth century: when Hilbert first made a name for himself by giving a fairly simple but totally nonconstructive proof of a theorem on existence of finite systems of invariants, generalizing a result which had previously been proved laboriously by a constructive argument by P. Gordan, Gordan's reaction was "Das ist nicht Mathematik, das ist Theologie!" Gordan later warmed up to Hilbert's methods, and admitted, "I have convinced myself that theology also has its advantages." [Kli72, p. 930].

Even a good formalization of a theory is generally an imperfect representation of the intuitive theory. For example, ZF (and even ZFC) set theory is inadequate to settle the Continuum Hypothesis. In fact, Gödel's Incompleteness Theorem guarantees that essentially any formalization of a mathematical theory is inadequate in some respect.

Constructivism is a rather different, although not unrelated, idea: that mathematical objects only exist when they have been given an explicit finite construction. In the most common version, the natural numbers (individually, not as a completed totality) are regarded as "given" and all other mathematical objects must be explicitly constructed from them. This is the form of constructivism espoused by Brouwer, but its origins go back at least to Kronecker, to whom is attributed the famous statement

"God made the integers [natural numbers]; all else is the work of man."[146]

Intuitionists find this form of constructivism natural and/or attractive since they regard the natural numbers (or, more precisely, specific small natural numbers) as being mathematical objects whose mental construction is self-evident and needs no further explanation or justification. Thus constructivism can be regarded as a natural logical consequence, or a somewhat extreme special case, of intuitionism. However, one could embrace constructivism without accepting the overall philosophy of intuitionism, e.g. it would be easy to instead justify constructivism on the basis of some type of formalism. Indeed, most constructivists greatly restrict acceptable modes of reasoning, in a manner arguably more in line with formalism than intuitionism.

The specification of natural numbers, and not, say, real numbers, points and lines, or sets, as *the* fundamental objects of mathematics out of which everything else must be built, can seem a bit strange and arbitrary, but constructivists have extensive arguments purporting to justify it. See ().

Among mathematicians, the most forceful modern-day proponent of constructivism was Errett Bishop, who showed that a quite remarkable amount of classical analysis can be developed within constructivist mathematics; see [BB85]. (Much of this was developed before Bishop, notably by Brouwer; see, for example, [Hey66]. This development has continued since Bishop; see, for example, [BV06].) Bishop's book also contains a lengthy exposition of his philosophy of mathematics, which, it must be noted, differed significantly from Brouwer's. For a less dogmatic introduction to intuitionism and constructivism, see [Hey66] or [Dum00]. Heyting actually presents intuitionist logic in an intriguing way, as a generalization of ordinary logic where statements do not just have a truth value of 0 or 1, but a truth value in a more general Boolean algebra. (This is similar to, but not the same as, *fuzzy logic*, which despite the name is a mathematically sound theory where truth values can be numbers between 0 and 1; fuzzy logic has yet to become a part of mainstream mathematics.)

The development of the theory of imaginary and complex numbers is an interesting case study for the philosophies of intuitionism and constructivism, although the discussion is purely academic since historically the entire development of the complex numbers predated the formulation of intuitionism or any of the other modern philosophies of mathematics. Imaginary numbers were contemplated at least as far back as the 16th century, when Cardano, and, subsequently and more clearly, Bombelli, discovered that even to represent the real roots of a cubic polynomial with real coefficients by an algebraic formula, square roots of negative numbers must sometimes be used. They, and most other mathematicians for centuries afterward, regarded such square roots to be "nonexistent" as numbers and "meaningless" (many, including Cardano, even felt the same way about negative real numbers), but found that they could be used algebraically as though they made sense in order to obtain real results, without apparent logical contradiction. (Not all mathematicians of the era had this problem; in the seventeenth century, John Wallis accepted the legitimacy of complex numbers and even came close to describing the complex plane; cf. [Kli72, p. 594-595].) It was only much later that anything resembling an actual construction was done of imaginary and complex numbers, and that they began to be accepted as actual mathematical objects. Modern mathematicians consider the terms "real" and "imaginary" as historical artifacts which are still used merely for convenience; today, we would say there is nothing inherently more (or less) "real" about $\sqrt{2}$ than about $i = \sqrt{-1}$, although $\sqrt{2}$ may be more *familiar* to many people (but see I.7.1.11.).

[146]This was said at a meeting of the Berliner Naturforscher-Versammlung in 1886; cf. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, Bd. 2, p. 19. Bell [Bel87, p. 33] says, "As he said this in an after-dinner speech, perhaps he should not be held to it too strictly."

> "That this subject [of imaginary magnitudes] has hitherto been considered from the wrong point of view and surrounded by a mysterious obscurity, is to be attributed largely to an ill-adapted notation. If, for instance, $+1$, -1, $\sqrt{-1}$ had been called direct, inverse, and lateral units, instead of positive, negative, and imaginary (or even impossible) such an obscurity would have been out of the question."
>
> *C. F. Gauss*[147]

An intuitionist might well say (although I doubt most would) that CARDANO and BOMBELLI had made a "mental construction" of square roots of negative numbers, despite their reservations about whether these objects represented anything "real", and that imaginary numbers were part of mathematics from that time onward. (One must be careful here: as brilliant a mathematician as FREGE made a "mental construction" of the set of all sets, and worked for ten years on a theory based on this notion before becoming aware of Russell's Paradox.) A constructivist, on the other hand, probably would say that complex numbers did not become mathematical objects until the nineteenth century, and some would say that CARDANO and BOMBELLI were not justified in using them at all; in fact, if strict constructivist methods had been insisted on, several hundred years of mathematical progress probably would not have happened.

Constructivists disagree about whether Goldbach's Conjecture (that every even number greater than 2 is a sum of two primes) has a truth value at present. Most mathematicians would argue that this statement is definitely either true or false, and not logically indeterminate, since if it is false there is a finitely constructible even number for which it fails. The fact that it is not now known which it is, and may never be known, is not really relevant for most mathematicians. (On the other hand, the same argument can be made about the consistency of ZF set theory.)

For perhaps a clearer example, suppose x is a fixed number of 100,000 digits, say with no prime factors of less than 10,000 digits (beyond our current ability to test primality in practice). It is hard to imagine even an ardent constructivist saying that the statement "x is prime" has no definite truth value, since there is an obvious process of fixed finite length which will definitively establish the truth or falsity of the statement, even though the process is not one which we are currently able to carry out. What about the statement

> "359334085968622831041960188598043661065388726959079837 is prime."

(which is now known to be true)? Was this statement already true 100 years ago, or did it only become true when modern computers made it possible to determine the number's primality? What will happen in the future if some catastrophe leaves mankind without computers – will this statement still be true?

My view is that constructivist mathematics is good and interesting mathematics, and the techniques and results are of both theoretical and practical importance. A constructivist proof of a theorem is usually "better" (albeit sometimes less "elegant"), and says more, than a nonconstructive one. But constructivist mathematics is by no means all of mathematics!

This is the fundamental point of disagreement I have with (some) constructivists, who believe that constructivist mathematics is the only legitimate mathematics. For example, one finds the following grandiose claim made as an apparently serious assertion:

[147] [Mor93, 1730].

> "A presentation of today's intuitionist mathematics sets out from meaningful mathematics and, following a clear, bright road of mathematical reasons, arrives at a firm vantage from which conventional mathematics is seen to be generally false and set theory, analysis, arithmetic, algebra – all the fields of modern mathematics – reappear, scoured clean of fallacy and error."
>
> *D. C. McCarty*[148]

(I should not accuse all constructivists of this exclusivity attitude – for example, the philosophy of constructivism described in [BV06] is a healthy one with which I almost entirely agree.)

This quote also illustrates a point made in [Luc00]: intuitionists claim that legitimate mathematical arguments are simply ones that mathematicians find convincing; but most mathematicians find classical mathematics convincing, and (many) intuitionists think they are wrong to do so.

> "Intuitionists can no more argue that classical mathematicians are wrong to argue as they do than a jazz pianist can argue that Bach is wrong to compose as he did."
>
> *J. R. Lucas*[149]

Until I see that the Law of the Excluded Middle and other nonconstructive arguments lead to some contradiction or absurdity, I see no reason not to use them, although I think it is very worthwhile to see how much of mathematics can be developed and proved without them. In this respect my opinion about the Law of the Excluded Middle (called the "excluded muddle" in [Bel87]) is not very different from my attitude about the Axiom of Choice.

It is sometimes asserted that intuitionists/constructivists have created a "new" or "different" mathematics. I disagree with this assertion; my conception of what mathematics is is broad enough to comfortably include intuitionist mathematics as well as "classical" mathematics, and also potentially more mathematics based on other logical and language schemes.

Frege, although he is considered a logicist, also seemed to leave open the possibility of multiple logics in a statement which also admirably summarized the axiomatic method:

> "It cannot be demanded that everything be proved ... but we can require that all propositions used without proof be expressly declared as such, so that we can see distinctly what the whole structure rests upon. After that we must try to diminish the number of these primitive laws as far as possible by proving everything that can be proved. Furthermore, I demand – and in this I go beyond Euclid – that all methods of inference employed be specified in advance ... "
>
> *Gottlob Frege*[150]

Formalism

The crudest expression of formalism is the assertion that mathematics is simply a "game" of symbol manipulation according to agreed rules. As such, there is no "meaning" or "truth" in mathematics, only formal derivation of new statements from agreed starting assumptions or previously derived consequences.

[148] [Sha05, p. 357]
[149] [Luc00, p. 171]
[150] *Grundgesetze der Arithmetik*, v. 1, introduction, translated by M. Furth; cf. [Gra00, p. 231].

"Mathematics, like dialectics, is an organ of the higher sense, in its execution it is an art like eloquence. To both nothing but the form is of value; neither cares anything for content. Whether mathematics considers pennies or guineas, whether rhetoric defends truth or error, is perfectly immaterial to either."

Goethe[151]

"The results of systematic symbolical reasoning must *always* express general truths, by their nature; and do not, for their justification, require each of the steps of the process to represent some definite operation upon quantity. The *absolute universality of the interpretation of symbols* is the fundamental principle of their use."

William Whewell[152]

"If the question of 'truth' of axioms isn't the mathematician's business to discuss whose is it?"

Hilary Putnam[153]

There are considerably more complex and nuanced versions of formalism, however, which cannot be discredited nearly as easily as the crude version of the previous paragraph. HILBERT is regarded as the father of formalism, but HILBERT'S formalism is almost unrecognizably different from the description above. HILBERT maintained that there are indeed real mathematical objects such as numbers and geometric objects like points and lines, with some sort of Platonic or intuitive existence, but that the right way to derive the properties of these real objects is to abstract them into theories applying to not only the real objects but also to additional "ideal" objects which really have no meaning or existence on their own, but about which formal statements can be derived, in particular giving valid statements about the real objects one began with. When phrased in this way, formalism more or less just appears to be a codification of the abstraction process commonly accepted and used throughout mathematics.

The set of real numbers is a good and relevant example of expanding a set of "meaningful" objects into an ideal theory which can be formalized and axiomatized. Only certain real numbers can be said to have intuitive meaning, arguably only finitely many; in fact, in principle, at most countably many real numbers can even have an explicit (finite) description or characterization. But uncountably many "ideal" numbers must be added to make a theory that can be axiomatized in the standard way we describe in this book. HILBERT actually regarded any infinite set, considered as a "completed" object, to be an idealization, but did not shy away from considering and working with them formally.

HILBERT went on to state that the only arguments which should be accepted as proofs are "finitary" arguments which can be completely and explicitly carried out according to the most basic logical principles. In this respect, the limitations he argued for might not be terribly different than what constructivists insist on, although HILBERT was a severe critic of BROUWER and the intuitionists/constructivists; indeed, HILBERT regarded any kind of potential mathematical objects not leading to a logical contradiction as being things whose existence could be legitimately assumed, or could at least be worked with as though they really existed. HILBERT did not reject the role of intuition and non-finitary arguments in developing and understanding mathematics, but regarded such arguments as no substitute for finitary proofs.

[151] *Sprüche in Prosa, Natur IV, 946*; cf. R. Moritz, *Memorabilia Mathematica*, 1307.
[152] *The Philosophy of the Inductive Sciences* (London, 1858), Part I, Bk. 2, chap. 12, sect. 2; cf. R. Moritz, *Memorabilia Mathematica*, 1212.
[153] *Philos. Sci.* 27(1960), p. 207.

Hilbert's goal was the formalization of all of mathematics, with finite mechanical processes to determine the truth or falsity of every mathematical statement. Although his ultimate hopes were dashed by Gödel's Incompleteness Theorem, Hilbert's program led to the subject of *metamathematics*, which could be regarded as the mathematical theory of mathematical theories (including mathematical logic), and which is an important and flourishing branch of mathematics today.

> "[Gödel's Incompleteness Theorem] means that man can never eliminate the necessity of using his own intelligence, no matter how cleverly he tries."
>
> *Paul Rosenbloom*[154]

For a working mathematician, there is little practical difference between logicism and formalism; both assert that the only valid mathematical arguments are completely formal ones, and that intuition and plausibility arguments have no place in mathematics proper (at least officially). There are, however, large philosophical differences between the two, which have led to vigorous criticism of each by proponents of the other. For example, logicists regard mathematics as part of logic, while formalists consider logic, at least "mathematical logic," to be just a part of mathematics. Formalists, at least at the basic level described above, deny that there is any notion of mathematical "truth"; logicists maintain that there is, but that mathematical "truth" is really just logical "truth." Formalists (basic ones, not Hilbert) say that mathematics is not really "about" anything, a point of view vehemently rejected by logicists, who view logic, and hence mathematics, as being "about" everything.

The Practice of Mathematics

> "There is probably no other science which presents such different appearances to one who cultivates it and one who does not, as mathematics. To [the noncultivator] it is ancient, venerable, and complete; a body of dry, irrefutable, unambiguous reasoning. To the mathematician, on the other hand, his science is yet in the purple bloom of vigorous youth, everywhere stretching out after the "attainable but unattained," and full of the excitement of nascent thoughts; its logic is beset with ambiguities, and its analytic processes, like Bunyan's road, have a quagmire on one side and a deep ditch on the other, and branch off into innumerable by-paths that end in a wilderness."
>
> *C. H. Chapman, 1892*[155]

> "Math is about taking ingredients, putting them together, seeing what you can make out of them, and then deciding whether it's tasty or not."
>
> *Eugenia Cheng*[156]

Mathematics, as practiced by human mathematicians, consists of two rather distinct and almost independent aspects:

1. Ascertaining what is true and significant in mathematics, including good formulation of mathematical ideas and statements.

[154] Raymond Smullyan, *The Lady or the Tiger?*, 1982, p. 226.
[155] *Bulletin of the New York Mathematical Society*, 2, 1892, p. 61; cf. [Bel92, p. v].
[156] *How to Bake* π; cf. *New York Times*, May 3, 2016, p. D4.

2. Proving that the statements thought to be true are in fact true.

Both aspects are essential. They are not entirely independent; a third aspect which bridges the other two, of comparable importance, is:

1.5. Understanding why the statements which are true are true.

Some mathematicians would say that 1.5 and 2 are really the same thing, but they are not (at least not exactly). One can sometimes prove a statement without having much real understanding of why it is true, except in a formal sense. And one can often gain a pretty good "understanding" of how a mathematical subject works and of what must be true without being able to actually prove it. In fact, much of calculus had this status for about 200 years.

> "Primarily, mathematics is a method of inquiry known as postulational thinking. The method consists in carefully formulating definitions of the concepts to be discussed and in explicitly stating the assumptions that shall be the basis for reasoning. From these definitions and assumptions conclusions are deduced by the application of the most rigorous logic man is capable of using. ... [But to] describe mathematics as only a method of inquiry is to describe da Vinci's 'Last Supper' as an organization of paint on canvas. Mathematics is, also, a field for creative endeavor. In divining what can be proved, as well as in constructing methods of proof, mathematicians employ a high order of intuition and imagination."
>
> *Morris Kline*[157]

> "The logician cuts up, so to speak, each demonstration into a very great number of elementary operations; when we have examined these operations one after the other and ascertained that each is correct, are we to think we have grasped the real meaning of the demonstration? Shall we have understood it even when, by an effort of memory, we have become able to reproduce all these elementary operations in just the order in which the inventor had arranged them? Evidently not; we shall not yet possess the entire reality; that I know not what, which makes the unity of the demonstration, will completely elude us ...
>
> If you are present at a game of chess, it will not suffice, for the understanding of the game, to know the rules of moving the chess pieces. That will only enable you to recognize that each move has been made conformably to these rules, and this knowledge will truly have very little value. Yet this is what the reader of a book on mathematics would do if he were a logician only. To understand the game is wholly another matter; it is to know why the player moves this piece rather than that other which he could have moved without breaking the rules of the game. It is to perceive the inward reason which makes of this series of moves a sort of organized whole. This faculty is still more necessary for the player himself, that is, for the inventor."
>
> *Henri Poincaré*[158]

[157] [**?**, p. 4].
[158] "The Value of Science," in *The Foundations of Science*, ed. and trans. G. Halsted, The Science Press, 1946; cf. [**?**, p. 290-291].

> "We are not very pleased when we are forced to accept a mathematical truth by virtue of a complicated chain of formal conclusions and computations, which we traverse blindly, link by link, feeling our way by touch. We want first an overview of the aim and of the road; we want to understand the idea of the proof, the deeper context."
>
> *Hermann Weyl*[159]

It can be argued that 1, and even 1.5, are not really part of mathematics at all. This is exactly what (basic) formalism does: it gives a good, precise, and rather accurate description of 2, but has nothing to say about 1 or 1.5, and in effect excludes them from mathematics proper, and also implies that the history of mathematics has no place in the study of present-day mathematics; indeed, full acceptance of formalism (or logicism) would imply that humans did virtually no actual mathematics before the late nineteenth century (Euclid perhaps excluded). In fact, in 1901 Russell [**?**] only half-jokingly wrote:

> "Pure mathematics was discovered by Boole, in a work which he called *The Laws of Thought* (1854). This work abounds in asseverations that it is not mathematical, the fact being that Boole was too modest to suppose his book the first ever written on mathematics."

However, 1 and 1.5 are certainly parts, and important parts, of human mathematics, and a philosophy of mathematics which does not recognize and embrace this is an inadequate and empty one. Russell [**?**, p. 280] acknowledges this point:

> "It is a peculiar fact about the genesis and growth of new disciplines that too much rigour too early imposed stifles the imagination and stultifies invention. A certain freedom from the strictures of sustained formality tends to promote the development of a subject in its early stages, even if this means the risk of a certain amount of error."

André Weil [Wei62] elaborated on and at the same time qualified this view:

> "The mathematician who first explores a promising new field is privileged to take a good deal for granted that a critical investigator would feel bound to justify step by step; at times when vast territories are being opened up, nothing could be more harmful to the progress of mathematics than a literal observance of strict standards of rigor. Nor should one forget ... that the so-called 'intuition' of earlier mathematicians, reckless as their use of it may sometimes appear to us, often rested on a most painstaking study of numerous special examples, from which they gained an insight not always found among modern exponents of the axiomatic creed. At the same time, it should always be remembered that it is the duty, as it is the business, of the mathematician to prove theorems, and that this duty can never be disregarded for long without fatal effects. The experience of many centuries has shown this to be a matter on which, whatever our tastes or tendencies, whether 'creative' or 'critical,' we mathematicians dare not disagree."

[159] *Unterrichtsblätter für Mathematik und Naturwissenschaften*, 38, 177-188 (1932). Translation by Abe Shenitzer appeared in The American Mathematical Monthly, v. 102, no. 7 (August-September 1995), p. 646.

Felix Klein [Kle04, p. 207-208] similarly expounded on the nature of mathematics and formalism:

> "You can hear often from non mathematicians, especially from philosophers, that mathematics consists exclusively in drawing conclusions from clearly stated premises; and that, in this process, it makes no difference what these premises signify, whether they are true or false, provided only that they do not contradict one another. But a person who has done productive mathematical work will talk quite differently. In fact those persons are thinking only of the crystallized form into which finished mathematical theories are finally cast. The investigator, however, in mathematics, as in every other science, does not work in this rigorous deductive fashion. On the contrary, he makes essential use of his phantasy and proceeds inductively, aided by heuristic expedients. One can give numerous examples of mathematicians who have discovered theorems of the greatest importance, which they were unable to prove. Should one, then, refuse to recognize this as a great accomplishment and, in deference to the above definition, insist that this is not mathematics, and only the successors who supply polished proofs are doing real mathematics? After all, it is an arbitrary thing how the word is to be used, but no judgment of value can deny that the inductive work of the person who first announces the theorem is at least as valuable as the deductive work of the one who first proves it. For both are equally necessary, and the discovery is the presupposition of the later conclusion."

> "There seems to be a great danger in the prevailing overemphasis on the deductive-postulational character of mathematics. ... A serious threat to the very life of science is implied in the assertion that mathematics is nothing but a system of conclusions drawn from definitions and postulates that must be consistent but otherwise can be created by the free will of the mathematician. If this description were accurate, mathematics could not attract any intelligent person. It would be a game with definitions, rules, and syllogisms, without motive or goal."

Courant and Robbins[160]

M. Postnikov [ZD07, p. 166] described the stages of development of new mathematics:

> "In general, mathematical work consists of three stages. The first stage is the most pleasant: thinking up the theorem and an idea for proving it. The second step – less pleasant, but still with an element of satisfaction – is to organize the proof, fill the gaps, check and simplify the computations, derive the necessary technical lemmas. The distasteful third stage – which, alas!, takes up 90% of one's time – consists in writing out the proof on paper in an understandable form. It is not surprising that many mathematicians shortchange this last stage, as a result of which their work becomes a mysterious puzzle, and one often finds it easier to think up a proof oneself than to figure out the author's text."

Many mathematicians would take issue with the distaste Postnikov expressed about the last step. I personally find considerable satisfaction and pleasure in producing a readable exposition of a mathematical result; admittedly it is much easier with the advent of word processing to do this today, and Postnikov's last two steps are generally more or less done simultaneously now. I also find that on average the quality of exposition in current mathematics publications is significantly better than in the past. But the separation of the development of the ideas and the formal exposition is something about which I think mathematicians would widely agree.

[160] What is Mathematics, 1941, p. xvii.

"Mathematics is regarded as a demonstrative science. Yet this is only one of its aspects. Finished mathematics presented in a finished form appears as purely demonstrative, consisting of proofs only. Yet mathematics in the making resembles any other human knowledge in the making. You have to guess a mathematical theorem before you prove it; you have to guess the idea of the proof before you carry through the details. You have to combine observations and follow analogies; you have to try and try again. The result of the mathematician's creative work is demonstrative reasoning, a proof; but the proof is discovered by plausible reasoning, by guessing. If the learning of mathematics reflects to any degree the invention of mathematics, it must have a place for guessing, for plausible inference."

G. Polya[161]

"[I]ntuition comes to us much earlier and with much less outside influence than formal arguments which we cannot really understand unless we have reached a relatively high level of logical experience and sophistication. ... In the first place, the beginner must be convinced that proofs deserve to be studied, that they have a purpose, that they are interesting."

G. Polya[162]

"If we say that a theorem has no meaning except as a conclusion from axioms, then do we say that Gauss did not know the fundamental theorem of algebra, Cauchy did not know Cauchy's integral formula, and Cantor did not know Cantor's theorem?"

Reuben Hersh[163]

"The object of mathematical rigor is to sanction and legitimize the conquests of intuition, and there was never any other object for it."

J. Hadamard[164]

"I believe the search for an explanation, for understanding, is what we should really be aiming for. Proof is simply part of that process, and sometimes its consequence."

M. Atiyah[165]

"Proofs really aren't there to convince you that something is true – they're there to show you why it is true."

Andrew Gleason[166]

[161] [Pól90a, p. vi].
[162] [Pól90a]; cf. [BB08, p. vii].
[163] *Some proposals for reviving the philosophy of mathematics*
[164] [Pól90a]; cf. [BB08, p. 10].
[165] *The Princeton Companion to Mathematics*, p. 1003.
[166] B. H. Yandell, *The Honors Class*, A K Peters, 2002, p. 150.

> "But the curious fact remained that, despite all the criticisms of the foundations of the subject [early Calculus], there could be no doubt but that the mathematical procedure was substantially right. In fact, the subject was right, though the explanations were wrong. It is this possibility of being right, albeit with entirely wrong explanations as to what is being done, that so often makes external criticism – that is so far as it is meant to stop pursuit of a method – singularly barren and futile in the progress of science. The instinct of trained observers, and their sense of curiosity, due to the fact that they are obviously getting at something, are far safer guides. Anyhow the general effect of the success of the Differential Calculus was to generate a large amount of bad philosophy, centring round the idea of the infinitely small."
>
> *Alfred North Whitehead*[167]

On the other hand, it is quite remarkable that most, if not all, true mathematical statements can be proved simply through formal deductions from accepted axioms, and elegant formal, or at least quasi-formal, proofs are generally very satisfying if not always very illuminating. I personally find this aspect of mathematics rather mysterious, but enormously appealing, and it gives formalism, especially the Hilbert version, much credibility as a *partial* explanation of the nature of mathematics.

> "Mathematical understanding develops by a mysterious interplay between intuitive insight and symbolic manipulation."
>
> *Robert Goldblatt*[168]

It has been said () that a typical mathematician is a Platonist during the week and a formalist on Sunday. He or she works all week formulating the problems and understanding the answer, all the time convinced that the problems and results concern real if abstract objects, and then at the end tries to write down a proof which is at least an outline of a formal argument, often leaving out or obscuring most of the ideas and intuition leading to the statement or proof in the first place. (It is fortunate that most mathematicians willingly, even eagerly, reveal such details in conversations and lectures, and sometimes in written surveys; published manuscripts are only part of the important communication which goes on in the mathematical community.)

This somewhat schizophrenic interplay of two disparate views of the nature of mathematics not only aids the progress of mathematics, but is psychologically useful too, and also convenient, as the next quote observes:

> "On foundations we believe in the reality of mathematics, but of course when philosophers attack us with their paradoxes we rush behind formalism and say: 'Mathematics is just a combination of meaningless symbols.' ... Finally we are left in peace to go back to our mathematics and do it as we have always done, with the feeling that each mathematician has that he is working with something real."
>
> *Jean Dieudonné*[169]

---

[167] [**?**, p. 169].
[168] [Gol98, p. v]
[169] J. Dieudonné, The work of Nicholas Bourbaki, *American Mathematical Monthly*, **77** (1970), 134-145.

"The major improvement in the modern viewpoint on axiomatics has been directed toward the elimination of intuition from proofs. It has been found that the only safe way to avoid intuition is to make its use impossible. This is accomplished by conscientiously refusing to know anything at all about the entities with which you are dealing be they called numbers, points, lines, or what have you, beyond what is stated explicitly about them in the axioms. Thus the entities with which a branch of mathematics is concerned, enter, in the first instance, as completely abstract, formless objects. Then a collection of axioms stating certain facts about these abstract objects is announced as the basis of the mathematical structure. These axioms are to be considered not as hints or clues as to the nature of the abstract objects with which you are concerned, but rather as the complete statement of *all* you know about them."

R. Kershner and L. Wilcox[170]

We conclude with an extensive quote which nicely summarizes the process of doing and communicating mathematics, and the healthy role formalism can play in the process:

"... [L]et us imagine that we wish to convince someone of the correctness of a certain statement. We may try to convince him by citing some authority in which he believes (sometimes called 'proof by intimidation'), by getting him to 'see' ('proof by intuition'), by appealing to his emotions ('proof by waving the red flag'), by confusing and tricking him with words and faulty logic ('elastic inference'), or by other methods familiar to everyone. These methods have two serious disadvantages: (a) They can be used to establish false statements just as easily as to establish true ones, and (b) They do not lead to universal agreement.

If we wish to convince a person of the correctness of a statement in such a way that our success does not depend on his weaknesses or our cleverness and so that our argument would be equally convincing to others, we must adopt a method of proof that meets with general agreement. Whatever this method is, it must certainly involve making statements, since it is hard to conceive of convincing anyone of anything without communicating with him. So we must carry on proof by means of statements. But suppose that the listener objects to some of our statements? Then we must convince him of them before we can continue. Now we can only do this by making other statements. Again if the listener does not agree with these, we should have to try to justify them – of course by making still other statements. Evidently, we cannot convince anyone of anything unless we can get him to agree to something! Apparently, we cannot prove all statements any more than we could define all terms. We must begin by assuming some statements without proof. ... [I]f we wish to prove any theorems we must begin by assuming some axioms.

But what if our listener does not agree that our axioms are truly laws? Is there no way out if we cannot obtain agreement on any axioms we propose? There is a way out. Namely we may ask our listener to agree merely that *if* our axioms are indeed laws, then our theorems are also laws. If we take this tack we eliminate argument about the truth of both the axioms and the theorems. The only question we ask our listener to agree on is whether the theorems really do follow from the axioms! Now this is precisely the procedure we follow in mathematics. Any formal mathematical theory begins with certain axioms and derives from them certain theorems. The mathematician

[170] *The Anatomy of Mathematics*, 1950, p. 26.

claims only that the theorems follow from the axioms, and he has no objection if someone prefers to adopt different axioms.

But suppose that our listener objects to the manner in which we derive the theorems from the axioms? Then we will have to come to some agreement with him as to what procedures are legitimate. And if we cannot agree with him on method of proof? Then we take the same way out as before. We say to him, 'Let us merely agree that if these axioms are accepted and if these methods of proof are used, then these theorems can be obtained.' In constructing a formal mathematical theory we state in advance the acceptable methods of deriving theorems from axioms. Then all we claim is that the theory is derived from the axioms according to these rules. If the axioms and the rules are acceptable to anyone, then the theorems should be also. If the axioms are applicable to any particular situation (i.e. are true in a particular case), then the theorems can be applied to that situation.

When we construct a theory on the basis of explicitly stated axioms and rules of proof, we have not really eliminated all possibility of controversy. There may, for one thing, be differences of opinion as to whether we have correctly applied the rules of proof. But this is just a question of whether we have made a mistake. Mistakes may be hard to find, but such differences of opinion can be solved by sufficiently careful examination of the theory. Unsolvable disagreement is still possible, however, on whether the axioms and rules of proof should have been adopted at all, whether the theory that results from them is a good theory, and so on. Such questions are not answered by the theory itself, but they are placed outside the theory when we agree to argue on the basis of the axioms and the methods of proof. Hence we may expect universal agreement *within* the theory, but no universal agreement *about* it. ... Evidently, we cannot eliminate disagreement or controversy, but we can construct a theory in such a way that disagreement is possible only about certain parts of it – namely the axioms and the methods of proof. This is a great advantage because it leads to universal agreement over a considerable area, avoids arguing about matters that can be agreed upon, and identifies the really controversial issues."

Kenneth O. May[171]

Proof Beyond a Reasonable Doubt

I.7.1.21. There is an important trend in the standard of rigor in mathematics which is emerging (or, more accurately, reemerging) largely as a result of the advent of powerful computers which can perform extensive and sophisticated mathematical computations. This point of view is articulated rather persuasively, and sometimes entertainingly, in [BB08], [BBG04], and [BD09]. I am, however, not yet personally convinced that this approach to rigor will, or should, be the future of mathematics. See also [Kra11] for an interesting discussion.

I have adopted the legal term "proof beyond a reasonable doubt" for this approach, although mathematicians undoubtedly have a different idea of what this phrase means than the courts. The idea is that it can sometimes be established with very high probability that a result is true, even in the absence of what would be considered a rigorous proof, and sufficiently high probability of truth should be enough for mathematicians to accept the result, although the search for a rigorous proof should not be abandoned.

Such "proof beyond a reasonable doubt" normally consists of a combination of an intuitive, heuristic, and/or nonrigorous argument for the plausibility of the result, along with extensive experimental or calculation evidence supporting the result (similar to standards for accepting a scientific theory).

[171] *Fundamental Mathematics*, preliminary edition, 1955, p. 75-78.

I.7.1.22. Such arguments are not new, and were quite common before the increased rigor of the nineteenth century. To illustrate, we consider a well-known example due to Euler. The problem was to compute

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \ .$$

This series was already "known" to converge (although the notion of convergence of an infinite series in the modern sense was not yet well understood) and that the numerical value of the sum was approximately 1.645; but the problem was to find what we would call a closed form expression for the exact sum.

Euler combined together some ingenious arguments and "known" facts to convince himself that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

(an outline of his argument is given below). He recognized that his argument was not a rigorous proof (even by the standards of the time), but simply gave a strong candidate for the exact value of the sum. He then calculated both the sum and $\frac{\pi^2}{6}$ to a number of decimal places (about 20), and found that the numbers agreed at least to the limits of his calculation. He took this agreement as definitive verification that the formula is correct (as indeed it is; cf. I.7.1.29.). Whether he regarded the argument plus numerical calculation as a "proof" of the formula is somewhat unclear; he subsequently made efforts to make parts of his argument more rigorous (see [Pól90a, p. 17-21] for a further discussion). Indeed, he did say:

> "The kind of knowledge which is supported only by observations and is not yet proved must be carefully distinguished from the truth; it is gained by induction, as we usually say. Yet we have seen cases in which mere induction led to error. Therefore, we should take great care not to accept as true such properties of the numbers which we have discovered by observation and which are supported by induction alone. Indeed, we should use such a discovery as an opportunity to investigate more exactly the properties discovered and to prove or disprove them; in both cases we may learn something useful."
>
> *L. Euler*[172]

It should be noted that although the modern definition of convergence of a sequence or series was not detailed until the nineteenth century, it would be unfair and inaccurate to suggest that Euler did not understand the difference between convergent and divergent series; in fact, he was acutely aware of the distinction. He said he had grave doubts as to the use of divergent series, but that he had never been led into error by using his definition of "sum".[173] Of course, few other mathematicians of the time, or any other time, have had Euler's awesome intuition and insight.

I.7.1.23. I am sure few if any modern mathematicians would regard what Euler did as a rigorous proof of the formula. After all, there are known examples of potential formulas which are supported by heuristic arguments and which are accurate to more decimal places than Euler calculated, but turn out not to be exact (some examples can be found in [**?**]; see also III.9.4.3. and V.14.4.4.).

[172] *Opera Omnia*, ser. 1, vol. 2, p. 459; cf. [Pól90a, p. 3].
[173] Letter to N. Bernoulli; cf. [**?**, p. 322].

However, suppose EULER had had access to a supercomputer and found that the sum agreed with $\frac{\pi^2}{6}$ to a million, or a billion, decimal places. Would we take this as compelling evidence of the truth of the formula? I imagine most mathematicians still would not. But is the chance that the result fails in light of such evidence really greater than the chance that a standard proof of some standard theorem in mathematics contains a subtle flaw so far unnoticed by mathematicians? One cannot say *for certain* that even the best-known "rigorous" mathematical proofs cannot have some inadequacy or error.

I.7.1.24. An interesting example concerns the classification of finite simple groups, which was one of the crowning achievements of mathematics in the late twentieth century. Even the statement of the result is far too complicated to give here, and the full proof (at least at present) consists of many research papers totaling about 15,000 printed pages. See [GLS94] for a survey of the project and result.

Group theorists take a curious combination of positions on this statement. They generally claim that the result is proved, i.e. a theorem, while at the same time acknowledging that it is virtually certain that some parts of the 15,000-page proof contain errors or at least gaps. *****UPDATE*****

The Role of Computers

> "The computer is a precision tool. It should not be used as a bludgeon or a substitute for thought."
>
> *Forman S. Acton*[174]

I.7.1.25. It is unquestionably true that computers have had a significant and growing influence on the practice of mathematics. It is a matter of some disagreement whether this influence has been entirely beneficial.

There is universal agreement that computers can be very useful in many parts of mathematics, especially applied mathematics. They have long been crucial for numerical calculations needed to turn theoretical results into practical solutions to many problems. They are also valuable in "experiments" testing the plausibility of conjectures, and in helping mathematicians analyze behavior of complex mathematical systems and develop new conjectures. They can even find and identify reasonable conjectures (and not just numerical ones) which would be difficult or impossible for a human mathematician to come up with, and, perhaps even more importantly, refute some apparently reasonable conjectures.

More controversial is the use of computers to actually carry out parts of arguments constituting actual proofs (or at least purported proofs) of mathematical results. The most famous case of this is the proof of the Four-Color Theorem; see (), () for extensive discussions of this example. This theorem is significant at least in part because it shows that there are theorems in mathematics whose statement is simple enough to be understood by almost anyone, but whose proof is necessarily so long and complicated as to be beyond the ability of any human mathematician to completely understand it or even completely verify its correctness. (There is, of course, the possibility that a simple proof of the Four Color Theorem will eventually be found, but although there have been some simplifications in the proof in the forty-some years since its publication, so far no version comes close to being simple enough that it can be checked by a human mathematician.)

Computers and computation have also had an influence on the mindset of mathematics (or at least of mathematicians). An algorithmic approach to much of mathematics has come more into vogue. There have even been predictions that the "age of infinity" in mathematics is drawing to a close! Such an assertion is

[174] [Act90, p. 246].

preposterous, in my opinion: while the algorithmic and computational parts of mathematics will undoubtedly grow greatly in importance and sophistication, the mathematics of infinity, including most of analysis, is and will remain far too important and central to fade away. What will instead happen is that the interplay between these two complementary parts of mathematics will become ever greater and more essential.

> "[A]lthough we are all conscious of the importance of the digital computer (sometimes to the point of hysteria), and indeed are inundated with advertising hyperbole from the most diverse quarters about all the wonders supercomputers will accomplish, many are much less conscious of what is ultimately an even more important fact: the computer is as much a *problem* as it is a *tool*. We must understand the nature and limitations of this most powerful of human tools. It is important to know what *cannot be computed* and the dangers of what can be miscomputed."
>
> *Felix Browder*[175]

Fraud in Mathematics

The nature of mathematics, at least in its present state, has one nice consequence: almost uniquely among academic subjects, there is almost no possibility of fraud in mathematics. (This may change with the spread of large-scale computation as an integral part of mathematics.) Just about the only type of fraud conceivable in mathematics today is plagiarism, and even this is usually hard to get away with since the world of mathematics is generally so small and interconnected that experts in a subdiscipline usually have a pretty good idea of what all the other experts are doing, although it is still possible for someone to represent an unpublished idea of a colleague or student as their own. And nowadays most new mathematical results are quickly written up and posted on the arXiv in preprint form, unlike in sciences where confidentiality of a paper is often jealously guarded until actual publication; as a result priority for new results in mathematics can rarely be disputed. (Correctness and completeness of purported proofs of new results may be another matter!)

Euler's Argument

We now outline Euler's argument for the above formula.

I.7.1.26. He began with the already "known" formula

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

which is just the Maclaurin (Taylor) series, and is valid for all $x \in \mathbb{R}$ (the validity of this formula had not yet been proved to modern standards). Thus if we set

$$f(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!}$$

we have that

$$f(x) = \frac{\sin x}{x} \text{ for } x \neq 0 \ .$$

[175] *Mathematics and the Sciences*, [AK88].

I.7.1.27. Next, he observed that if $p(x)$ is a polynomial of degree m with roots $a_1, \dots, a_m$, all nonzero (but not necessarily distinct), then

$$p(x) = p(0) \prod_{n=1}^{m} \left(1 - \frac{x}{a_n}\right) .$$

Thus, since the roots of $f(x)$ are $\{\pm\pi, \pm 2\pi, \pm 3\pi, \dots\}$, he concluded by analogy that

$$\sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n+1)!} = f(x) = f(0) \prod_{n=1}^{\infty} \left[\left(1 + \frac{x}{n\pi}\right)\left(1 - \frac{x}{n\pi}\right)\right] = \prod_{n=1}^{\infty} \left(1 - \frac{x^2}{n^2\pi^2}\right)$$

since $f(0) = 1$. (This formula turns out to be correct, but the proof of correctness needs a good bit more work than EULER did.)

I.7.1.28. Finally, he multiplied out the right side as though it was a finite product, and compared the coefficients of powers of x on both sides of the equation. In particular, the coefficient of x^2 on the right side is $-\sum_{n=1}^{\infty} \frac{1}{n^2\pi^2}$, and on the left it is $-\frac{1}{6}$, giving the desired formula. (Comparison of coefficients of other powers of x gave additional formulas for other infinite series.)

A Proof of Euler's Formula

I.7.1.29. For completeness, we give a slick proof of the formula, from [LeV56] (where originality is not claimed). This proof and a couple of others, including one closer in spirit to EULER'S argument, can be found in [AZ10]. The argument uses some theorems from integration theory; the special cases used can easily be established directly.

Set

$$f(x, y) = \frac{1}{1 - xy} .$$

Let R be the unit square $[0, 1]^2$ in $\mathbb{R}^2$. We will calculate

$$I = \iint_R f \, dA$$

two ways. Note that this is an improper Riemann integral since f has a singularity at $(1, 1)$; but the integrand is nonnegative on R, so the meaning of the integral is unambiguous.

The infinite series (geometric series)

$$\sum_{k=0}^{\infty} (xy)^k$$

with nonnegative terms converges pointwise to f on $R \setminus \{(1, 1)\}$; thus by the Monotone Convergence Theorem () we have

$$I = \sum_{k=0}^{\infty} \left[\iint_R (xy)^k \, dA\right] .$$

By the continuous version of Fubini's Theorem (IX.3.4.7.), these integrals can be calculated as iterated integrals:

$$\iint_R (xy)^k \, dA = \int_0^1 \int_0^1 x^k y^k \, dy \, dx = \left[\int_0^1 x^k \, dx\right] \left[\int_0^1 y^k \, dy\right] = \frac{1}{(k+1)^2}$$

and thus

$$I = \sum_{k=0}^{\infty} \frac{1}{(k+1)^2} = \sum_{n=1}^{\infty} \frac{1}{n^2} \ .$$

On the other hand, I can be calculated by a change of variable

$$x = u - v, \quad y = u + v$$

i.e. by the Change of Variables Theorem () with $\phi(u, v) = (u - v, u + v)$ we have

$$I = \iint_R f \, dA = \iint_D (f \circ \phi) |J_\phi| \, dA$$

where $D = \phi^{-1}(R)$ is the diamond with vertices $(0, 0)$, $\left(\frac{1}{2}, \frac{1}{2}\right)$, $\left(\frac{1}{2}, -\frac{1}{2}\right)$, $(1, 0)$. We have that

$$J_\phi(u, v) = det \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = 2$$

for all (u, v) (the change of variables is linear, making things easy), so we have

$$I = 2 \iint_D (f \circ \phi) \, dA \ .$$

We have that

$$(f \circ \phi)(u, v) = \frac{1}{1 - u^2 + v^2}$$

and by Tonelli's Theorem () the double integral over D can be computed by an iterated integral:

$$I = 2 \left[\int_0^{1/2} \left(\int_{-u}^{u} \frac{1}{1 - u^2 + v^2} \, dv \right) du + \int_{1/2}^{1} \left(\int_{-(1-u)}^{1-u} \frac{1}{1 - u^2 + v^2} \, dv \right) du \right] \ .$$

(Note that the singularity of f at $(1, 1)$ "disappears" under this transformation.) We have

$$\int \frac{1}{1 - u^2 + v^2} \, dv = \frac{1}{\sqrt{1 - u^2}} \arctan\left(\frac{v}{\sqrt{1 - u^2}} \right) + C$$

for $0 \leq u < 1$, so we have

$$I = 4 \int_0^{1/2} \frac{1}{\sqrt{1 - u^2}} \arctan\left(\frac{u}{\sqrt{1 - u^2}} \right) du + 4 \int_{1/2}^{1} \frac{1}{\sqrt{1 - u^2}} \arctan\left(\frac{1 - u}{\sqrt{1 - u^2}} \right) du \ .$$

The first integral can be evaluated by the substitution

$$s = \arctan\left(\frac{u}{\sqrt{1 - u^2}} \right), \quad ds = \frac{1}{\sqrt{1 - u^2}} \, du$$

to obtain

$$\int_0^{1/2} \frac{1}{\sqrt{1 - u^2}} \arctan\left(\frac{u}{\sqrt{1 - u^2}} \right) du = \int_0^{\arctan(1/\sqrt{3})} s \, ds = \int_0^{\pi/6} s \, ds = \frac{1}{2} \left(\frac{\pi}{6} \right)^2 \ .$$

Similarly, the second integral can be calculated by the substitution

$$t = \arctan\left(\frac{1-u}{\sqrt{1-u^2}}\right) = \arctan\left(\sqrt{\frac{1-u}{1+u}}\right), \quad dt = -\frac{1}{2}\frac{1}{\sqrt{1-u^2}}\,du$$

to obtain

$$\int_{1/2}^{1} \frac{1}{\sqrt{1-u^2}} \arctan\left(\frac{1-u}{\sqrt{1-u^2}}\right) du = 2\int_0^{\pi/6} t\,dt = \left(\frac{\pi}{6}\right)^2 .$$

Thus we have

$$I = 4\cdot\frac{1}{2}\left(\frac{\pi}{6}\right)^2 + 4\left(\frac{\pi}{6}\right)^2 = 6\left(\frac{\pi}{6}\right)^2 = \frac{\pi^2}{6} .$$

Mathematicians on Philosophy

I.7.1.30. Mathematicians tend to look down on philosophy, sometimes excessively, as a subject unworthy of their time and intellect. The only aspect of philosophy which is of any widespread interest among mathematicians is the foundations of mathematics, and even this interest is pretty limited. Here are some of the opinions expressed about philosophy; I emphasize that I do not personally endorse them, but just present them in the interest of full disclosure:

> "[P]hilosophy proper is a subject, on the one hand so hopelessly obscure, on the other so astonishingly elementary, that there knowledge hardly counts."
>
> *G. H. Hardy*[176]

> "Philosophers may be unable to understand how we can know mathematical truths, but they have a bad track record at being able to understand anything."
>
> *J. R. Lucas*[177]

> "Philosophy, from the earliest times, has made greater claims and achieved fewer results than any other branch of learning."
>
> *Bertrand Russell*[178]

> "Mathematics makes a steady advance, while philosophy continues to flounder in unending bafflement at the problems it confronted at the outset."
>
> *Michael Dummett*[179]

[176] *Mathematical Proof*, in [**?**, p. 173].
[177] [Luc00, p. 19].
[178] *Our Knowledge of the External World*, cf. [**?**, p. 3].
[179] [**?**].

"... People who study philosophy too long, and don't treat it simply as part of their early education and then drop it, become, most of them, very odd birds, not to say thoroughly vicious; while even the best of them are reduced by this study ... to complete uselessness as members of society."

Plato[180]

A joke told by philosophers: a man in a cafe orders a cup of coffee without cream. The waiter replies, "I'm sorry, sir, but we're out of cream. How about a cup of coffee without milk?"

Should Mathematicians Care About Foundations?

"Mathematics changes slowly. After ten years, some problems have been solved and some new problems have arisen, but the subject is recognisably the same. However, over a century it changes in ways which mathematicians at the beginning of that century could not conceive. ...

Given this, it seems presumptuous to seek a permanent foundation for mathematics. It is a happy accident that almost all of mathematics, as we now know it, can be deduced from one set of axioms; but, just as history has taught us that there is not one geometry but many, so it is not unlikely that the future will present us with different mathematics deduced from radically different axiom schemes. The art of the mathematician will still consist in the rigorous proof of unexpected conclusions from simple premises, but those premises will change in ways that we cannot possibly imagine."

T. W. Körner[181]

I.7.1.31. The question of which, if any, of the philosophies of the foundations of mathematics is correct has rather little to do with the day-to-day work of most mathematicians. Mathematicians generally just take for granted that mathematics has some sort of logical, consistent foundation, whether it is understood yet or not. I think it is fair to say that the only foundational issue that does, or probably even should, concern every mathematician is the consistency of mathematics.

"God exists since mathematics is consistent, and the Devil exists since we cannot prove it."

André Weil[182]

Even if it were shown that, say, ZF set theory is inconsistent, most mathematicians would be at least moderately interested, but not particularly upset; this would just mean that ZF set theory was not the right theory to base mathematics on after all. They would then simply go on doing their mathematics as usual.

The only thing that would really upset mathematicians is if mathematics itself (e.g. real analysis as axiomatized in (), or geometry, algebra, etc.) were discovered to be inconsistent. Mathematicians mostly just take on faith that this is an impossibility; indeed, belief in some form of realism would preclude this possibility. After all, the very existence of mathematics as a subject is entirely based on the belief that

[180] *The Republic*, VII.3; cf. [**?**].
[181] [Kör04, p. 356].
[182] In H. Eves, *Mathematical Circles Adieu*, Boston: Prindle, Weber and Schmidt, 1977.

if one starts with correct statements and applies valid, careful reasoning, correct conclusions will always be obtained. It seems inconceivable to us that an inconsistency could show up in mathematics, especially since the massive edifice already created has produced no hint of inconsistency. At the very core of every mathematician is a certain conviction of the consistency of mathematics, reinforced by all the mathematics they see and do throughout their careers.

> "If mathematical thinking is defective, where are we to find truth and certitude?"
>
> *D. Hilbert*[183]

> "We believe that mathematics is destined to survive, and that the essential parts of the majestic edifice will never collapse as a result of the sudden appearance of a contradiction; but we do not maintain that this opinion rests on anything but experience. That is not much, it may be said. But for twenty-five centuries mathematicians have been correcting their errors, and seeing their science enriched and not impoverished as a consequence; and this gives them the right to contemplate the future with equanimity."
>
> *N. Bourbaki*[184]

I.7.2. Mathematical Theories

I.7.3. Languages

Any theory must be expressed in some sort of language. Actually, it is generally accepted that there must be at the outset some sort of *metalanguage* which is simply understood, and in which certain phrases and terms have meaning. The metalanguage is not part of the theory, or really part of mathematics at all; it is just a vehicle for communication and specification of the theory. In [Kle02], KLEENE calls this the *observer's language* (with the term "metalanguage" reserved for an observer's language with certain technical restrictions), which seems apt terminology.

In addition, a mathematical theory must have a specified language in which it is expressed, called the *object language* in [Kle02]. At a minimum, the language of a theory must define some collection of "statements" which are considered "meaningful" for the theory (although no actual meaning in the usual sense of the term need be attached to them), and there needs to be a way to combine together statements or transform them into new statements according to specific rules.

There does not seem to be any necessity that the language consist of finite one-dimensional strings of discrete symbols (human spoken language is an obvious exception), but languages in mathematics, called *formal languages*, are normally taken to be of this sort: an *alphabet* of symbols is specified, and certain finite strings of these symbols are the "statements." This restriction is necessary (at least at present) to make formulation and analysis of mathematical theories manageable. It can be argued that formal languages can satisfactorily encode anything which could be called a "language", although it must be noted that even ordinary written language is at best an imperfect representation of human spoken language; in spoken language there is a lot of subtle information conveyed in timing, stress and intonation, facial expressions, gestures, and body language which is not captured in the written language (even transmission of spoken language over a telephone or radio causes significant losses of information).

[183] *On the Infinite*, cf. [**?**].
[184] cf. [**?**, p. 132].

In a good formal language we need to have the ability to make *definitions*. In effect, a definition is simply the establishment of an "abbreviation" for a (generally more complicated) phrase. In principle, the defined term can always be replaced by the phrase it abbreviates without loss of utility, and thus from a strictly logical point of view definitions are unnecessary. However, in practice it is very useful, often essential, to be able to make definitions: otherwise, statements involving even moderately complex mathematical ideas would become unmanageably or even incomprehensibly complicated.

I.7.4. Logic

> "Logic is the systematic study of the conditions of valid inference."
>
> *Abraham Wolf*[185]

> "Logic is the systematic study of the structure of propositions and of the general conditions of valid inference by a method which abstracts from the content or *matter* of the propositions and deals only with their logical *form*."
>
> *Alonzo Church*[186]

> Logic is a discipline which is not merely descriptive, but normative as well, being concerned more with how we *ought* to reason than with how we habitually do so. ... We shall seek to construct an ideal theory of correct reasoning, based on a small number of general principles which we postulate as 'laws of thought.' "
>
> *G. T. Kneebone*[187]

> "Few persons care to study logic, because everyone conceives himself to be proficient enough in the art of reasoning already."
>
> *C. S. Peirce*[188]

As with languages, a distinction must be made between what may be called *metalogic* (this term is not commonly used in the literature), which is an underlying feature, and *logic*, which is a part of the mathematics. Unlike with language, this distinction is often not carefully made and has led to much confusion and disagreement over the last century, with arguments about whether logic was part of mathematics or vice versa (or neither). The subject of *mathematical logic* has since become established.

Metalogic can be thought of as the rules of inference which are assumed to underly not only mathematics but all rational discourse. An example of a principle of metalogic is transitivity of implication:

> If P implies Q and Q implies R, then P implies R.

(Here P, Q, and R are statements.) Another example, called *modus ponens*, is

> If P is true and P implies Q, then Q is true.

To rephrase *modus ponens* without using the loaded term "true", we could say instead:

> If we assume or conclude P and we assume or conclude (P implies Q), then we conclude Q.

[185] *Encyclopaedia Britannica*, fourteenth edition, 1929; cf. [**?**, p. 5].
[186] *Encyclopaedia Britannica*, ? edition, 1959; cf. [**?**, p. 6].
[187] [**?**, p. 8].
[188] *Collected Papers*; cf. [**?**].

I.7.5. Mathematical Structures and Models

I.7.6. Exercises

I.7.6.1. Here is a fun logic puzzle. You have 12 steel balls which appear identical, but one of them is a fake whose weight is slightly different than the other eleven, either slightly more or slightly less. The difference is too slight to be felt by hand, but you have a balance scale which is sensitive enough to detect it and sturdy enough to handle several balls on each side. (A balance scale has two trays and will tell which side is heavier or whether they are the same.) In three weighings (comparisons), determine which is the fake and whether it is too heavy or too light. [Hint: begin by dividing the balls into three groups, and number the balls in each group A1–A4, B1–B4, C1–C4. In the first "weighing," compare the A balls with the B balls. The trickiest part of the problem is to decide how to do the second "weighing" if the first one does not balance.]

I.8. Printed or Online?

When this book project began, I envisioned the result as a traditional printed volume. But I fairly quickly became convinced that it should be an online book instead. An online book has at least four enormous advantages over a printed one:

(1) An online book is much more widely, easily, and freely available to potential readers. Production, reproduction, and distribution costs are almost nonexistent. The path from author to reader is direct without having to pass through a publisher.

(2) Unlike with a printed book, there is no practical limit to the length of an online book. Redundancy (within reason) does not have to be avoided.

(3) An online book does not have to be organized linearly.

(4) Revisions, corrections, and additions can be made immediately and painlessly.

Advantage (1) is obvious and needs no further comment. As to (2), an author of a printed book usually has to make difficult and delicate decisions about what to include and what to omit, and how much detail should be included. A printed book with too much material or detail not only becomes physically unwieldy and expensive to produce, but also creates psychological problems for the reader: it can be difficult for a reader to plow through a mass of detail to extract the essentials he/she needs, and an abundance of ancillary material can be intimidating to a reader not yet familiar with the subject and can make the theory appear to be more complex and difficult than necessary.

An online book, if properly organized, avoids most (but not all) of these problems. Extra details, technicalities, and side threads can be included in such a way that they are invisible to readers not specifically interested in looking at them. The book can consist of a large number of separate sections connected only by links organized by a complicated directed graph. Each section can have a set of links to the prerequisite sections and another set of links to potential successors. (However, one serious danger in a non-linear organization of a mathematics book is the possibility of circularity in definitions and arguments; great care must be exercised to avoid circularity.)

A reader of a book organized in this way will probably need some guidance through the material, a "road map." Actually, several road maps will be required for readers of various backgrounds and interests, and I plan to provide some. If this book is used as a text or supplementary text for a course, the instructor will have to settle on a road map appropriate for the course, or provide his/her own. Every instructor has a slightly different idea about how the material of the subject should be organized and presented. This book by and large represents my idea, but I am trying to make the organization flexible enough that an instructor who wants a different flow through the material can develop a satisfactory road map for the desired journey.

One advantage related to (4) is that the book can be placed online before it is "finished," so long as it is far enough along to be useful. (The currently posted version is already probably about the 1000'th edition; it is likely that most of these editions have not been read by anyone but me!) In fact, my expectation is that I will never regard the book as finished. I have a dream that this book will eventually turn into an all-encompassing treatment of real analysis. Of course, I am far from having the personal expertise, or enough time and energy left in my lifetime, to accomplish this goal by myself. I hope eventually that others will contribute to the project by adding treatments of additional material and/or alternative treatments of material already covered. I hope also that experts interested in other parts of mathematics, especially parts of mathematics that are mature enough that a reasonably definitive exposition can be given, will undertake

similar projects in their specialties, and that eventually a massive and interconnected body of knowledge and understanding of a large part of mathematics will exist. Previous attempts have been made along this line, notably by BOURBAKI, but I do not regard my efforts or goals to be at all comparable or duplicative of such work. The current and rapidly expanding treatment of mathematics in Wikipedia and other websites is laudable and very useful, but also not comparable in goals. (Wikipedia articles on some other subjects, especially ones with political or nationalistic overtones, are notoriously unreliable; however, I have found most Wikipedia articles on mathematics to be quite good, although they should not be taken as gospel. Articles on mathematicians and/or the history of mathematics are not always as good.)

It must be said that even in an online book of potentially infinite length, the necessity of making decisions about what to include and what to omit is not completely avoided unless one plans to cover all of mathematics. The decision is particularly critical in regard to material which is ancillary to the main topic of the book, such as the intricacies of set theory or topology in my case. I have necessarily employed subjective standards based on my personal interests and limitations, and I have included more detail than necessary on some topics I like, and (at least so far) much too little detail on other topics. At least I can rationalize some of these decisions by saying I have not *yet* finished a complete treatment of the subject. Or I suppose I could just invoke the words of DESCARTES [**?**]:

> "I hope that posterity will judge me kindly, not only as to the things of which I have explained, but also as to those which I have intentionally omitted so as to leave to others the pleasure of discovery."

I expect that many, if not most, sections of this book will eventually have "short" ("appetizer") and "long" ("main course") versions. The short version will generally be more of a survey of the high points of the specific topic, with most proofs and technical details suppressed, and sometimes with definitions and/or statements of results stated somewhat informally. (Current versions of many sections probably fall somewhere between the ultimate short and long versions.) I will insist, though, that even the short treatments be mathematically "honest."

There are two common kinds of mathematical "dishonesty" frequently encountered in lower-level (and sometimes higher-level) texts:

(1) Suggesting that a certain mathematical statement is true more generally than it actually is.

(2) Suggesting that a certain mathematical statement is "obvious" and does not require proof, when in fact it does.

A variation of (2), related to (1), is

(2′) Suggesting that an informal definition of something (e.g. tangent line, area) is mathematically adequate.

A third kind of mathematical dishonesty, somewhat less common but still frequently seen, and arguably more serious than (1) and (2), is

(3) Giving a purported "proof" of something which is not complete or not rigorous, while suggesting it is.

> "Few will deny that even in the first scientific instruction in mathematics the most rigorous method is to be given preference over all the others. Especially will every teacher prefer a consistent proof to one which is based on fallacies or proceeds in a vicious circle, indeed it will be morally impossible for the teacher to present a proof of the latter kind consciously and thus

> in a sense deceive his pupils. Notwithstanding these objectionable so-called proofs, so far as the foundation and the development of the system is concerned, predominate in our textbooks to the present time. Perhaps it will be answered, that rigorous proof is found too difficult for the pupil's power of comprehension. Should this be anywhere the case, – which would only indicate some defect in the plan or treatment of the whole, – the only remedy would be to merely state the theorem in a historic way, and forego a proof with the frank confession that no proof has been found which could be comprehended by the pupil; a remedy which is ever doubtful and should only be applied in the case of extreme necessity. But this remedy is to be preferred to a proof which is no proof, and is therefore wholly unintelligible to the pupil, or deceives him with an appearance of knowledge which opens the door to all superficiality and lack of scientific method."
>
> *H. Grassmann, 1904*[189]

In my opinion, as a general rule a written argument should not be called a "proof" unless it is complete and rigorous (although I must admit that I myself have not always adhered to this principle!)

I certainly do not mean to suggest that it is necessary, or even desirable, that all the technicalities, ramifications, or potential pathologies connected with a particular topic being discussed in an elementary text be treated or even mentioned. However, I believe that mathematical honesty is possible even in, say, a Business Calculus text, although maintaining mathematical honesty while at the same time keeping the discussion at a reasonable and appropriate level of detail and sophistication is a fine art and not easily accomplished.

I.8.1. How Much Generality?

Another constantly encountered problem in preparing an exposition of real analysis (or of many other parts of mathematics) is the question of whether to do things in the utmost generality, or stick to an important special case. There is no simple answer to this dilemma, and I have found it impossible (as well as undesirable) to be entirely consistent.

There are several considerations, which are often in tension:

1. It is desirable to have results recorded in the greatest generality that can be reasonably stated, for future reference, even if the immediate applications one has in mind require only a special case.
2. It is often easier to understand and appreciate a proof if the statement of the result contains the exact hypotheses needed to make the proof work.

On the other hand:

3. The proof of the most general case of a result can often be considerably harder or more delicate than that of a crucial special case.
4. Even the statement of the most general result can be much harder to write or digest than a clean statement of a special case.
5. It may seem farfetched in some cases that increased generality would ever be useful, especially if it requires more work to state or prove than the result of obvious importance.

I have had to make many decisions of this type in the manuscript, and have each time had to settle on the best tradeoff between generality, difficulty, and usefulness. My decisions may not have always been the wisest, and will be subject to review and possible revision in the future.

[189] cf. [Mor93, #538].

I.8.2. History vs. Logic

> "Once a good building is completed its scaffolding should no longer be seen."
>
> *C. F. Gauss*[190]

> "Whoever wants to get to know a science shouldn't just grab the ripe fruit – he must also pay attention to how and where it grew."
>
> *J. C. Poggendorf*[191]

There are two competing schools of thought as to how mathematics, especially "finished" mathematics such as Real Analysis, should be presented:

(1) The ideas and results should be presented in the most efficient, elegant, and logical order, without regard to their origins or history.

(2) The historical development and context of all results and ideas should be carefully traced and presented.

The ultimate exponent of approach (1) is Bourbaki. There are numerous books on Analysis taking the second approach. Most textbooks basically take the first approach, with some lip service to the second.

New mathematics is usually not very elegant at first. The first proof of a result is rarely the most efficient one or done in the optimum generality, and historically it has often happened that the first "proof" of an important result is not even complete or correct. In fact, Besicovitch once said that a mathematician should be judged by how many bad proofs he has produced (he meant to imply a positive correlation). It usually takes time and effort, often quite a bit of it, before the "proper" approach or point of view for a part of mathematics is developed.

> "The ways of discovery must necessarily be very different from the shortest way, indirect and circuitous, with many windings and retreats. It's only at a later stage of knowledge, when a new domain has been sufficiently explored, that it becomes possible to reconstruct the whole theory on a logical basis, and show how it might have been discovered by an omniscient being, that is, if there had been no need for discovering it!"
>
> *George Sarton*[192]

I have tried to steer somewhat of a middle ground between the two approaches, although the result is of necessity much closer to the first than the second since I am a mathematician, not a historian. But I have made considerable efforts to accurately put many things in historical context, not only because it is beneficial in understanding where the mathematics (and terminology) came from, but also because much of the history is fascinating. I have particularly tried to correctly trace the history of some things which I discovered have been commonly misportrayed in modern references. I am certainly not the first to try to do this: for example, in [Kle04, p. 234], Klein points out that what is (even today!) commonly called the "Maclaurin series" had been previously obtained by Taylor, explicitly stated as a special case of the Taylor series, and was acknowledged as such by Maclaurin himself. Klein states:

[190] cf. [Rem98, p. viii].
[191] cf. [Rem98, p. vii].
[192] *The Study of the History of Mathematics*, p. 19.

> "Errors of this sort are common. ... It is difficult, if not impossible, to overcome such deeprooted absurdities. At best, one can only spread the truth in the small circle of those who have historical interests."

I hope this is unduly pessimistic!

Despite my efforts along these lines, I am sure my amateur historical research is highly inadequate in many cases and probably inaccurate in some respects. I will endeavor to gradually improve it as I learn more.

Chapter II

The Formalities of Mathematics

In this chapter, we have collected together discussions of the basic language and tools of mathematics (not just analysis). These vary widely in sophistication, ranging from elementary (from the point of view of modern mathematics), e.g. formal statements, basic set theory, functions, and relations, to very advanced, e.g. the intricacies of the Axiom of Choice, ordinals, and large cardinals. The sections in this chapter can be regarded primarily as reference material for our treatment of analysis.

II.1. Formal Statements, Negations, and Mathematical Arguments

"Mathematics is often considered a difficult and mysterious science, because of the numerous symbols which it employs. Of course, nothing is more incomprehensible than a symbolism which we do not understand. Also a symbolism, which we only partially understand and are unaccustomed to use, is difficult to follow. In exactly the same way the technical terms of any profession or trade are incomprehensible to those who have never been trained to use them. But this is not because they are difficult in themselves. On the contrary they have invariably been introduced to make things easy. So in mathematics, granted that we are giving any serious attention to mathematical ideas, the symbolism is invariably an immense simplification."

A. N. Whitehead[1]

"Every human pastime–music, cooking, sports, art, theoretical physics–develops an argot to spare its enthusiasts from having to say or type a long-winded description every time they refer to a familiar concept in each other's company. The problem is that as we become proficient at our job or hobby we come to use these catchwords so often that they flow out of our fingers automatically, and we forget that our readers may not be members of the clubhouse in which we learned them."

Steven Pinker[2]

"The domain over which the language of analysis extends its sway, is, indeed, relatively limited, but within this domain it so infinitely excels ordinary language that its attempt to follow the former must be given up after a few steps. The mathematician, who knows how to think in this marvelously condensed language, is as different from the mechanical computer as heaven from earth."

A. Pringsheim[3]

"The prominent reason a mathematician can be judged by none but mathematicians, is that he uses a peculiar language. The language of mathesis is special and untranslatable. In its simplest form it can be translated, as, for instance, we say a right angle to mean a square corner. But you go a little higher in the science of mathematics, and it is impossible to dispense with a peculiar language. It would defy all the power of Mercury himself to explain to a person ignorant of the science what is meant by the single phrase 'functional exponent.' ... But to one who has learned this language, it is the most precise and clear of all modes of expression. It discloses the thought exactly as conceived by the writer, with more or less beauty of form, but never with obscurity."

Thomas Hill[4]

[1] *Introduction to Mathematics*, 1911, p. 59-60; cf. [Mor93, 1218].
[2] *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century.*
[3] *Jahresberichte der Deutschen Mathematiker Vereinigung*, Bd. 13, p. 367; cf. R. Moritz, *Memorabilia Mathematica*, 1211.
[4] *North American Review*, v. 85, p. 224-225; cf. R. Moritz, *Memorabilia Mathematica*, 1210.

HOBBES: "[The page] is so covered over with the scab of symbols, that I had not the patience to examine whether it be well or ill demonstrated."

WALLIS: "Is it not lawful for me to write Symbols, till you can understand them? Sir, they were not written for you to read, but for them that can."

1656[5]

Mathematical statements are typically given in a hybrid of symbolic language and standard English. For example:

"For every real number x there is a unique integer n such that $n \leq x < n + 1$."

(Of course, comparable statements can be made equally well using other languages than English.) It is fairly easy to understand what a statement like this means, but many mathematical statements, particularly in analysis, are considerably more complicated, such as:

"For every real number a and every $\epsilon > 0$ there is a $\delta > 0$ such that, for every real number x with $|x - a| < \delta$, $|f(x) - f(a)| < \epsilon$."

(Some statements are much more complicated than this!) We need to have a systematic way to analyze the meaning of a statement like this. In particular, it is often important to be able to write (and understand) the negation of such a statement. A very useful tool for these problems is to rewrite mathematical statements in a language-independent formal way. This section describes how to do this.

The presentation in this section is intended for students preparing to be working mathematicians, and we will not strictly adhere to the (sometimes rather arcane) conventions of the subject of mathematical logic. Thus our notation and terminology will not necessarily agree with that found in texts on logic.

II.1.1. Formal Statements

II.1.1.1. A formal statement equivalent to the second one above is

$$(\forall a \in \mathbb{R})(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in \mathbb{R})[(|x - a| < \delta) \Rightarrow (|f(x) - f(a)| < \epsilon)] \ .$$

We will explain what this formal statement means and how to get it.

II.1.1.2. A formal statement (or proposition) will be a string of symbols; however, not every string of symbols will be regarded as a formal statement. The allowable symbols will be of three types: *constants*, *variables*, and *logical symbols*. A formal statement will then be, roughly speaking, a string of allowable symbols which is in a form to which a truth value (true or false) can potentially be attached, perhaps depending on the values assigned to the variables in the statement. We will in due course be more precise.

[5] [KK03, p. 85-86].

Constants

II.1.1.3. Constants are symbols standing for previously defined specific objects. The objects may be elements, sets, ordered pairs, functions, relations, binary operations, or any other type of object specifically defined in the theory being discussed (all the types of objects listed are officially sets, so one could just say that the objects are all sets to be precise). Additional constants can be defined in the course of the discussion. Note that some standard objects in mathematics have names consisting of compound symbols, such as sin, $\frac{d}{dx}$, or $\int(\cdot)\,dx$, which are regarded as single constants. Note also that a function which is a constant in this sense (i.e. simply a specified function) is generally not a "constant function" in the function sense of taking the same value at any input.

Variables

II.1.1.4. Variables are similar to constants, but are symbols which have not been previously defined to represent specific objects, and thus can potentially represent many different objects. The *type* of object that a particular variable represents can be specified, and can be any of the types of objects that can be constants.

The distinction between variables and constants can be a subtle one. For example, in the statement in II.1.1.1., the f might be either a constant or a variable depending whether it had previously been defined to be a specific function. Even more subtle is the specification of variable names which are treated essentially like constants although they remain variables; this is discussed in II.1.7.7..

Logical Symbols

II.1.1.5. There are only a few logical symbols we will use. These fall into the categories of *relations*, *set operations*, *connectives*, and *quantifiers*, and one additional symbol for negation.

Relations:

$=$ ("equals")

$\in$ ("is a member of" or "belongs to")

$\subseteq$ ("is a subset of")

Set Operations:

$\cup$ ("union")

$\cap$ ("intersection")

$\setminus$ ("relative complement")

Connectives:

$\wedge$ ("and")

$\vee$ ("or")

$\Rightarrow$ ("implies")

Quantifiers:

$\exists$ ("there exists")

$\forall$ ("for all")

Negation:

$\neg$ ("not")

The symbols $\wedge$, $\Rightarrow$, and $\neg$ are sometimes written &, $\rightarrow$, and $\sim$ respectively. The set operation $\setminus$ is also sometimes written $\sim$ or $-$.

II.1.1.6. The set-theoretic operations are not strictly speaking logical symbols, but constant operations (functions) defined in set theory. However, they are so universal that we can regard them as logical symbols. Other set-theoretic operations such as symmetric difference can also be defined and given symbols. The relations $\in$ and $\subseteq$ are also technically constant relations defined in set theory.

II.1.1.7. The symbols we have given here are somewhat redundant; not all are needed, since some can be defined in terms of others, as described below. But it is useful to have all of them since equivalent statements written without them can be considerably more complicated. A few other useful logical symbols such as $\Longleftrightarrow$ and ! will be defined later in terms of the standard ones.

Parentheses

II.1.1.8. Parentheses are technically not symbols of the formal language, and are used merely to make the formal statements more comprehensible. They can be avoided entirely if certain conventions of interpretation are strictly adhered to. But they are very desirable to have in practice, and we will make use of them freely. For clarity, nested parentheses can be written as (), [], or rarely even {}, which all have the same meaning (or lack thereof).

Terms

II.1.1.9. DEFINITION. A *simple term* of a certain type is a constant or variable of the specified type.
A *compound term* of level 2 of a certain type is a function whose value is of the specified type, whose arguments are simple terms of the appropriate types.
A *compound term* of level $n \geq 2$ of a certain type is a function whose value is of the specified type, whose arguments are simple terms or compound terms of level $< n$ of the appropriate type.
A simple term or a compound term of any (finite) level is called a *term*.

The level of a compound term is of little if any significance; even the distinction between simple and compound terms is unimportant. The important thing is that terms are objects of a specified type built up from constants and variables through a finite number of steps.

II.1.1.10. EXAMPLES. Examples of simple terms of type "real number" are 0, $\sqrt{2}$, π, x where x is a variable of type "real number." If f is a variable which is of type "real-valued function of three real arguments", and we recall that $+$ is a constant of type "function from pairs of real numbers to real numbers", then such expressions as $f(x, y, 5)$ and $x+z$ (where x, y, z are variables of type "real number") are compound terms of level 2 of type "real number", and $f(x + y, x + 2, f(x, x, y))$ is a compound term of level 3 of type "real number."

Atomic Statements

II.1.1.11. DEFINITION. An *atomic statement*, or *simple statement*, is an expression of the form $R(x_1, \ldots, x_n)$, where R is an n-ary constant or logical relation ($n \geq 2$) and $x_1, \ldots, x_n$ are terms (simple or compound) of the appropriate types.

The words *sentence* and *proposition* are often used interchangeably with *statement*.

II.1.1.12. EXAMPLES. Expressions of the form $x = y$, $(x + 5) \in A$, $f(x, y, \sin(z)) > 0$, etc., are atomic statements, where x, y, z are variables of type "real number" and A is a constant or variable of type "set of real numbers."

General Formal Statements

II.1.1.13. DEFINITION. A *formal statement* is an expression built up from atomic statements by a finite number of connectives, quantifications, and negations. Specifically:

Every atomic statement is a formal statement.

If P and Q are formal statements, then the following are also formal statements:

$P \wedge Q$.

$P \vee Q$.

$P \Rightarrow Q$.

$\neg P$.

$(\forall x)(P)$, where x is a variable not already quantified in P.

$(\exists x)(P)$, where x is a variable not already quantified in P.

(Quantification is discussed in detail in the next subsection.)

The *level* of a formal statement can be defined as in II.1.1.9.: atomic statements have level 1, and every connective, quantification, or negation adds one to the level. Every formal statement has level n for some finite n.

First-Order vs. Higher-Order Statements

II.1.1.14. There is an important point, which, however, need not concern us very much at this level. In most contexts, there is a natural set of objects being discussed, usually called the *domain of discourse*. In most of this book, the domain of discourse would be the real numbers.

II.1.1.15. While constants can be not only objects in the domain of discourse, but also subsets of this domain, functions and relations on this domain, etc., logicians often (in fact, usually) restrict variables to represent only actual elements of the domain of discourse. Statements containing only variables taking values in the domain of discourse are called *first-order statements*, and statements with variables ranging over larger sets are *higher-order statements*. Levels of higher-order statements can be defined similarly to II.1.1.9.; for example, a statement containing one or more variables ranging over subsets of the domain of discourse or functions from the domain of discourse to itself is a *second-order statement*. Thus, for example, the statements of II.1.1.1. and II.1.3.11. are first-order statements, but the statement in II.1.3.12. is second-order.

II.1.1.16. The distinction between first-order and higher-order statements can be eliminated by expanding the domain of discourse. However, in practice the domain of discourse is usually established by a set of axioms assumed in the discussion, so cannot be easily expanded. (Also, even expanding the domain of discourse does not convert second-order theories to first-order theories.)

II.1.1.17. A *first-order theory* is a theory in which all statements are first-order. "First-order logic" can be used in first-order theories. "Higher-order logic" is needed for higher-order theories; higher-order logic is not as well behaved, or as well understood, as first-order logic, and indeed "logic" means "first-order logic" to many logicians.

One common criticism of second-order logic, made particularly in [**?**], is that second-order logic is essentially first-order set theory, and hence is "mathematical" and not "logic"; in [**?**] second-order logic is called "set theory in sheep's clothing." See [**?**] for an attempt to refute this criticism.

While restriction to first-order logic may make sense to logicians, it does not work well for mathematics in general; second-order logic is really needed to properly describe real analysis and even ordinary arithmetic the way mathematicians do it. See e.g. [Luc00] or [Sha05, Chapter 25] for details. As J. Barwise [**?**, p. 5] puts it:

> As logicians, we do our subject a disservice by convincing others that logic is first-order and then convincing them that almost none of the concepts of modern mathematics can really be captured in first-order logic."

II.1.1.18. We will feel free to use higher-order statements (i.e. quantify as we see fit) in our development of analysis, and just assume enough higher-order logic exists to justify this policy (this assumption is in fact justifiable despite the incompleteness of higher-order logic).

II.1.2. Meaning of Statements

II.1.2.1. The "meaning" of a formal statement is not something that can really be defined. We will limit our discussion to the interpretation of truth values for formal statements.

II.1.2.2. The truth value of an atomic formal statement cannot be determined (in most cases) unless values for the variables are specified. Once they are, the statement is *true* if and only if the resulting n-tuple is one which satisfies the defining relation (technically, the relation is a set of n-tuples, and the relation is "satisfied" by a particular n-tuple if it is in the relation). A statement is *false* if it is not true.

II.1.2.3. Truth values for compound statements depend on the truth values of the constituent parts in a precise way which we will now describe. The first three are simple and fairly obvious:

$(P \wedge Q)$ is true if and only if both P and Q are true.

$(P \vee Q)$ is true if and only if P is true or Q is true or both. (Thus $\vee$ is sometimes called "inclusive or", as opposed to the "exclusive or" used primarily in computer science.)

$\neg P$ is true if and only if P is false.

Implication

II.1.2.4. The meaning of $\Rightarrow$ is less obvious. $(P \Rightarrow Q)$ is read "P implies Q," "If P then Q," or "P only if Q." Mathematicians have found through experience that the proper meaning of $(P \Rightarrow Q)$ is

$$[(\neg P) \vee Q] \ .$$

The idea is that if P is true, then Q must necessarily be true, but if P is false, no conclusion about the truth of Q may be drawn. So the implication is only violated if P is true but Q is false; the implication is satisfied (not violated) in all other cases. If P is false, the implication is not violated irrespective of whether Q is true or false, and if Q is true, the implication is not violated irrespective of whether P is true or false.

This interpretation leads to some jarring conclusions in cases where there appears to be no logical relationship between P and Q. Thus mathematicians would regard the implication

"If $1 = 2$, then the sky is green."

as being "true." But there is no other reasonable interpretation in terms of truth values for $(P \Rightarrow Q)$ than $[(\neg P) \vee Q]$.

II.1.2.5. The connective $\Rightarrow$ can thus be defined in terms of $\neg$ and $\vee$, and so is strictly speaking redundant. But we will use it freely since it is natural and many statements are simpler and more comprehensible if it is used.

Converse, Inverse, and Contrapositive

II.1.2.6. DEFINITION. Let P and Q be statements.

(i) The *converse* of the statement $P \Rightarrow Q$ is $Q \Rightarrow P$.

(ii) The *inverse* of the statement $P \Rightarrow Q$ is $(\neg P) \Rightarrow (\neg Q)$.

(iii) The *contrapositive* of the statement $P \Rightarrow Q$ is $(\neg Q) \Rightarrow (\neg P)$.

II.1.2.7. It is important and basic that the converse of a statement does not have the same meaning; there is no implication either way between a statement and its converse, as almost any example shows. It is a common mistake to confuse a statement and its converse, which requires practice to learn to avoid.

However:

II.1.2.8. PROPOSITION. A statement and its contrapositive are logically equivalent (if ordinary logic is used).

PROOF: The statement $P \Rightarrow Q$ means $(\neg P) \vee Q$, and its contrapositive $(\neg Q) \Rightarrow (\neg P)$ means $[\neg(\neg Q)] \vee (\neg P)$, and $\neg(\neg Q)$ is equivalent to Q in ordinary logic. 👍

II.1.2.9. Note that similarly, the converse and inverse of a statement are logically equivalent; indeed, the inverse is the contrapositive of the converse.

II.1.2.10. So to prove a statement, it is just as good to prove the contrapositive, and this is frequently easier or more natural. In fact, we will often just prove the contrapositive of a statement without comment.

If and Only If

II.1.2.11. We can define one other convenient logical connective, which can be regarded as merely an abbreviation if desired. If P and Q are statements, we write $P \Longleftrightarrow Q$, read "P if and only if Q", to mean

$$[(P \Rightarrow Q) \wedge (Q \Rightarrow P)] \ .$$

Rewriting this without the symbol $\Rightarrow$, we obtain

$$([(\neg P) \vee Q] \wedge [P \vee (\neg Q)]) \ .$$

The statement $P \Longleftrightarrow Q$ is also logically equivalent to

$$([P \wedge Q] \vee [(\neg P) \wedge (\neg Q)]) \ .$$

II.1.2.12. The following table summarizes the behavior of truth values under connectives and negation:

P	Q	$\neg P$	$(P \wedge Q)$	$(P \vee Q)$	$(P \Rightarrow Q)$	$(P \Longleftrightarrow Q)$
T	T	F	T	T	T	T
T	F	F	F	T	F	F
F	T	T	F	T	T	F
F	F	T	F	F	T	T

II.1.3. Quantification and Closed Statements

II.1.3.1. The quantifiers $\exists$ and $\forall$ are called the *existential quantifier* and the *universal quantifier* respectively. They are used to modify a statement P by specifying the sense in which a certain variable is used in P. A variable appearing in P which has not been quantified is called *free*; free variables give a degree of incompleteness in a statement. For example, if we just state "$x > 3$", we cannot pass judgment on the truth of the statement without knowing what the x is supposed to stand for. But if we write $(\forall x)(x > 3)$ ("For

all x, $x > 3$") or $(\exists x)(x > 3)$ ("There exists an x such that $x > 3$"), we have a self-contained statement and we know what we are talking about, and it makes sense to say the statements are false and true respectively (in the real numbers).

Note, however, that if the statement $x > 3$ appears in the course of a discussion, and x has been previously defined in the discussion to be a specific number (e.g. suppose there has been a previous line "Let $x = \sqrt{10}$."), then the x in the statement $x > 3$ is a constant, not a variable, and a truth value can be unambiguously assigned (true in this case).

Quantified variables can be regarded as "dummy variables" whose names are really just placeholders in an expression. The x in $(\forall x)(x > 3)$ plays the same role as the x in $f(x) = x^2 + 1$ or $\int_0^1 x^2\,dx$, or the n in $\sum_{n=1}^{4}(2n - 1)$. The symbol used is unimportant; it can be replaced by any other variable name which does not already appear in the expression without changing the meaning.

II.1.3.2. A variable in a statement P which has been quantified is called a *bound variable*. A bound variable cannot be quantified again.

Logicians use the term "bound" to refer only to a variable quantified by an existential quantifier. In fact, universal quantifiers are normally not used at all in mathematical logic; they are often just understood, or if needed can be defined in terms of existential quantifiers and negations (). But universal quantifiers are very useful in practice in mathematics, and it is good form to use them explicitly rather than to let them just be understood or implicit; statements with explicit universal quantifiers (where needed) are generally much clearer and more comprehensible.

Restricted Quantification

II.1.3.3. When quantifying variables, we often (even usually) want to place explicit restrictions on the values the variables are allowed to take. Thus we often want to write expressions like $(\forall n \in \mathbb{N})$ or $(\exists \delta > 0)$. With proper interpretation, such restricted quantification is allowable and causes no problems; a restricted quantification of a statement will just be an ordinary quantification of a related but more complicated statement.

II.1.3.4. In the usual case where we have a specified domain of discourse, e.g. the real numbers $\mathbb{R}$, the expressions $(\forall x \in \mathbb{R})$ or $(\exists x \in \mathbb{R})$ mean exactly the same thing as the simple expressions $(\forall x)$ and $(\exists x)$ respectively, and writing $(\forall x \in \mathbb{R})$ or $(\exists x \in \mathbb{R})$ simply adds to the clarity of the statement by specifying what type of variable x is.

II.1.3.5. In the case where A is a constant (previously specified) subset of the domain of discourse, we take $(\forall x \in A)(P)$ to mean

$$(\forall x)[(x \in A) \Rightarrow P]$$

and we take $(\exists x \in A)(P)$ to mean

$$(\exists x)[(x \in A) \wedge P]\ .$$

II.1.3.6. Similarly, if R is a relation on the domain of discourse, and we write $R(x)$ to denote R with x used as one of the arguments, along with other terms, then by $(\forall x R(x))(P)$ we mean

$$(\forall x)[R(x) \Rightarrow P]$$

and by $(\exists x R(x))(P)$ we mean

$$(\exists x)[R(x) \wedge P] \ .$$

For example, if the domain of discourse is $\mathbb{R}$, the expression $(\exists \delta > 0)(P)$ means

$$(\exists \delta)[(\delta > 0) \wedge P] \ .$$

Order of Quantifiers

II.1.3.7. In a statement with more than one quantifier, the order of quantifiers is crucial; interchanging the order of a universal quantifier and an existential quantifier can drastically alter the meaning of a statement. For example, consider the following statement:

"For every natural number n there is a real number x with $x > n$."

The corresponding formal statement is

$$(\forall n \in \mathbb{N})(\exists x \in \mathbb{R})(x > n) \ .$$

This statement is true (e.g. can take $x = n + 1$).

But suppose the order of the quantifiers is reversed, i.e. we take the formal statement

$$(\exists x \in \mathbb{R})(\forall n \in \mathbb{N})(x > n)$$

which in English says

"There is a real number x such that for every natural number n, $x > n$."

This statement is false (). The point is that in the first statement, only the existence of an x satisfying $x > n$ for one particular, though unspecified, n is asserted (the x can depend on the n), whereas in the second it is asserted that there exists one x which satisfies the inequality simultaneously for all n.

II.1.3.8. There is an implication one way, however: for any statement P, and for any free variables x and y in P, we have

$$[(\exists x)(\forall y)(P)] \ \Rightarrow \ [(\forall y)(\exists x)(P)] \ .$$

II.1.3.9. The order of two consecutive universal quantifiers, or two consecutive existential quantifiers, can be reversed without changing the meaning of the statement. If two consecutive quantifiers of the same type have identical restrictions, they can be combined into one abbreviated quantification expression. For example, the expressions $(\forall x \in \mathbb{R})(\forall y \in \mathbb{R})$ and $(\forall y \in \mathbb{R})(\forall x \in \mathbb{R})$ have an identical meaning, and can be abbreviated as $(\forall x, y \in \mathbb{R})$. We can similarly write $(\exists x, y \in \mathbb{R})$ to mean $(\exists x \in \mathbb{R})(\exists y \in \mathbb{R})$ or $(\exists y \in \mathbb{R})(\exists x \in \mathbb{R})$.

Closed Statements

II.1.3.10. Definition. A *closed statement* is a formal statement P in which every variable is quantified (i.e. every variable in P is bound, or P contains no free variables).

Closed statements are the ones which are "self-contained" and to which a well-defined truth value can (potentially) be attached. Restricted quantification is allowable in closed statements; such a statement can be rewritten as a more complicated closed statement with only ordinary quantification.

II.1.3.11. In principle, the statement of any theorem is a closed statement, and it is good mathematical form to write the statement of every result in the form of, or in a form easily translatable into, a closed statement. In practice it may sometimes be overly pedantic to do this, and we sometimes abbreviate when no confusion is likely. We also sometimes for stylistic reasons write theorem statements in a form not exactly of a closed statement. For example, many theorems are phrased in the form of the next statement, or slight variations of it:

THEOREM. Let x and y be real numbers. If $x < y$, then there is a rational number r with $x < r < y$.

The corresponding closed formal statement is

$$(\forall x \in \mathbb{R})(\forall y \in \mathbb{R})[(x < y) \Rightarrow ((\exists r \in \mathbb{Q})[(x < r) \wedge (r < y)])]$$

which is a little awkward to express directly in English.

II.1.3.12. Sometimes the formal statement is even more different from the English one. For example:

THEOREM. Every nonempty subset of $\mathbb{N}$ has a smallest element.

A corresponding formal statement would be:

$$(\forall S \subseteq \mathbb{N})[(\exists n)(n \in S) \Rightarrow (\exists m)[(m \in S) \wedge (\forall n \in S)(m \leq n)]] \ .$$

This statement explicitly incorporates the meaning of "nonempty" and "smallest element." It also illustrates that variable objects which can remain unnamed in the English statement must often be given names in the formal statement.

II.1.3.13. One mistake often made by beginners (and sometimes by non-beginners too!) is to write a statement where the same letter is used to mean two or more different things in the expression, i.e. to stand for two or more different variables or constants. The statement in II.1.3.12. does not have this mistake since the two uses of n occur within disjoint quantifications (convince yourself of this!), but it would have been safer, and perhaps clearer, to use another letter in place of one of the uses of n. It is always safest to use a separate letter or symbol for every different object in a statement. See II.1.7.7. for further discussion of a related point.

II.1.3.14. There is a related (mostly nonmathematical) issue concerning style. In a mathematical statement or argument, it is a good idea to use easily distinguishable symbols for unrelated objects; using the same letter in upper and lower case, or in different typefaces, can be confusing to a reader. It is often good practice to use related symbols for related objects: for example, it is usually good practice to use a lower-case letter to denote an element of a set whose name is the same upper-case letter (e.g. $x \in X$, $n \in \mathbb{N}$). Some extension of this is often also good practice, e.g.

"Let $x_n \in X$ for all $n \geq N$."

(Here N is a previously specified number.) But this can be taken to ridiculous lengths. For example, I have seen the following statement in a book (which shall remain anonymous):

"Let $A_a \in \mathfrak{A}_a$ for all $a \in \mathbb{A}$."

Here $\mathbb{A}$ was an index set, a an index, and $\mathfrak{A}_a$ a set for each a. Although there is nothing mathematically wrong with such a statement, *DON'T DO THIS!* (I admit we have been forced by standard conventions to do something like this in Section XVII.6.3., but it is still to be avoided whenever possible.)

II.1.3.15. There is a wide choice of possible symbols available for use in mathematical statements. In principle any sort of symbol can be used, although in practice one should use symbols which can be easily recognized, distinguished, and remembered (and pronounced!) Besides standard mathematical symbols, one should generally stick to Latin or Greek letters, although other symbols are sometimes used. German *fraktur* is common (more common in older references). Hebrew letters, notably $\aleph$, are standard in certain contexts. I admit to occasionally using Cyrillic letters. And in a paper with a Danish coauthor, we used the Danish letter æ in a context where it was appropriate, although we refrained from also using the Danish letter ø, as in

"Let $\emptyset$ be a set, and $\text{ø} \in \emptyset$."

I recommend exercising restraint in using uncommon letters or symbols.

II.1.3.16. A closed statement need not have any variables at all. Ones which do not tend to be simple formulas like elementary statements about arithmetic (e.g. $2+2=4$), but they are not necessarily so simple. One example is the following result, which caused a minor stir when it was discovered by F. Cole in 1903 since it provided a counterexample to an old conjecture of Mersenne in number theory (cf. [Bel87, p. 228]):
Theorem. $2^{67}-1 = 193707721 \times 761838257287$.

II.1.4. Negations

II.1.4.1. It is particularly important to be able to write the negation of a mathematical statement. In one sense, it is a triviality to do this: just insert the word "not" (and perhaps adjust the statement slightly to make it grammatically acceptable), i.e. the negation of the statement P is $\neg P$. But this begs the question: in order to have a *useful* phrasing of the negation of a statement (e.g. in order to disprove the original statement, or to prove it by contradiction), we need a way of writing the negation affirmatively without the words "no" or "not". (In fact, it is most useful to write every mathematical statement affirmatively without such words when possible, and correspondingly to write formal statements without using the symbol $\neg$, and it turns out it is almost always possible to do this.)

It is often quite difficult to do this for statements written in English (try it for the two statements in the introduction of this section!) But there is a systematic way to affirmatively write the negation of a formal statement (although it involves what could be regarded as a bit of a subterfuge for atomic statements). So a procedure for finding the negation of a mathematical statement expressed in English would be:

(1) Rewrite the statement as a formal statement.

(2) Find the negation of the formal statement.

(3) If desired, convert the negated formal statement into an English statement.

II.1.4.2. We should emphasize that this discussion of negations uses ordinary (or "classical") logic. Negations behave rather differently if intuitionistic logic is used (cf. ()).

Negations of Atomic Statements

II.1.4.3. It is quite simple to affirmatively write the negation of an atomic statement: just replace the relation in the statement with its complementary relation. (A relation is a subset of a Cartesian product, and the complementary relation is just the set-theoretic complement in the Cartesian product.) Sometimes there is a simple formula for the complementary relation: for example, if the domain of discourse is $\mathbb{R}$, the negation of the statement $(x > 3)$ is $(x \leq 3)$. But it is not always so simple: for example, in a general domain of discourse there is no simple way to write the negation of the sentence $(x = y)$ except as $(x \neq y)$. (If the domain of discourse is $\mathbb{R}$, the negation can be written $[(x < y) \vee (x > y)]$, but this is rather special to $\mathbb{R}$.)

Negations of Compound Statements

II.1.4.4. The negation of $\neg P$ is P, i.e. P is logically equivalent to $\neg(\neg P)$ for any statement P.

II.1.4.5. The negation of $P \vee Q$ is

$$[(\neg P) \wedge (\neg Q)] \ .$$

To see this, note that $P \vee Q$ is true if and only if P is true or Q is true or both. Thus $P \vee Q$ is false if and only if both P and Q are false.

An almost identical argument (or an application of II.1.4.4.) shows that the negation of $P \wedge Q$ is

$$[(\neg P) \vee (\neg Q)] \ .$$

So if P and Q are sentences whose negations can be written without use of $\neg$ (e.g. if P and Q are atomic sentences), then the negations of $P \vee Q$ and $P \wedge Q$ can be written also without use of $\neg$.

II.1.4.6. Example. The negation of $[(x > 3) \vee (x \leq 1)]$ is $[(x \leq 3) \wedge (x > 1)]$, which can be abbreviated $(1 < x \leq 3)$.

II.1.4.7. Since $(P \Rightarrow Q)$ is logically equivalent to $[(\neg P) \vee Q]$, the negation of $(P \Rightarrow Q)$ is

$$[P \wedge (\neg Q)] \ .$$

So if P and Q are sentences whose negations can be written without use of $\neg$ (e.g. if P and Q are atomic sentences), then the negation of $P \Rightarrow Q$ can be written also without use of $\neg$.

II.1.4.8. Since we have $\neg(P \vee Q)$ logically equivalent to $[(\neg P) \wedge (\neg Q)]$, we have that $(P \vee Q)$ is logically equivalent to $\neg[(\neg P) \wedge (\neg Q)]$. Thus $(P \vee Q)$ can be *defined* to mean $\neg[(\neg P) \wedge (\neg Q)]$, i.e. $\vee$ can be defined in terms of $\neg$ and $\wedge$. Similarly, we have that $(P \wedge Q)$ is logically equivalent to $\neg[(\neg P) \vee (\neg Q)]$, i.e. $\wedge$ can be defined in terms of $\neg$ and $\vee$. We will not want to do this, for two reasons: (1) both $\vee$ and $\wedge$ are natural and useful in writing comprehensible formal statements, and (2) we actually want to avoid using $\neg$ as much as possible.

II.1.4.9. The negation of $(P \iff Q)$ can be found by writing it as $[P \vee (\neg Q)] \wedge [Q \vee (\neg P)]$, whose negation is $[(\neg P) \wedge Q] \vee [(\neg Q) \wedge P]$. $(P \iff Q)$ can also be rewritten as $[P \wedge Q] \vee [(\neg P) \wedge (\neg Q)]$, whose negation is $[P \vee Q] \wedge [(\neg P) \vee (\neg Q)]$.

Negations of Quantified Statements

II.1.4.10. The negation of $(\exists x)(P)$ is

$$(\forall x)(\neg P) \ .$$

To see this, note that $(\exists x)(P)$ is true if and only if there is at least one x making P true. The negation is that every x makes P false, i.e. every x makes $\neg P$ true.

Similarly (or by an application of II.1.4.4.), the negation of $(\forall x)(P)$ is

$$(\exists x)(\neg P) \ .$$

Thus to take the negation of a statement with a quantifier, just take the negation of the inside statement and change the quantifier from existential to universal or vice versa.

II.1.4.11. If a statement has two or more consecutive quantifiers, change each but leave the order of the quantifiers unchanged. Thus the negation of the statement $(\forall x)(\exists y)(P)$ is

$$(\exists x)(\forall y)(\neg P) \ .$$

Negations of Statements with Restricted Quantifiers

It can be confusing to try to reason out the negation of a statement with restricted quantifiers. It is easy to make the mistake of changing too much – don't fail to avoid making this mistake!

II.1.4.12. Proposition. The negation of the statement $(\forall x \in A)(P)$ is

$$(\exists x \in A)(\neg P) \ .$$

The negation of the statement $(\exists x \in A)(P)$ is

$$(\forall x \in A)(\neg P) \ .$$

Proof: These can be reasoned out just as ordinary quantifiers, but we can also argue formally on the basis of the meaning of the restricted quantifiers. Since $(\forall x \in A)(P)$ means

$$(\forall x)[(x \in A) \wedge P]$$

we have that the negation is

$$(\exists x)[[\neg(x \in A)] \vee (\neg P)]$$

which by definition is the same as

$$(\exists x)[(x \in A) \Rightarrow (\neg P)]$$

and this is the meaning of $(\exists x \in A)(\neg P)$. The second half is the same argument run backwards, with P and $\neg P$ interchanged. 👍

By an almost identical argument, we obtain:

II.1.4.13. PROPOSITION. If R is a relation, then the negation of $(\forall x R(x))(P)$ is

$$(\exists x R(x))(\neg P)$$

and the negation of $(\exists x R(x))(P)$ is

$$(\forall x R(x))(\neg P)\ .$$

II.1.4.14. Thus, in taking the negation of a statement with restricted quantifiers, the inside statement is negated, universal and existential quantifiers are switched, but the restrictions on each quantifier stay the *same*. If there is more than one quantifier, the order of the quantifiers stays the same.

II.1.4.15. EXAMPLE. Consider the statement

$$(\forall a \in \mathbb{R})(\forall \epsilon > 0)(\exists \delta > 0)(\forall x \in \mathbb{R})[(|x - a| < \delta) \Rightarrow (|f(x) - f(a)| < \epsilon)]$$

of II.1.1.1.. Its negation is

$$(\exists a \in \mathbb{R})(\exists \epsilon > 0)(\forall \delta > 0)(\exists x \in \mathbb{R})[(|x - a| < \delta) \wedge (|f(x) - f(a)| \geq \epsilon)]$$

(using an application of II.1.4.7. on the statement inside the quantifiers).

II.1.4.16. By repeated applications of the techniques of this section, the negation of any formal statement P can be written. If P can be written without use of $\neg$, so can its negation.

II.1.5. Uniqueness Statements

II.1.6. Tautologies

II.1.6.1. A *tautology* is a statement form which is always true no matter what truth values are assigned to the constituents. Some standard tautologies (for ordinary logic) are:

$\neg[P \wedge (\neg P)]$ *(Consistency)*

$P \vee (\neg P)$ *(Law of the Excluded Middle)*

$[P \Rightarrow Q] \iff [(\neg P) \vee Q]$ *(Definition of $\Rightarrow$)*

Some tautologies discussed in the section on negation are:

$P \iff \neg(\neg P)$ *(Law of Double Negation)*

$[\neg(P \vee Q)] \iff [(\neg P) \wedge (\neg Q)]$

$[\neg(P \wedge Q)] \iff [(\neg P) \vee (\neg Q)]$

$[\neg(P \Rightarrow Q)] \iff [P \wedge (\neg Q)]$

$[\neg[(\forall x)(P)]] \iff [(\exists x)(\neg P)]$

$[\neg[(\exists x)(P)]] \iff [(\forall x)(\neg P)]$

Some other tautologies which can be verified by using truth tables:

$[(P \vee Q) \vee S] \iff [P \vee (Q \vee S)]$ *(Associativity of $\vee$)*

$[(P \wedge Q) \wedge S] \iff [P \wedge (Q \wedge S)]$ *Associativity of $\wedge$)*

$(P \vee Q) \iff (Q \vee P)$ *(Commutativity of $\vee$)*

$(P \wedge Q) \iff (Q \wedge P)$ *(Commutativity of $\wedge$)*

$[(P \vee Q) \wedge S] \iff [(P \wedge S) \vee (Q \wedge S)]$ *(Distributivity of $\wedge$)*

$[(P \wedge Q) \vee S] \iff [(P \vee S) \wedge (Q \vee S)]$ *(Distributivity of $\vee$)*

$[(P \Rightarrow Q) \wedge (Q \Rightarrow S)] \Rightarrow [P \Rightarrow S]$ *(Transitivity of $\Rightarrow$ or Syllogism Law)*

$[P \Rightarrow Q] \iff [(\neg Q) \Rightarrow (\neg P)]$ *(Law of Contrapositive)*

$[P \Rightarrow (Q \Rightarrow S)] \iff [(P \wedge Q) \Rightarrow S]$

$P \iff (P \wedge [Q \vee (\neg Q)])$

$P \iff (P \vee [Q \wedge (\neg Q)])$

$P \iff ([P \wedge Q] \vee [P \wedge (\neg Q)])$

$P \iff ([P \vee Q] \wedge [P \vee (\neg Q)])$

$P \iff ([Q \Rightarrow P] \wedge [(\neg Q) \Rightarrow P])$ *(Case Testing)*

$[(\exists x)(\forall y)(P)] \Rightarrow [(\forall y)(\exists x)(P)]$

$P \Rightarrow P$ (note that this is by definition equivalent to the Law of the Excluded Middle, and thus is not a tautology in intuitionist logic!)

$P \Rightarrow (P \vee Q)$

$(P \wedge Q) \Rightarrow P$

$(\neg P) \Rightarrow (P \Rightarrow Q)$

$Q \Rightarrow (P \Rightarrow Q)$

$[P \wedge (P \Rightarrow Q)] \Rightarrow Q$ *(Modus Ponens)*

II.1.6.2. A tautology of the form $A \iff B$, where A and B are statement forms involving P's, Q's, etc., is called a *tautological equivalence*. A tautological equivalence is a *logical equivalence*, i.e. any occurrence of a statement of form A can be replaced by the corresponding statement of form B, or vice versa, without changing the logical meaning of the statement. (A logical equivalence need not be a tautological equivalence, however.) Such substitutions allow for simplification or modification of statements in arguments.

II.1.6.3. A tautology of form $A \Rightarrow B$ can be used similarly, allowing replacement of a statement of form A by the corresponding one of form B (but not conversely). The converted statement is not (necessarily) logically equivalent to the original, but is a logical consequence. This type of substitution can also be used in arguments; however, such an argument is then not reversible.

Use of *modus ponens* is a particularly important example: if one shows that P is true and that $(P \Rightarrow Q)$ is true, it can be concluded that Q is true. This logical argument is used constantly. Actually, *modus ponens* is more accurately a *law of inference*, not just a tautology.

II.1.6.4. If intuitionist logic is used, not all of these tautologies hold. For example, intuitionists do not accept the Law of the Excluded Middle (and hence not $P \Rightarrow P$ either!) They (mostly) accept $P \Rightarrow [\neg(\neg P)]$, but not the converse. Several of the other tautologies listed above are not valid in intuitionist logic; we leave it to the reader to identify which ones.

II.1.7. Definitions

II.1.7.1. Definitions in mathematics are of two rather different sorts:

(1) A precise specification of the meaning of a word or phrase.

(2) A specification that a certain symbol (or compound symbol) is to stand for a particular mathematical object (i.e. the specification of a constant in the sense of II.1.1.3.).

II.1.7.2. Examples. (i) Examples of definitions of the first kind are the definitions of *partial order* (), *field* (), *metric space* (), *measure* (), *Riemann integral* (), $\cdots$
(ii) Examples of definitions of the second type are expressions of the form

(a) "Let $f(x) = x^2 + 1$."

(b) "Let $x = \lim_{k \to \infty} x_k$."

(c) "Let B be the set of all upper bounds of the set A [where A has been previously defined]."

II.1.7.3. A definition of the first kind is simply the establishment of an "abbreviation" for a (generally more complicated) phrase. In principle, the defined term can always be replaced by the phrase it abbreviates without loss of utility, and thus from a strictly logical point of view such definitions are unnecessary. However, in practice it is very useful, often essential, to be able to make these definitions: otherwise, statements involving even moderately complex mathematical ideas would become unmanageably or even incomprehensibly complicated.

Definitions of the first kind are often of words which are standard terms used widely by mathematicians, and are frequently even the basic subject matter of whole parts of mathematics. But definitions of this kind can also be conveniently used to create a name for a concept of no obvious interest beyond the specific discussion at hand but which must be referred to multiple times, such as a concept which will shortly be shown to be equivalent to some other well-known concept.

II.1.7.4. Definitions of the second kind are another story. In principle, they too can be avoided, but in practice it is often hard to see how arguments can be reasonably phrased without their use.

II.1.7.5. Definitions of the second kind are normally made as part of a mathematical argument, and are operative only for the duration of the argument. Definitions of the first kind can be made in the same way as part of an argument, but are more commonly stand-alone statements.

Note that a definition is not a formal mathematical statement as described in this section, although it may resemble a formal statement in form.

II.1.7.6. Sometimes a definition of the second kind is dependent on other results for it to make sense. For example, in II.1.7.2.(iib), it must be previously (or subsequently) shown that $\lim_{k\to\infty} x_k$ exists, and that limits are unique.

A common instance where a definition must be justified is a situation where the object defined apparently or potentially depends on some choices. It must be shown that the definition always gives the same result no matter how the choices are made. In this situation, we say the resulting object is *well defined*. For example, we might make a definition (cf. ())

> "Let $f(x) = \lim_{k\to\infty} f(x_k)$, where (x_k) is any sequence in E converging to x."

To show that $f(x)$ is well defined, it must be shown not only that $\lim_{k\to\infty} f(x_k)$ always exists for every such choice of (x_k), but also that the same limit is always obtained no matter how the x_k are chosen (so long as they are in E and converge to x).

Although from a logical standpoint justifications should be put before the definition, the justification for the legitimacy of a definition is stylistically often put after the definition itself.

Ambiguity of Terminology

Ideally, every term used in mathematics should have a unique, unambiguous precise definition. But in practice there are many instances where the same name is used in mathematics to refer to quite different concepts in different contexts (this is probably inevitable since there is only a limited supply of suitable words in English to describe mathematical concepts, and is not even entirely undesirable since there is often some similarity between distinct concepts). Some common examples are *closed*, *complete*, *free*, *independent*, *regular*, *normal*, and *separable*. In different instances there may be a greater or lesser philosophical resemblance between the usages; but there is still a tendency for the mathematically uninitiated to make too much of the coincidence of terminology by taking the names too literally.

In most if not all languages where mathematical terms have been named, there are similar instances where the same name is used in different contexts. But the ambiguities are not always the same. For example, French mathematicians use the word *fermé* to refer to closed sets in topology, but *algebriquement clos* for algebraically closed fields. On the other hand, the French word *nœud* is used to mean both *node* and *knot* (and I have seen it translated incorrectly in English translations of French mathematics books!) Thus a French-speaking neophyte in mathematics may confuse different mathematics concepts than an English-speaking one.

Separated Quantification

II.1.7.7. There is a variant of type two definitions which is frequently employed, which is technically not a true definition at all. In mathematical arguments, we often write phrases such as

"Let x be an element of A."

(Here, A is a previously defined set; in order to legitimately make such a statement, it must be known, or already shown, that A is nonempty.) This is not really a definition, since it does not specify exactly what x is. The x remains a variable, but from that point on it is treated essentially like a constant in that it must always refer to the same (variable) object (this is related to the discussion in II.1.3.13.). If we then show that a statement $P(x)$ containing the variable x is true, without further specifying what x is or making assumptions about it beyond that it is an element of A, what we have shown is that the statement $(\forall x \in A)[P(x)]$ is true.

II.1.7.8. What is technically being done here is that quantification of a variable is stylistically moved to the beginning of a section of a discussion in which it is used repeatedly, which we call *separated quantification*. Separated quantification can be eliminated, but the price of elimination in many cases is a horribly complex compound quantified statement which is difficult to understand. So separated quantification is very important to use to make mathematical discussions readable and understandable.

II.1.7.9. Sometimes "definitions" of this sort are made implicitly. For example, if (x_k) is known to be a bounded sequence, one can legitimately say:

"Since (x_k) is bounded, there is an M such that $x_k \leq M$ for all k."

This implicitly defines M to be a (variable) number satisfying the inequality for all k.

It is good mathematical form to phrase things in this way. A bad form mistake beginners often make is to write something like

"Since (x_k) is bounded, $x_k \leq M$ for all k."

where the M is a variable which has not yet been specified, i.e. the variable name M has not yet appeared in the discussion. (It is a more serious mistake if the name M has already been used to mean something else.) A reader seeing a statement like this must assume that the M refers to something which has already been defined, and wonders "What is M?" Some time and effort then must be wasted trying to find or figure out what the M refers to, and if it is concluded that the statement is intended to be an implicit "definition" of M, the conclusion is only an educated guess, so there is a lack of precision and rigor.

II.1.7.10. There are subtle and potentially confusing instances where it is unclear whether an implicit separated definition is being made, for example in the common calculus statement

$$\int x\,dx = \frac{x^2}{2} + C\ .$$

Is the C now a quantified variable? We often write

$$\int x\,dx = \frac{x^2}{2} + C\ , \quad \int x^2\,dx = \frac{x^3}{3} + C\ .$$

Are the C's supposed to be the same? If we have these two statements, what about

$$\int (x + x^2)\, dx \ ?$$

Is this $\frac{x^2}{2}+\frac{x^3}{3}+C$, or $\frac{x^2}{2}+\frac{x^3}{3}+2C$? It is standard to write the first, since by convention the C in each statement is not a specific constant, but simply denotes that any constant can be added. This convention, although universal, is a little unfortunate. Another instance of violation of usual ambiguity rules of mathematical language is in the big-O and little-o notation ().

II.1.8. Languages and Relational Systems

We should really have begun the entire discussion of formal statements by first specifying a language and a domain of discourse. We did not do this at first because we didn't want to make the discussion too abstract. But there is merit in considering systems of formal statements based on various languages, or, from a different point of view, systems where the language and statements are restricted to be of a certain form.

II.1.8.1. DEFINITION. A *relational system* is a triple $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$, where Ω is a set, $\mathcal{R}$ is a set of relations on Ω, and $\mathcal{F}$ a set of functions on Ω (with codomain Ω). The relations and functions may be n-ary for any $n \geq 1$, which may vary from relation to relation, and may have domains which are proper subsets. The set Ω is called the *domain of discourse* of the relational system.

Every function is just a special kind of relation, so it seems unnecessary to explicitly also include a set of functions in the definition; but the relations and functions are used in very different ways, so it is necessary to specify both $\mathcal{R}$ and $\mathcal{F}$ as part of the definition. Note that $\mathcal{R}$ and/or $\mathcal{F}$ need not contain all relations or functions on Ω.

The discussion of the previous part of this section concerned the relational system where Ω was some universal set containing all objects under discussion (in many cases Ω could be taken to be the real numbers $\mathbb{R}$), and $\mathcal{R}$ and $\mathcal{F}$ the sets of all relations and all functions on Ω.

II.1.8.2. DEFINITION. Let $\mathcal{S}$ be a relational system. A *language* for the relational system $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$ is:

(i) A set of *symbols*

(ii) A set of *terms*

(iii) A set of *formulas*

where the set of symbols includes at least one constant symbol for each element of Ω and each relation in $\mathcal{R}$ and function in $\mathcal{F}$, called its *name* (an element of Ω can have more than one name); a set of *variables*; a set of logical symbols including $\wedge$, $\Rightarrow$, and $\forall$; and parentheses. The terms satisfy

(1) Each constant or variable symbol is a term.

(2) If $f \in \mathcal{F}$ is a function of n arguments, and $t_1, \ldots, t_n$ are terms, then $f(t_1, \ldots, t_n)$ is a term.

A variable x *occurs* in the term $f(t_1, \ldots, t_n)$ if it equals or occurs in one of the t_k.
The formulas are certain finite strings of symbols including all strings of the form

(a) $R(t_1, \ldots, t_n)$, where R is an n-ary relation in $\mathcal{R}$ and each t_k is a term involving only constants which are names of elements of Ω and variables.

(b) $\forall x(P)$, where P is a formula and x is a variable not quantified in P. The variable x is then *quantified* in $\forall x(P)$.

(c) $P_1 \wedge \cdots \wedge P_n$, where $P_1, \ldots, P_n$ are formulas, and if x is a variable which is quantified in some P_k, then x does not appear unquantified in any other P_j.

(d) $P_1 \Rightarrow P_2$, where P_1 and P_2 are formulas, and if x is a variable which is quantified in one P_k, then x does not appear unquantified in the other P_j.

A *statement*, or *closed statement* for emphasis, is a formula in which all variables are quantified, i.e. in which there are no *free variables* (variables which occur and are not quantified).

If every term in the language is obtained by a finite number of applications of (1) and (2), and every formula built up from a finite number of applications of (a), (b), (c), and (d), the language is a *simple language* for $\mathcal{S}$. Each term and each formula then has a well-defined *level*, and is built up from terms or formulas of strictly lower level.

In some references the word "sentence" is used in place of "formula" and in some other references "sentence" is used in place of "statement." Because of the nonuniformity of usage, we have avoided the word "sentence" entirely. In the earlier part of this section, the word "statement" was used a little informally; in many of the places it was used we should have instead said "formula."

II.1.8.3. A language can be a language for more than one relational system. But it is very hard for a language to be a simple language for more than one relational system, although it can be if the names of elements and relations are chosen in a different way. Conversely, a relational system can have many associated languages, but a simple language for a relational system is essentially uniquely determined by the relational system.

II.1.8.4. Suppose $\mathcal{L}$ is a language for a relational system $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$, which we will take to be a simple language. It makes sense to say whether or not closed statements in $\mathcal{L}$ (and *only* closed statements!) are "true" in $\mathcal{S}$:

II.1.8.5. DEFINITION. Suppose $\mathcal{L}$ is a simple language for a relational system $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$. If t is a term in $\mathcal{L}$, then

(i) If t is a constant c which is the name of an element $\omega \in \Omega$, then t is *interpretable in* $\mathcal{S}$ as ω.

(ii) If $t = f(t_1, \ldots, t_n)$ and $t_1, \ldots, t_n$ are interpretable in $\mathcal{S}$ as $\omega_1, \ldots, \omega_n$ respectively, and $(\omega_1, \ldots, \omega_n)$ is in the domain of f, then t is *interpretable in* $\mathcal{S}$ as $f(\omega_1, \ldots, \omega_n)$.

If t is interpretable in $\mathcal{S}$ via a finite string of applications of (i) and (ii), then t is *interpretable* in $\mathcal{S}$; otherwise t is not interpretable in $\mathcal{S}$.

II.1.8.6. DEFINITION. Suppose $\mathcal{L}$ is a simple language for a relational system $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$. Then

(i) If P is the statement $R(t_1, \ldots, t_n)$ involving no variables, and $t_1, \ldots, t_n$ are interpretable in $\mathcal{S}$ as $\omega_1, \ldots, \omega_n$ respectively, and $(\omega_1, \ldots, \omega_n) \in R$ (i.e. $R(\omega_1, \ldots, \omega_n)$ holds in Ω), then P *holds* (or *is true*) in $\mathcal{S}$.

(ii) If P is the formula $R(t_1, \ldots, t_n)$, with all variables occurring in $t_1, \ldots, t_n$ included among $x_1, \ldots, x_m$, $c_1, \ldots, c_m$ are names of elements of Ω, and P holds when, for each k, c_k is substituted for each instance of x_k in P, then P *holds* (or *is true*) for $(c_1, \ldots, c_m)$ in $\mathcal{S}$.

(iii) If P is a formula containing only unquantified variables among $x_1, \ldots, x_m, x_{m+1}, \ldots, x_p$ and quantified variables among $x_{p+1}, \ldots, x_r$, $c_1, \ldots, c_m$ are names of elements of Ω, and for every set of constants $c_{m+1}, \ldots, c_r$ in $\mathcal{L}$ which are names of elements of Ω, P holds for $(c_1, \ldots, c_r)$, then

$$\forall x_{m+1} \cdots \forall x_p(P)$$

holds (or *is true*) for $(c_1, \ldots, c_m)$ in $\mathcal{S}$. If $m = 0$, then the statement

$$\forall x_1 \cdots \forall x_p(P)$$

holds (or *is true*) in $\mathcal{S}$.

(iv) If $P_1, \ldots, P_n$ are formulas containing only unquantified variables among $x_1, \ldots, x_m$, with no variable quantified in some P_k and free in another P_j, $c_1, \ldots, c_m$ are names of elements of Ω, and each sentence P_j holds for $(c_1, \ldots, c_m)$, then

$$P_1 \wedge \cdots \wedge P_n$$

holds (or *is true*) for $(c_1, \ldots, c_m)$ in $\mathcal{S}$. If $m = 0$, i.e. if $P_1, \ldots, P_n$ are statements, then the statement

$$P_1 \wedge \cdots \wedge P_n$$

holds (or *is true*) in $\mathcal{S}$.

(v) If P_1 and P_2 are formulas containing only the free variables $x_1, \ldots, x_m$, none of which are quantified in one and free in the other, $c_1, \ldots, c_m$ are names of elements of Ω for which P_2 holds in $\mathcal{S}$ or P_1 does not hold in S, then

$$P_1 \Rightarrow P_2$$

holds (or *is true*) for $(c_1, \ldots, c_m)$ in $\mathcal{S}$. If $m = 0$, i.e. if P_1 and P_2 are statements, then the statement

$$P_1 \Rightarrow P_2$$

holds (or *is true*) in $\mathcal{S}$.

If P is a statement in $\mathcal{L}$, and P is built up from a finite chain of formulas which hold in $\mathcal{S}$, then P holds in $\mathcal{S}$; otherwise P *does not hold* (or is *false*) in $\mathcal{S}$.

II.1.8.7. There are two quite distinct ways a statement can fail to hold:

(1) At some point, a term is not interpretable.

(2) All terms are always interpretable, but a relation does not hold.

II.1.8.8. Unless the set $\mathcal{R}$ of relations is empty (in which case $\mathcal{L}$ has no formulas) there are statements which hold in $\mathcal{S}$: if P is any formula, and $x_1, \ldots, x_m$ are the unquantified variables in P, then the statement

$$\forall x_1 \cdots \forall x_m (P \Rightarrow P)$$

holds in $\mathcal{S}$. It can happen that every statement in $\mathcal{L}$ holds in $\mathcal{S}$, but only if $\mathcal{R} = \emptyset$, or if $\mathcal{R}$ consists only of the full relation and, for every n, every n-ary function in $\mathcal{F}$ has full domain Ω^n. Such a relational system is called *degenerate*. In a nondegenerate relational system there is a statement $\Phi = R(t_1, \ldots, t_n)$ containing no variables which is false (and there are also statements which are true, e.g. $\Phi \Rightarrow \Phi$).

II.1.8.9. If we fix such a Φ, then for any formula P containing free variables among $x_1, \ldots, x_m$, and every set $c_1, \ldots, c_m$ of names of elements of Ω, the formula $P \Rightarrow \Phi$ holds for $(c_1, \ldots, c_m)$ in $\mathcal{S}$ if and only if P does not hold for $(c_1, \ldots, c_m)$ in $\mathcal{S}$. We can thus interpret this formula as $\neg P$, i.e. we can extend $\mathcal{L}$ by adding the symbol $\neg$, and formulas $\neg(P)$, where the formula $\neg(P)$ in the extended language is translated to mean $P \Rightarrow \Phi$ in the simple language.

II.1.8.10. We can then further extend $\mathcal{L}$ by adding the symbols $\vee$ and $\exists$, where

$$P_1 \vee \cdots \vee P_n$$

is translated to mean

$$\neg((\neg(P_1)) \wedge \cdots \wedge (\neg(P_n)))$$

(this will be a formula of the extended language if no variable is quantified in one of the P_k and unquantified in another P_j), and

$$\exists x_1 \cdots \exists x_m (P)$$

is translated to mean

$$\neg(\forall x_1 \cdots \forall x_m (\neg(P)))$$

(this will be a formula of the extended language if no x_k is quantified in P). More complicated formulas of the extended language can be built up from repeated applications of these translations.

II.1.8.11. We may thus extend a simple language $\mathcal{L}$ for a nondegenerate relational structure $\mathcal{S}$ to an *extended simple language* $\bar{\mathcal{L}}$, which includes all the standard logical symbols with their usual meanings, for which every formula has a well-defined translation into a formula of $\mathcal{L}$, and it makes sense to say whether statements in $\bar{\mathcal{L}}$ are true or false in $\mathcal{S}$.

II.1.8.12. In the extended simple language, the usual rules for taking negations hold. We can define an additional logical symbol $\Longleftrightarrow$ by translating $P \Longleftrightarrow Q$ as $(P \Rightarrow Q) \wedge (Q \Rightarrow P)$, and we say P and Q are *equivalent* in $\mathcal{S}$ if $P \Longleftrightarrow Q$ holds in $\mathcal{S}$. Then

$$\neg(\neg(P)) \Longleftrightarrow P$$

for every statement P. A statement form like this which holds in $\mathcal{S}$ for every statement P is called a *tautology* in $\mathcal{S}$. The other tautologies in II.1.6.1. are also tautologies in $\mathcal{S}$.

II.1.9. Mathematical Arguments and Formal Proofs

II.1.9.1. Informally, a mathematical argument consists of a derivation of a mathematical statement (conclusion) from a set of initial assumptions, using acceptable mathematical reasoning. The argument, if logically sound, constitutes a *proof* of the conclusion.

II.1.9.2. The initial assumptions can be divided into two kinds: general assumptions underlying an entire mathematical theory, called *axioms*, and assumptions particular to the individual statement being proved, called *hypotheses*. There is not always a clear dividing line between axioms and hypotheses, and anyway the distinction between them is unimportant for the purpose of making and understanding mathematical arguments.

II.1.9.3. In practice, mathematical arguments are usually written out in a combination of English (or another language) and formal symbolism.

A formal mathematical argument

II.1.10. Exercises

II.1.10.1. Let $P(x)$ and $Q(x)$ be statements with a free variable x.

(a) Show that the statement $\forall x(P(x) \wedge Q(x))$ is logically equivalent to $[\forall x(P(x))] \wedge [\forall x(Q(x))]$.

(b) Show that the statement $\forall x(P(x) \vee Q(x))$ is *not* logically equivalent to $[\forall x(P(x))] \vee [\forall x(Q(x))]$.

(c) Show that the statement $\exists x(P(x) \vee Q(x))$ is logically equivalent to $[\exists x(P(x))] \vee [\exists x(Q(x))]$.

(d) Show that the statement $\exists x(P(x) \wedge Q(x))$ is *not* logically equivalent to $[\exists x(P(x))] \wedge [\exists x(Q(x))]$.

(e) Which implications are valid in (b) and (d)?

II.1.10.2. (a) Show that $P \Rightarrow P$ and $P \Rightarrow (Q \Rightarrow Q)$ are tautologies.

(b) Are $(P \Rightarrow P) \Rightarrow Q$ and $P \Rightarrow (P \Rightarrow Q)$ tautologies?

(c) Is $\Rightarrow$ "associative," i.e. is

$$[(P \Rightarrow Q) \Rightarrow R] \Longleftrightarrow [P \Rightarrow (Q \Rightarrow R)]$$

a tautology?

(d) There are three possible interpretations of the statement $P \Rightarrow Q \Rightarrow R$: as $(P \Rightarrow Q) \Rightarrow R$, $P \Rightarrow (Q \Rightarrow R)$, or as $(P \Rightarrow Q) \wedge (Q \Rightarrow R)$. What is the logical relationship between these three interpretations? Which is the "best" interpretation?

(e) Discuss the same question as in (d) for the statement $P \Longleftrightarrow Q \Longleftrightarrow R$.

II.1.10.3. If $\mathcal{S}$ is nondegenerate, a statement Φ is false if and only if $\Phi \Rightarrow P$ holds for every statement P. Show that any two false statements are equivalent.

II.1.10.4. Give a careful justification for all the assertions in II.1.8.11. and II.1.8.12..

II.1.10.5. [Bar00] Is

$$(P \Rightarrow Q) \vee (Q \Rightarrow P)$$

a tautology? Does this mean that for any statements P and Q, either P implies Q or Q implies P? In particular, does it mean "If a number is prime, then it is odd, or if a number is odd, then it is prime"?

II.1.10.6. [L55] Let Q be any statement. Let P be the statement "If this statement is true, then Q." Thus P asserts: "If P is true, then Q" or $P \Rightarrow Q$. So P is true unless P is true and Q is false. But if P is true, then Q is true since this is what "P is true" means! Thus Q is true. Does this prove that every statement is true?

II.1.10.7. Here is an interesting logic problem. A convention of logicians tries a game. They all sit in a room, the lights are turned out, and a hat is placed on each one, either red or green. There are more green hats than red, but several reds (the logicians are not told anything about the number of red hats). The lights are turned back on. A bell rings every five minutes, and whenever a logician figures out the color of his/her hat (and can prove it), they leave the room the next time the bell rings and get a prize. They are not allowed to look at their own hat or discuss hat color with anyone, but of course they can each see the color of everyone else's hat. They all know that everyone in the room is an excellent logician and can draw any logical conclusions possible in any situation (and that everyone can see and is not color-blind).

For a long time nothing happens. Then someone comes into the room and says, "There are people here wearing red hats and green hats!" and then leaves again. Nothing happens for a while after this except that the the bell rings several times. Then suddenly the next time the bell rings all the logicians with red hats get up and leave. The next time the bell rings, all the people with green hats leave too.

Explain why. Why did the situation change after the visitor's statement, even though the visitor didn't appear to tell the logicians anything they didn't already know? [Hint: What does "common knowledge" mean?]

Versions of this problem have been extensively discussed for a long time. See [Bar00] for a long list of references.

II.1.10.8. Here is a somewhat similar logic problem. You and a friend are thrown in prison in a foreign country. You are in cells on opposite sides of the prison with no opportunity to communicate. The warden explains to each of you that you have an opportunity to be released. He says there are either 18 or 20 statues on the prison grounds. Each of you can see some of them through your cell window; every statue is visible to exactly one of you. Each morning the warden will come to you and ask how many statues there are in all. You can either answer or decline to answer. If you answer correctly, you will both be released immediately, but if you answer incorrectly the process ends and you will never be released. If you decline to answer, the warden will go to your friend and ask the same question with the same consequences. If your friend also declines to answer, the warden will come back to you the next morning and begin the process again. You both know that you will be asked first and your friend second each day.

You can see 12 statues from your window. You also know your friend is an excellent logician and can deduce anything it is possible to deduce, and presumably they know the same about you. How can you get yourselves released? [Hint: a selective memory is helpful; each of you should set aside what you know at the outset about how many statues the other sees and act only on information obtained from the question process, plus the number of statues you see yourself.]

II.1.10.9. The Prisoner's Dilemma. A prisoner is told on Sunday that he will be hanged at dawn some day in the coming week, and that he will not know in advance which day it will be. The prisoner reasons: "They can't wait until Sunday, since if I'm not hanged by Saturday I will know in advance that it will be Sunday. Then they can't wait until Saturday, since if I'm not hanged by Friday I know it will have to be Saturday since I already know it can't be Sunday. By the same argument they can't wait until Friday, or Thursday, etc. Thus they cannot hang me this week at all." The hangman comes at dawn on Thursday and catches the prisoner by complete surprise. Explain why the prisoner's reasoning was incorrect.

> A professor was giving a lecture. At one point, a student interrupted him and said, "Excuse me, professor, but I don't understand the second equation on the left-hand board." The professor stared at him for a moment, and then continued his lecture. The student said, "But, professor, you didn't answer my question." The professor replied, "That was not a question." [Supposedly a true story; the professor will remain anonymous.]

Theorems, Corollaries, Propositions, Lemmas

The terms *Theorem*, *Corollary*, *Proposition*, and *Lemma* all mean essentially the same thing, with nuances of difference. All refer to mathematical statements which must be, and are, proved.

The terms are largely interchangeable, and the same statements can be found in various books called by any of the names. But there are conventions which are generally followed, although they are subjective and subject to interpretation. So the following descriptions should be taken with a grain of salt, and are intended only as a guide:

II.1.10.10. A *Theorem* is a result which (i) is of major significance and/or (ii) is difficult to prove. Many have widely-used official names, e.g. "The Fundamental Theorem of Calculus." It is good practice to use the term "Theorem" sparingly, to really call attention to the big results.

II.1.10.11. A *Corollary* is a result which follows immediately or easily from another result. Sometimes a Corollary of a Theorem is simply a crucial special case of the theorem. A result which is a combination of two or more Theorems can also be called a Corollary.

II.1.10.12. A *Proposition* is a statement which (i) is less significant than a Theorem and/or (ii) has an easy or obvious proof. Propositions are not necessarily insignificant: fundamental immediate consequences of definitions and other observations are often called Propositions, with the proof often omitted if it is thought obvious.

II.1.10.13. A *Lemma* is a statement which is not of particular interest in and of itself, but which is of use in proving something else of greater interest. Statements of lemmas are often rather technical, and proofs can be complicated (although neither is a necessary attribute). Lemmas are roughly used in two ways. It is sometimes convenient to break a long proof involving many steps into a sequence of Lemmas to make the overall argument easier to follow and understand. And sometimes a subsidiary result which must be used repeatedly in the main argument can be conveniently isolated as a Lemma.

The best description of a Lemma I have heard came when I was a student in a course on Automata Theory. It was a mathematics course, with theorems and proofs, but many of the students were graduate students in Computer Science, with a different outlook than mathematicians. One day the instructor wrote a Lemma on the board, and one of the students asked, "What's a Lemma?" Another CS student immediately replied, "It's a subroutine."

Germans sometimes use the nicely descriptive term *Hilfssatz* ("helper theorem") instead of *Lemma* (although *Lemma* is often used too, and its plural is written *Lemmata*, which is proper Greek).

There are statements commonly called Lemmas which are significant Theorems, e.g. "Urysohn's Lemma" (XIII.7.6.13.) (which is probably called a "Lemma" simply to distinguish it from "Urysohn's Metrization Theorem", although it is a crucial tool in the proof of the Metrization Theorem). And in subjects like homological algebra, basic results like the Snake Lemma (XV.3.1.10.) are called "Lemmas" since their primary importance is as tools in other arguments.

II.1.10.14. The terms are occasionally used in a somewhat different sense. For example, the name "Fermat's Last Theorem" has been used for nearly two hundred years, even though it only became a theorem (or, perhaps more accurately, became known to be a theorem, i.e. was proved) quite recently. And "Zorn's Lemma" (II.6.2.9.) is not really a Lemma, but a useful equivalent restatement of the Axiom of Choice.

II.1.10.15. Certain theorem-level results have common names using other terms, such as the "Product Rule," "l'Hôpital's Rule," the "Sequential Criterion for Continuity," the "Ratio Test," the "Cauchy Integral Formula," the "Principle of the Argument," etc.

Is a two-part lemma a dilemma?

II.2. Set Theory

II.2.1. New Sets from Old

The Power Set of a Set

Ordered Pairs and Cartesian Products

The Union of a Set

The Constituent Set of a Set

II.2.1.1. DEFINITION. Let X be a set. Inductively set

$$\mathcal{C}_1(X) = \cup X = \{y : \exists x \in X \text{ with } y \in x\}$$
$$\mathcal{C}_{n+1}(X) = \mathcal{C}_1(\mathcal{C}_n(X))$$
$$\mathcal{C}(X) = \cup_{n=1}^{\infty} \mathcal{C}_n(X) \ .$$

The set $\mathcal{C}(X)$ is called the *constituent set* of X.

Intuitively, $X \cup \mathcal{C}(X)$ is the collection of all things X is "made up of." The constituents of X, i.e. the elements of $\mathcal{C}(X)$, are the things that the elements of X are made of. (It is a matter of convention not to assume that $X \subseteq \mathcal{C}(X)$, i.e. to define $\mathcal{C}(X)$ as we have instead of as $X \cup (\cup_{n=1}^{\infty} \mathcal{C}_n(X))$, but our definition turns out to be more notationally useful.)

II.2.1.2. $\mathcal{C}(X)$ is a set by the Axiom of Union and the Axiom of Infinity. However, the cardinality of $\mathcal{C}(X)$ is quite independent of $card(X)$, as the following examples show (but see II.2.1.6.).

II.2.1.3. EXAMPLES. (i) $\mathcal{C}_1(\emptyset) = \mathcal{C}(\emptyset) = \emptyset$. $\mathcal{C}_1(\{\emptyset\}) = \mathcal{C}(\{\emptyset\}) = \emptyset$. $\mathcal{C}_1(\{\{\emptyset\}\}) = \mathcal{C}(\{\{\emptyset\}\}) = \{\emptyset\}$. $\mathcal{C}_1(\{\{\{\emptyset\}\}\}) = \{\{\emptyset\}\}$ and $\mathcal{C}_2(\{\{\{\emptyset\}\}\}) = \{\emptyset\}$, so $\mathcal{C}(\{\{\{\emptyset\}\}\}) = \{\emptyset, \{\emptyset\}\}$.
(ii) If σ is a successor ordinal (II.8.2.8.), then $\mathcal{C}_1(\sigma) = \mathcal{C}(\sigma) = \sigma - 1$. If σ is a limit ordinal, then $\mathcal{C}_1(\sigma) = \mathcal{C}(\sigma) = \sigma$.
(iii) If x is a set of cardinality κ, and $X = \{x\}$, then $card(X) = 1$ but $card(\mathcal{C}_1(X)) = \kappa$ since $\mathcal{C}_1(X) = x$. The cardinality of $\mathcal{C}(X)$ may be even larger.
(iv) If the elements of X are not sets (i.e. are urelements ()), then $\mathcal{C}(X) = \emptyset$ no matter how large $card(X)$ is.

II.2.1.4. If $x \in X$, then $\mathcal{C}_n(x) \subseteq \mathcal{C}_{n+1}(X)$ for each n, and thus $\mathcal{C}(x) \subseteq \mathcal{C}(X)$. More generally, if $x \in \mathcal{C}(X)$, then $\mathcal{C}(x) \subseteq \mathcal{C}(X)$. If $Y \subseteq X$, then $\mathcal{C}_n(Y) \subseteq \mathcal{C}_n(X)$ for every n, so $\mathcal{C}(Y) \subseteq \mathcal{C}(X)$. Similarly, if $Y \subseteq \mathcal{C}(X)$, then $\mathcal{C}(Y) \subseteq \mathcal{C}(X)$. In particular, we have $\mathcal{C}(\mathcal{C}(X)) \subseteq \mathcal{C}(X)$, but they need not be equal (II.2.1.3.(ii)).

II.2.1.5. Let X be a set. Since $\mathcal{C}_1(\mathcal{P}(X)) = X$, $\mathcal{C}(\mathcal{P}(X)) = X \cup \mathcal{C}(X)$.

II.2.1.6. If X contains no urelements, then $X \subseteq \mathcal{P}(\mathcal{C}_1(X))$, and hence

$$card(X) \leq 2^{card(\mathcal{C}_1(X))} \leq 2^{card(\mathcal{C}(X))}$$

so $card(\mathcal{C}(X))$ cannot be "too much smaller" than $card(X)$.

II.3. Relations and Functions

The idea of a relation is a crucial one, both within and outside of mathematics. Mathematicians have formalized the intuitive idea of a relation, and special kinds of relations abound throughout mathematics. The three most important kinds of relations in mathematics are *functions*, *equivalence relations*, and *partial orderings*. The first two will be discussed in this section, and partial orderings described in the next section.

II.3.1. General Relations

To formalize the notion of a relation between elements of a set X and a (possibly different) set Y, we can think of the relation as consisting of certain ordered pairs (x, y). Thus the formal definition is:

II.3.1.1. DEFINITION. A *relation* from a set X to a set Y is a subset $\mathcal{R}$ of $X \times Y$. If $(x, y) \in X \times Y$, write $x\mathcal{R}y$ if $(x, y) \in \mathcal{R}$. If $X = Y$, $\mathcal{R}$ is called a *relation on* X.

The set

$$\{x \in X : x\mathcal{R}y \text{ for some } y \in Y\}$$

is called the *domain* of $\mathcal{R}$, denoted $\mathrm{Dom}(\mathcal{R})$; and the set

$$\{y \in Y : x\mathcal{R}y \text{ for some } x \in X\}$$

is called the *range* of $\mathcal{R}$, denoted $\mathrm{Range}(\mathcal{R})$.

II.3.1.2. EXAMPLES.

(i) The relation $\leq$ can be regarded as a relation on $\mathbb{R}$: it is technically the set

$$\{(x, y) \in \mathbb{R}^2 : x \leq y\} \subseteq \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} \ .$$

(ii) Let S be the set of students at Ivory State University, and C the set of classes offered this semester. Define a relation $\mathcal{E}$ from S to C by letting $s\mathcal{E}c$ if student s is enrolled in class c. (There are obviously an enormous variety of variations on this example.) The domain is probably not quite all of S, since there may be students not taking any classes this semester; the range may or may not be all of C, depending on whether classes with zero enrollment are counted.

(iii) Let X and Y be any sets. Then $\emptyset$ and $X \times Y$ are relations from X to Y, called the *empty relation* and *full relation* respectively.

(iv) If X is a set, the *identity relation* on X is the relation $\mathcal{I}$ (often written $\mathcal{I}_X$) defined by $x\mathcal{I}y$ if and only if $x = y$. The domain and range of $\mathcal{I}$ are all of X.

II.3.1.3. DEFINITION. Let $\mathcal{R}$ and $\mathcal{S}$ be relations from X to Y. $\mathcal{S}$ is an *extension* of $\mathcal{R}$ if $\mathcal{R} \subseteq \mathcal{S}$, i.e. $x\mathcal{R}y$ implies $x\mathcal{S}y$.
If $A \subseteq X$, the *restriction* of $\mathcal{R}$ to A is

$$\mathcal{R}|_A = \mathcal{R} \cap (A \times Y) = \{(x, y) : x \in A,\ y \in Y,\ x\mathcal{R}y\} \ .$$

Note that $\mathcal{R}$ is an extension of $\mathcal{R}|_A$. However, not every extension comes from a restriction in this way.

II.3.1.4. Definition. Let $\mathcal{R}$ be a relation from X to Y. The *inverse relation* $\mathcal{R}^{-1}$ is the relation from Y to X defined by

$$\mathcal{R}^{-1} = \{(y, x) : (x, y) \in \mathcal{R}\}$$

(in other words, $y\mathcal{R}^{-1}x$ if and only if $x\mathcal{R}y$.) The domain of $\mathcal{R}^{-1}$ is the range of $\mathcal{R}$, and vice versa; and $(\mathcal{R}^{-1})^{-1} = \mathcal{R}$.

As a relation, $\mathcal{R}^{-1}$ appears to be only superficially or notationally different from $\mathcal{R}$; however, it can have rather different properties, as we will see later.

Composition

Relations can be composed:

II.3.1.5. Definition. Let $\mathcal{R}$ be a relation from a set X to a set Y, and $\mathcal{S}$ a relation from Y to a set Z. The *composition* of $\mathcal{S}$ and $\mathcal{R}$, written $\mathcal{S} \circ \mathcal{R}$, is the relation from X to Z defined by

$$\mathcal{S} \circ \mathcal{R} = \{(x, z) \in X \times Z : x\mathcal{R}y \text{ and } y\mathcal{S}z \text{ for some } y \in Y\} .$$

(Note the order in which $\mathcal{R}$ and $\mathcal{S}$ are written.)

II.3.1.6. Examples.

(i) Let $\mathcal{E}$ be the relation from S to C defined in II.3.1.2.(ii), and let $\mathcal{L}$ be the relation from C to the set B of buildings at Ivory State with $c\mathcal{L}b$ if class c meets at least once a week in building b. Then $\mathcal{L} \circ \mathcal{E}$ is the relation $\mathcal{R}$ from S to B with $s\mathcal{R}b$ if and only if student s has at least one class a week in building b.

(ii) If $\mathcal{I}$ is the identity relation on a set X, and $\mathcal{R}$ is any relation from X to Y, then $\mathcal{R} \circ \mathcal{I} = \mathcal{R}$.

(iii) If $\mathcal{F}$ is the full relation from X to Y, and $\mathcal{G}$ the full relation from Y to Z, then $\mathcal{G} \circ \mathcal{F}$ is the full relation from X to Z (provided $Y \neq \emptyset$.)

(iv) If $<$ is the usual strict inequality relation on $\mathbb{R}$, then $(< \circ <) = <$. However, the relation $< \circ <$ on $\mathbb{N}$ is not the same as $<$ on $\mathbb{N}$: $m(< \circ <)n$ if and only if $n - m \geq 2$. Thus the sets on which a relation is defined are very important to specify carefully.

(v) The composition of $<$ and $>$ on $\mathcal{R}$ is the full relation on $\mathcal{R}$.

(vi) Let P be the set of all people. Define relations $\mathcal{B}$ and $\mathcal{S}$ on P by saying $x\mathcal{B}y$ if x is the brother of y, and $x\mathcal{S}y$ if x is the sister of y. Then (disregarding modern surgical procedures) $\mathcal{B} \circ (\mathcal{S}^{-1})$ is the empty relation on P.

II.3.1.7. The domain of $\mathcal{S} \circ \mathcal{R}$ is a subset of the domain of $\mathcal{R}$, and the range is a subset of the range of $\mathcal{S}$. Either or both containments can be proper. See II.3.4.1..

Images and Preimages

II.3.1.8. DEFINITION. Let $\mathcal{R}$ be a relation from X to Y, and let $A \subseteq X$. Then the *image* of A under $\mathcal{R}$ is

$$\mathcal{R}(A) = \{y \in Y : x\mathcal{R}y \text{ for some } x \in A\} .$$

If $B \subseteq Y$, the *preimage* (or *inverse image*) of B under $\mathcal{R}$ is

$$\mathcal{R}^{-1}(B) = \{x \in X : x\mathcal{R}y \text{ for some } y \in B\} .$$

II.3.1.9. The notation is consistent: $\mathcal{R}^{-1}(B)$ is the image of B under the relation $\mathcal{R}^{-1}$. Thus images and preimages can be considered on the same basis. We prefer the term "preimage" to "inverse image," especially in the case of functions, where "inverse image" can mistakenly seem to imply the existence of an inverse function (). Some authors use the notation $\mathcal{R}[A]$ and $\mathcal{R}^{-1}[B]$ for $\mathcal{R}(A)$ and $\mathcal{R}^{-1}(B)$ respectively.

Images (and hence also preimages) behave nicely with respect to unions, but not quite so nicely under intersections:

II.3.1.10. PROPOSITION. Let $\mathcal{R}$ be a relation from X to Y, and $\{A_i : i \in I\}$ an indexed collection of subsets of X. Then

(i) $\mathcal{R}(\cup_i A_i) = \cup_i \mathcal{R}(A_i)$.

(ii) $\mathcal{R}(\cap_i A_i) \subseteq \cap_i \mathcal{R}(A_i)$.

The proof is left as an exercise (Exercise ().)

II.3.1.11. Simple examples show that equality does not hold in (ii) in general. For example, if $\mathcal{F}$ is the full relation from X to Y, and A is any nonempty subset of X, then $\mathcal{F}(A) = Y$. If A_1 and A_2 are nonempty but $A_1 \cap A_2 = \emptyset$, then

$$\emptyset = \mathcal{F}(\emptyset) = \mathcal{F}(A_1 \cap A_2) \neq \mathcal{F}(A_1) \cap \mathcal{F}(A_2) = Y .$$

II.3.2. Functions

Functions are officially relations of a special kind, although they are not usually thought of in this way.

II.3.2.1. DEFINITION. Let X and Y be sets. A *function* from X to Y is a relation f in which no two ordered pairs have the same first coordinate, i.e. if $x \in X$ and xfy_1 and xfy_2 for some $y_1, y_2 \in Y$, then $y_1 = y_2$.

Thus, if f is a function, then for every $x \in \mathrm{Dom}(f)$ there is a *unique* $y \in Y$ with xfy. We normally write $f(x) = y$ instead of xfy, and call y the *value* of f at x or the *image* of x under f.

The vast majority of functions considered in analysis are ones taking numerical values, but there is no reason to restrict attention only to numerical functions.

II.3.2.2. EXAMPLES.

(i) If X and Y are sets, the empty relation from X to Y and the identity relation on X are functions. The identity relation on X, when regarded as a function, is written ι_X. The full relation from X to Y is not a function unless Y is a singleton (or $\emptyset$).

(ii) Let X and Y be sets, and y_0 a fixed element of Y. Set $f(x) = y_0$ for all $x \in X$. Then f is a function, called the *constant function* y_0. By slight abuse of notation, we will often use the symbol y_0 as the name of this constant function, especially if Y is a set of numbers. Thus, for example, we will often use a phrase like "the constant function 5", meaning the function with $f(x) = 5$ for all x.

(iii) The relations $\mathcal{E}$ of II.3.1.2.(ii) and $\mathcal{B}$ and $\mathcal{S}$ of II.3.1.6.(vi) are not functions since a student can take more than one class and a person can have more than one brother or sister. The relation $\mathcal{L}$ of II.3.1.6.(i) is also probably not a function, since a class can meet in different buildings on different days (for example, it could have a lab or discussion session in another building.)

(iv) The set $\{(x, y) \in \mathbb{R}^2 : y = x^2\}$ is a typical function from $\mathbb{R}$ to $\mathbb{R}$ defined by a formula.

II.3.2.3. **Terminology:** It is especially important with functions to carefully specify the domain; thus we should always specify something like "f is a function from X to Y with domain A." Without further qualification, the phrase "f is a function from X to Y" will be taken to mean that $\mathrm{Dom}(f)$ is all of X. In this case, we often write $f : X \to Y$. When f is defined by a formula or rule, without explicit specification to the contrary the domain will always be taken to be all elements of X for which the formula or rule makes sense. Thus, when we write $f : X \to Y$, the domain X is determined by f. The set Y is called the *codomain* of f when written in this way. It is a subtle matter whether the Y should be regarded as determined by f, since it is not the range of f in general. See II.3.2.13..

Mappings

Functions are usually thought of as mappings:

II.3.2.4. DEFINITION. A *mapping* f from a set X to a set Y is an assignment to each $x \in X$ of a unique element $f(x)$ of Y. A mapping f is often represented pictorially:

II.3.2.5. This definition of a mapping is not quite precise enough mathematically, hence the formal definition of function given above. But there is a natural correspondence: a function can obviously be pictured as a mapping, and a mapping f from X to Y gives a function

$$\{(x, y) \in X \times Y : f(x) = y\} .$$

This set is often called the *graph* of the mapping f; formally, a function *is* its own graph. Thinking of functions and graphs as the same thing can be confusing, but is technically useful.

II.3.2.6. We are in the habit of defining functions, especially numerical functions, by algebraic formulas; we typically write something like "let $f(x) = \dfrac{\cos x}{x-2}$" to define a function, where we really mean

$$f = \left\{(x,y) \in \mathbb{R}^2 : y = \frac{\cos x}{x-2}\right\} .$$

By our convention, the domain of this function would be $\{x \in \mathbb{R} : x \neq 2\}$.

Images and Preimages

Since functions are relations, the definitions of image and preimage (II.3.1.8.) make sense. We rephrase them in function notation:

II.3.2.7. DEFINITION. Let $f : X \to Y$ be a function.
(i) If $A \subseteq X$, the *image* of A under f is

$$f(A) = \{f(x) : x \in A\} \subseteq Y .$$

(ii) If $B \subseteq Y$, the *preimage* (or *inverse image*) of B under f is

$$f^{-1}(B) = \{x \in X : f(x) \in B\} \subseteq X .$$

II.3.2.8. We prefer the term "preimage" to "inverse image," since there is not necessarily an inverse function $f^{-1} : Y \to X$ (there is an inverse *relation*, which is not a function in general). The notation $f^{-1}(B)$ can also be confusing, but is standard; note that this notation does *not* mean there is an inverse function! Preimages are defined for arbitrary functions and arbitrary subsets of the codomain. In the special case that there is an inverse function, the preimage of a set is indeed just the image under the inverse function, so the notation is consistent.

II.3.2.9. If $f : X \to Y$ and $a \in X$, there is a subtle difference between $f(a)$ and $f(\{a\})$: $f(a)$ is an element of Y, while $f(\{a\})$ is a subset of Y. In fact, $f(\{a\}) = \{f(a)\}$. The distinction is more pronounced for preimages: if $b \in Y$, then $f^{-1}(\{b\})$ is a subset of X which can be a singleton, but can also be empty (if b is not in the range of f) or it can have many elements. The notation $f^{-1}(b)$ does not make sense unless f has an inverse function f^{-1}. Some authors use the notation $f[A]$ and $f^{-1}[B]$ to denote images and preimages, and write $f^{-1}[b]$ for $f^{-1}(\{b\})$, which has some notational advantage.

II.3.2.10. If $f : X \to Y$ is a function, then $f(\emptyset) = \emptyset$, and if $A \subseteq X$ and $f(A) = \emptyset$, then $A = \emptyset$. Preimages are more complicated: $f^{-1}(\emptyset) = \emptyset$, but if $B \subseteq Y$, then $f^{-1}(B) = \emptyset$ does not imply that $B = \emptyset$ – it only means the intersection of B with the range of f is empty.
We have $f^{-1}(Y) = X$, but not necessarily $f(X) = Y$: $f(X)$ is the range of f. If $A \subseteq X$ and $B \subseteq Y$, we have $A \subseteq f^{-1}(f(A))$ and $f(f^{-1}(B)) \subseteq B$, but we do not have equality in general.

Preimages respect unions, intersections, and complements, but images only respect unions in general (cf. II.3.4.3.):

II.3.2.11. PROPOSITION. Let $f : X \to Y$ be a function.

(i) If $\{A_i : i \in I\}$ is a collection of subsets of X, then $f(\cup_{i\in I} A_i) = \cup_{i\in I} f(A_i)$ and $f(\cap_{i\in I} A_i) \subseteq \cap_{i\in I} f(A_i)$.

(ii) If $\{B_i : i \in I\}$ is a collection of subsets of Y, then $f^{-1}(\cup_{i\in I} B_i) = \cup_{i\in I} f^{-1}(B_i)$ and $f^{-1}(\cap_{i\in I} B_i) = \cap_{i\in I} f^{-1}(B_i)$.

(iii) If $B \subseteq Y$, then $f^{-1}(B^c) = [f^{-1}(B)]^c$ (i.e. $f^{-1}(Y \setminus B) = X \setminus f^{-1}(B)$).

PROOF: We prove as a sample the second statement of (ii). The other proofs are very similar and are left as an exercise. Note that in all cases except the second statement of (i), two containments must be shown.

Let $x \in f^{-1}(\cap_{i\in I} B_i)$. Then $f(x) \in \cap_{i\in I} B_i$, so for each $i \in I$, $f(x) \in B_i$ and so $x \in f^{-1}(B_i)$. Thus $x \in \cap_{i\in I} f^{-1}(B_i)$.

For the reverse containment, let $x \in \cap_{i\in I} f^{-1}(B_i)$. Then, for each $i \in I$, $x \in f^{-1}(B_i)$, so $f(x) \in B_i$. Thus $f(x) \in \cap_{i\in I} B_i$, so $x \in f^{-1}(\cap_{i\in I} B_i)$. 👍

Injective and Surjective Functions

II.3.2.12. A. SLOMSON points out that "one-to-one" is potentially misleading terminology, since every function is arguably "one-to-one" (i.e. single-valued), and proposes that a better term would be "two-to-two." See [**?**].

Is the Codomain Uniquely Determined?

II.3.2.13. There is a subtle problem with codomains of functions which is overlooked in many texts, which sometimes even make contradictory statements as a result. If we have a function $f : X \to Y$, by our convention (II.3.2.3.) X is the domain of f and is thus uniquely determined by f. The Y is called the codomain of f, but it is only a set containing the range of f in general. Should we regard the codomain as being intrinsic to f, i.e. uniquely specified by f? This is a trickier question than would first appear: no matter which convention we take, it will be impossible to cleanly answer both of the following questions:

(i) When are two functions equal?

(ii) When is a function surjective?

II.3.2.14. It is stated in many books that two functions $f : X \to Y$ and $g : Z \to W$ are equal if and only if $X = Z$, $Y = W$, and $f(x) = g(x)$ for all $x \in X = Z$. But suppose $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^2$, and $g : \mathbb{R} \to [0, +\infty)$ is defined by $g(x) = x^2$. Are f and g equal functions, or are they different functions? (And does "equal functions" mean "the same function"?)

On the other hand, if f and g are the same function, does it really make sense to say that g is surjective and f is not surjective? How can a function be both surjective and not surjective?

II.3.2.15. If a function is defined to be a set of ordered pairs, as is the usual definition in set theory, then two functions with the same domain taking the same values at all points of the domain, such as the f and g, consist of exactly the same ordered pairs, and then must be equal as sets by the Extension Axiom of set theory, whether or not the specified codomains are the same.

One can proceed differently: in Category Theory, every function (morphism) has a specified domain and codomain which are part of the definition of a morphism. Thus from a category theory point of view the

above f and g are different. This point of view gives a clean meaning to surjectivity (technically called an *epimorphism*, although the abstract category definition must be done differently, cf. II.11.2.7.).

II.3.2.16. Since we have used the ordered pair definition of function, we ought to, and will, adopt the convention that the codomain of a function is not intrinsic to the function, and thus two functions are equal if they have the same domain and take the same values at all points of the domain; thus we will regard the above f and g to be the same function. This is the most common convention in Analysis, and is a useful one: the codomain can be taken to just be any convenient target set containing the range. For example, the most common types of functions considered in basic analysis are real-valued functions, and it is most convenient to simply take the codomain to be $\mathbb{R}$ in all cases. The reader should be warned, however, that surjectivity of a function depends crucially on what codomain is specified.

II.3.2.17. The same logical question arises with more general relations: if $\mathcal{R}$ is a relation from X to Y, are the X and Y intrinsic to $\mathcal{R}$? In this case we do not conventionally assume even that X is the domain of the relation, so using the ordered pair definition we should not consider either X or Y to be intrinsic to $\mathcal{R}$. However, this problem will rarely if ever arise in this book.

Composition of Functions

Sequences

Indexed Collections of Sets

General Cartesian Products

II.3.3. Equivalence Relations

The second important class of relations are equivalence relations. There are three standard points of view for these relations which can, and will, be used interchangeably.

II.3.3.1. Definition. A relation on a set is an *equivalence relation* if it is reflexive, symmetric, and transitive.

It is common to use the symbol $\sim$ for an equivalence relation, although other similar symbols are sometimes used. Thus a relation $\sim$ on a set X is an equivalence relation if it has the following properties:

(i) $x \sim x$ for all $x \in X$.

(ii) If $x, y \in X$ and $x \sim y$, then $y \sim x$.

(iii) If $x, y, z \in X$, $x \sim y$, and $y \sim z$, then $x \sim z$.

Note that by (i), the domain and range of an equivalence relation on X are all of X.

II.3.3.2. Examples. (i) If X is any set, the identity relation (II.3.1.2.(iv)) is a (not very interesting) equivalence relation on X. The full relation (II.3.1.2.(iii)) on X is another equivalence relation (also not very interesting). The identity relation and the full relation are the "smallest" and "largest" equivalence relations on X.

(ii) On $\mathbb{R}^2$, set $(x_1, y_1) \sim (x_2, y_2)$ if and only if $x_1 = x_2$. Then $\sim$ is an equivalence relation on $\mathbb{R}^2$.

(iii) Fix $n \in \mathbb{N}$. On $\mathbb{Z}$, set $x \sim y$ if and only if $x - y$ is divisible by n. Then $\sim$ is an equivalence relation on $\mathbb{Z}$, called *congruence modulo* n. We usually write $x \equiv y \bmod n$ if $x \sim y$ for this relation.

(iv) For a generic example including (i) and (ii) (and, in slight disguise, (iii)), let X and Y be sets, and $f : X \to Y$ a function. If $x_1, x_2 \in X$, set $x_1 \sim_f x_2$ if and only if $f(x_1) = f(x_2)$. Then $\sim_f$ is an equivalence relation on X. We will see that every equivalence relation on X is of this form for some Y and f. By replacing Y by the range of f, there is no loss of generality in assuming that f is surjective, in which case the Y and f are uniquely determined (up to bijective equivalence) by the relation $\sim_f$ (II.3.3.9.).

Equivalence Classes

The crucial notion connecting equivalence relations with the other points of view is equivalence classes:

II.3.3.3. Definition. Let $\sim$ be an equivalence relation on a set X, and $x \in X$. Then the *equivalence class* (or $\sim$*-equivalence class*) of x is

$$[x] = \{y \in X : x \sim y\} \subseteq X \ .$$

Note that $x \in [x]$ since $\sim$ is reflexive.

II.3.3.4. Proposition. If $\sim$ is an equivalence relation on a set X, and $x, y \in X$, then either $[x] = [y]$ or $[x]$ and $[y]$ are disjoint. We have $[x] = [y]$ if and only if $x \sim y$.

Proof: Suppose $x \sim y$. If $z \in [y]$, i.e. $y \sim z$, then $x \sim z$ by transitivity, so $z \in [x]$ and $[y] \subseteq [x]$. Since also $y \sim x$ by symmetry, by interchanging x and y in the argument we obtain $[x] = [y]$. Now suppose $[x]$ and $[y]$ are not disjoint, and let $z \in [x] \cap [y]$. Then $x \sim z$ and $z \sim y$, so $x \sim y$ and $[x] = [y]$.

II.3.3.5. Thus the set X partitions into a disjoint union of equivalence classes. The equivalence classes completely determine the equivalence relation: two elements of X are equivalent if and only if they are in the same equivalence class.

II.3.3.6. Examples. (i) If $\sim$ is the identity relation on X, then each equivalence class is a singleton (and conversely), i.e. $[x] = \{x\}$ for each $x \in X$. If $\sim$ is the full relation on X, there is only one equivalence class: $[x] = X$ for every $x \in X$.

(ii) In the equivalence relation of II.3.3.2.(ii), we have

$$[(x_1, y_1)] = \{(x_1, y) : y \in \mathbb{R}\}$$

so the equivalence classes are the vertical lines in $\mathbb{R}^2$.

(iii) In the equivalence relation of II.3.3.2.(iii), we have

$$[x] = \{x + kn : k \in \mathbb{Z}\} \ .$$

If $n = 2$, the equivalence class of any even integer is the set of all even integers, and the equivalence class of any odd integer is the set of all odd integers. For general n, there are exactly n equivalence classes: $[0]$, $[1]$, $\ldots$, $[n-1]$.

(iv) In the equivalence relation of II.3.3.2.(iv), we have $[x] = f^{-1}(f(x))$ for any $x \in X$, i.e. the equivalence classes are just the (nonempty) preimages of elements of Y (more correctly, singleton subsets of Y). Thus if f is surjective, there is a natural parametrization of the equivalence classes by elements of Y.

II.3.3.7. Conversely, suppose $\{X_i : i \in I\}$ is a partition of a set X, i.e. the X_i are disjoint and nonempty and $\cup_{i\in I} X_i = X$. Define a relation $\sim$ on X by setting $x \sim y$ if and only if x and y are in X_i for the same i. Then $\sim$ is an equivalence relation on X whose equivalence classes are precisely the X_i. Thus there is a natural one-one correspondence between equivalence relations on X and partitions of X. Viewing equivalence relations as partitions is the second standard point of view.

II.3.3.8. Suppose $\{X_i : i \in I\}$ is a partition of a set X. Define a function $f : X \to I$ by setting $f(x) = i$ if $x \in X_i$. Then f is a well-defined function from X to I, which is surjective since all the X_i are nonempty, and the equivalence relation defined by the partition is exactly $\sim_f$ as defined in II.3.3.2.(iv). Conversely, if g is a function from X onto a set Y, the equivalence classes of the relation $\sim_g$ are indexed by the elements of Y as in II.3.3.6.(iv). Thus the third point of view of equivalence relations on a set X is via surjective functions from X to other sets. The target set corresponding to an equivalence relation can be identified with the set of equivalence classes of the relation.

We summarize the three points of view:

II.3.3.9. THEOREM. Let X be a set. There is a natural one-one correspondence between

(i) Equivalence relations on X.

(ii) Partitions of X.

(iii) Surjective functions from X to other sets (up to bijective equivalence).

In (iii), "up to bijective equivalence" means: if $f : X \to Y$ and $g : X \to Z$ are surjective functions, then f and g are bijectively equivalent if there is a bijection $\phi : Y \to Z$ with $g = \phi \circ f$ (so $f = \phi^{-1} \circ g$), i.e. the functions are the "same" up to renaming the elements of the target set.

II.3.4. Exercises

II.3.4.1. Let $\mathcal{R}$ be a relation from X to Y, and $\mathcal{S}$ a relation from Y to Z. Show that $\mathrm{Dom}(\mathcal{S} \circ \mathcal{R}) = \mathcal{R}^{-1}(\mathrm{Range}(\mathcal{R}) \cap \mathrm{Dom}(\mathcal{S}))$ and $\mathrm{Range}(\mathcal{S} \circ \mathcal{R}) = \mathcal{S}(\mathrm{Range}(\mathcal{R}) \cap \mathrm{Dom}(\mathcal{S}))$.

II.3.4.2. Here is a common but fallacious argument that a relation which is symmetric and transitive must also be reflexive, hence an equivalence relation. "Let $\mathcal{R}$ be a relation on a set X which is symmetric and transitive, and let $x \in X$. Let y be any element of X with $x\mathcal{R}y$. Then $y\mathcal{R}x$ by symmetry, and hence $x\mathcal{R}x$ by transitivity. Thus $\mathcal{R}$ is reflexive."

(a) What is wrong with this argument?

(b) Give an example of a relation on a set which is symmetric and transitive, but not reflexive.

II.3.4.3. (a) What is wrong with the following argument that the image of an intersection under a function is the intersection of the images?

Let $f : X \to Y$ be a function, and $\{A_i : i \in I\}$ an indexed collection of subsets of X. Let $y \in \cap_{i\in I} f(A_i)$. Then, for each $i \in I$, $y \in f(A_i)$, so there is an $x \in A_i$ with $f(x) = y$. Then $x \in \cap_{i\in I} A_i$, so $y \in f(\cap_{i\in I} A_i)$.

(b) Give an example of an f and two subsets A_1, A_2 of X and a $y \in f(A_1) \cap f(A_2)$ for which $y \notin f(A_1 \cap A_2)$. [Consider a constant function.]

(c) Use your example in (b) to show that $f(A^c) \neq [f(A)]^c$ in general.

(d) Under what conditions on f can we assert that the image under f of an intersection is always the intersection of the images? Under what conditions is the image of a complement always the complement of the image?

II.4. Ordered Sets and Induction

II.4.1. Partially Ordered Sets

Partially ordered sets arise throughout mathematics. A particularly important use in analysis and topology is to use partially ordered index sets to describe convergence.

II.4.1.1. Definition. A transitive, antisymmetric relation on a set X is called a *partial order* on X. A *partially ordered set* is a set with a specified partial order. Elements x and y ($x \neq y$) in a partially ordered set ($X \precsim$) are *comparable* if $x \precsim y$ or $y \precsim x$; otherwise they are *incomparable*. A partial order $\precsim$ on a set X is a *total order* if any two elements of X are comparable, i.e. if, for every $x, y \in X$, at least one of $x \precsim y$, $x = y$, or $y \precsim x$ holds.

The symbols $\leq$, $<$, $\prec$, $\preceq$, $\precsim$, $\ll$, and their reversals ($\geq$, etc.) are typically used for partial orders, although other similar symbols can be used.

II.4.1.2. A partial order need not be reflexive; in fact, it can happen that $x \not\precsim x$ for any x, for example with $<$ on $\mathbb{R}$. Such a partial order is called a *strict partial order*. If $\precsim$ is a partial order on a set X, there is an associated strict partial order $\prec$ defined by $x \prec y$ if and only if $x \precsim y$ and $x \neq y$, and similarly there is an associated reflexive partial order $\preceq$ defined by $x \preceq y$ if and only if $x \precsim y$ or $x = y$. We will usually consider reflexive partial orders for definiteness; there is no essential loss of generality in doing so. We will sometimes (slightly carelessly) not distinguish between partial orders whose associated reflexive partial orders coincide (and hence whose associated strict partial orders also coincide); such partial orders will be called *equivalent*.

II.4.1.3. If Y is a subset of X, then the restriction to Y of any partial order on X is a partial order on Y, called the *induced ordering*.

II.4.1.4. Examples.

(i) $<$, $\leq$, $>$, $\geq$ are total orders on $\mathbb{R}$, or any subset of $\mathbb{R}$.

(ii) For any set X, $\subseteq$ and $\supseteq$ are reflexive partial orders on $\mathcal{P}(X)$. These are not total orders if X has more than one element: if $x \neq y$, then $\{x\}$ and $\{y\}$ are incomparable.

(iii) On $\mathbb{R}^2$ (or any subset), define $(x_1, y_1) \precsim (x_2, y_2)$ if and only if $x_1 \leq x_2$ and $y_1 \leq y_2$. This is a reflexive partial order which is not a total order [$(1, 0)$ and $(0, 1)$ are incomparable], called the *ordinary ordering* on $\mathbb{R}^2$.

(iv) On $\mathbb{R}^2$ (or any subset), define $(x_1, y_1) \precsim (x_2, y_2)$ if and only if $(x_1, y_1) = (x_2, y_2)$, or $x_1 < x_2$ and $y_1 < y_2$. This is a different reflexive partial order which is also not a total order [$(1, 0)$ and $(1, 1)$ are incomparable], called the *strict ordering* on $\mathbb{R}^2$ (the name is well established but somewhat unfortunate, since this is not a strict partial order.)

(v) On $\mathbb{R}^2$ (or any subset), define $(x_1, y_1) \precsim (x_2, y_2)$ if and only if $x_1 < x_2$, or $x_1 = x_2$ and and $y_1 \leq y_2$. This is a reflexive total order, called the *lexicographic ordering* on $\mathbb{R}^2$.

Examples (iii)-(v) can be generalized to arbitrary Cartesian products of partially ordered sets (for (v), the index set must be ordered; see Exercise ().) The term "lexicographic ordering" is used because this ordering is an analog of the alphabetical ordering of words, regarding the alphabet as a totally ordered set. An important generalization of (iii) is:

(vi) Let X be a set, and $\mathbb{R}^X$ the set of all functions from X to $\mathbb{R}$. If $f, g \in \mathbb{R}^X$, set $f \leq g$ if $f(x) \leq g(x)$ for all $x \in X$. This is the *ordinary ordering* on $\mathbb{R}^X$. It is not a total order if X has more than one element.

A more complicated example, which is typical of a class of examples arising in analysis, is:

(vii) Let X and Y be sets, and $\mathcal{F}(X, Y)$ the set of functions from subsets of X to Y. Write $f \precsim g$ if g extends f, i.e. the domain of g contains the domain of f and $g = f$ on the domain of f. Then $\precsim$ is a partial order on $\mathcal{F}(X, Y)$ which is not a total order if Y has more than one element. In fact, in general, "almost all" pairs of functions are incomparable.

There is a lot of terminology in the theory of partially ordered sets which we need to introduce, mostly related to upper and lower bounds of subsets.

II.4.1.5. Definition. Let $(X, \precsim)$ be a partially ordered set, and $A \subseteq X$. An element $x \in X$ is an *upper bound* for A if $a \precsim x$ for every $a \in A$. The element x is the *supremum*, or *least upper bound*, of A, written $x = \sup(A)$, if x is an upper bound for A and $x \precsim y$ for every upper bound y of A. Lower bounds and greatest lower bounds are defined similarly; the greatest lower bound of A is also called the *infimum* of A, written $\inf(A)$. A is *bounded above* if it has an upper bound, and *bounded below* if it has a lower bound; A is *bounded* if it is both bounded above and bounded below.

II.4.1.6. It is obvious from antisymmetry that the supremum or infimum of a set, if it exists, is unique. Note that an upper or lower bound for a subset A is not a member of A in general. For example, if $X = \mathbb{R}$ and $A = (0, 1)$, then any $x \geq 1$ is an upper bound for A, and $\sup(A) = 1$; similarly, any negative number is a lower bound for A and $\inf(A) = 0$.

In a general partially ordered set, a subset which is bounded above need not have a supremum; examples are given in the exercises.

II.4.1.7. Definition. Let $(X, \precsim)$ be a partially ordered set. Then $(X, \precsim)$ is *complete* if every bounded subset of X has a supremum and infimum.

II.4.1.8. One of the most important properties of $\mathbb{R}$ is the *Completeness Axiom*:

Completeness Axiom. $(\mathbb{R}, \leq)$ is complete.

There are many other properties of $\mathbb{R}$ which are equivalent to the completeness axiom, e.g. (), (), (), ... The completeness axiom is the main feature which distinguishes $\mathbb{R}$ from other number systems such as $\mathbb{Q}$. See Exercise ().

II.4.1.9. There are other important complete partial orders. Among the examples of II.4.1.4., (i), (ii), (iii), (vi), and (vii) are complete (in (ii), the supremum of a collection of subsets is their union; see Exercise () for (vii)), while (iv) and (v) are not complete (Exercises ().)

II.4.1.10. There are two finite versions of completeness which are are important. A partially ordered set $(X, \precsim)$ is *directed* (technically, *upward directed*) if every pair of elements has an upper bound, and is a *lattice* if every pair of elements has a supremum and an infimum. The supremum and infimum of x and y are usually denoted $x \vee y$ and $x \wedge y$ respectively. Every totally ordered set is a lattice, and a lattice-ordered set is directed. Of the examples, (i), (ii), (iii), (iv), and (vi) are lattices, (iv) is directed but not a lattice, and (vii) is not directed (see Exercises ().) In a directed set [lattice], every finite subset has an upper bound [supremum and infimum] (Exercise ().) See II.4.4.1. for a generalization and discussion.

II.4.1.11. Definition. Let $(X, \precsim)$ be a partially ordered set. An element $x \in X$ is a *maximal element* if $y \in X$, $x \precsim y$ implies $y = x$. The element $x \in X$ is the *largest element* if $y \precsim x$ for all $y \in X$. Minimal elements and smallest elements are defined similarly.

II.4.1.12. A largest element, if it exists, is obviously unique, and is a maximal element. In a totally ordered set, the two things are the same; but in a general partially ordered set a maximal element need not be a largest element, and there can be more than one maximal element. For example, in $\mathcal{F}(X, Y)$ (II.4.1.4.(vii)), a function is a maximal element if and only if its domain is all of X, and there is no largest element (if Y has more than one element.) The situation is the same for smallest or minimal elements.

There is also a notion of morphism for partially ordered sets:

II.4.1.13. Definition. Let $(X, \precsim)$ and $(Y, \leq)$ be partially ordered sets. A function $f : X \to Y$ is *order-preserving* if $x_1 \precsim x_2$ implies $f(x_1) \leq f(x_2)$. The function f is an *order-isomorphism* if it is a bijection and both f and f^{-1} are order-preserving, i.e. $x_1 \precsim x_2$ if and only if $f(x_1) \leq f(x_2)$.

II.4.1.14. **Caution:** an order-isomorphism is more than an order-preserving bijection: for example, the identity map from $\mathbb{R}^2$ with the strict ordering () to $\mathbb{R}^2$ with the ordinary ordering () is an order-preserving bijection, but not an order-isomorphism.

II.4.1.15. A composition of order-preserving functions is order-preserving, so the collection of partially ordered sets and order-preserving functions forms a category.

II.4.1.16. We also slightly carelessly say that $(X, \precsim)$ and $(Y, \leq)$ are order-isomorphic if there is a bijection from X to Y which is an order-isomorphism between $\precsim$ and a partial order on Y equivalent to $\leq$. Thus, for example, we say that $(\mathbb{R}, \leq)$ and $(\mathbb{R}, <)$ are order-isomorphic.

II.4.2. Well-Ordered Sets and Transfinite Induction

There is a very special type of ordered set which is important in some applications:

II.4.2.1. Definition. A partially ordered set $(I, \precsim)$ is *well-ordered* if every nonempty subset has a smallest element.

We will generally use I and J as names of well-ordered sets since well-ordered sets are primarily used in analysis as index sets.

A well-ordered set I is totally ordered (if $i, j \in I$, then $\{i, j\}$ has a smallest element), and I itself has a smallest element. Every subset of a well-ordered set is well-ordered in the induced ordering.

The ordering on a well-ordered set is called a *well ordering*. While the term "well-ordered" is grammatically correct English, the term "well ordering" is an abuse of language (the word "well" used as an adjective is the opposite of "sick"), but the term is unfortunately "well established." (The terms come from German, where these grammatical problems do not exist.)

II.4.2.2. Examples.

(i) Any finite totally ordered set is well-ordered.

(ii) $\mathbb{N}$ is well-ordered.

(iii) $\mathbb{N}^2$ with lexicographic ordering is well-ordered.

(iv) $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ are not well-ordered since they have no smallest element. $[0, \infty)$ is not well-ordered either: it has a smallest element, but the subset $(0, \infty)$ does not.

II.4.2.3. Actually, well-ordered sets are quite unusual. While there are many of them, they all look somewhat the same except for length: if I and J are well-ordered, then one of them is order-isomorphic to an initial segment of the other, in a unique way (II.4.2.19.). In addition, any subset of a well-ordered set is well ordered in the induced ordering; in particular, a tail of a well-ordered set with an initial segment deleted is well ordered. Thus well-ordered sets all look like the picture in Figure II.1 except for length (the vertical segments represent points and are drawn this way only for visual clarity). Note that the segment to the right starting at any fixed element looks exactly like the whole set except (possibly) for length.

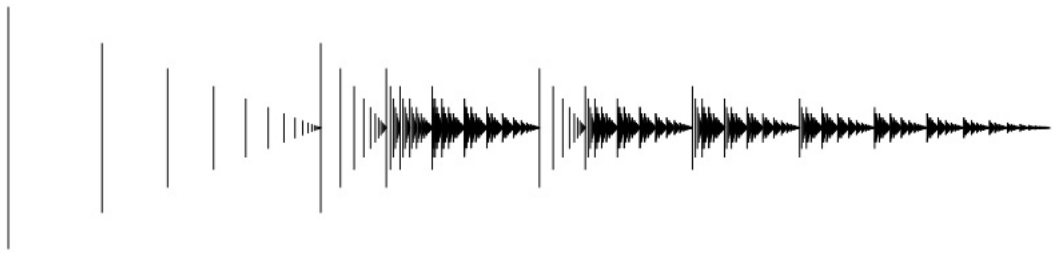


Figure II.1: A well-ordered set

Well-ordered sets are described in more detail in II.8.4..

II.4.2.4. Every element of a well-ordered set, except the largest element if there is one, has an immediate successor: for any i, the set $\{j : i \prec j\}$, if nonempty, has a smallest element. An element need not have an immediate predecessor, however; for example, in $\mathbb{N}^2$ with lexicographic order, $(2, 1)$ has no immediate predecessor. An element with no immediate predecessor is called a *limit element*.

II.4.2.5. A well ordering is a complete ordering: if $(I, \precsim)$ is well-ordered, and $A \subseteq I$ is nonempty, then A has a smallest element which is $\inf(A)$; and the set of upper bounds for A, if nonempty, has a smallest element which is $\sup(A)$.

Transfinite Induction

A proof procedure generalizing ordinary induction can be based on any well-ordered set.

II.4.2.6. Theorem. [Principle of Transfinite Induction] Let $(I, \leq)$ be a well-ordered set, and suppose $P(i)$ is a statement for each $i \in I$. Suppose, for each $i \in I$, the following *inductive implication* holds:

"If $P(j)$ is true for all $j < i$, then $P(i)$ is true."

Then $P(i)$ is true for all $i \in I$.

We have not been precise about the meaning of the term "statement", since to be so would require an extensive excursion into set theory and mathematical logic. Many books on mathematical logic such as [**?**] contain a careful discussion of this point. But any type of mathematical statements likely to be made by a working mathematician come within the scope of the principle.

II.4.2.7. We outline the proof (although we cannot be completely precise because of the imprecision in the term "statement".) Let J be the set of $j \in I$ such that $P(j)$ is true. If $J \neq I$, since I is well-ordered there is a smallest element $i \in I \setminus J$. But then $P(j)$ is true for all $j < i$, and therefore $P(i)$ is true by the inductive implication; this is a contradiction, so $I = J$.

II.4.2.8. The term "transfinite induction" is usually used in the case where I properly contains an initial segment order-isomorphic to $\mathbb{N}$, and especially in the case where I is uncountable.

One of the most frequently used consequences of the principle of transfinite induction is:

II.4.2.9. Corollary. Let $(I, \leq)$ be a well-ordered set, and $J \subseteq I$. If J has the property that, for every $i \in I$, $[j \in J \text{ for all } j < i]$ implies $i \in J$, then $J = I$.

Proof: Let $P(i)$ be the statement "$i \in J$", and apply II.4.2.6.. 👍

II.4.2.10. Transfinite induction in the special case $I = \mathbb{N}$ is usually called *complete induction*. Complete induction is not quite the same as ordinary induction, where statements $P(n)$ indexed by $\mathbb{N}$ are proved for all n by proving $P(1)$ and $[P(n) \Rightarrow P(n+1)]$ for all n. Instead, in complete induction, we assume $P(k)$ is true for all $k \in \mathbb{N}$, $k < n$, and deduce $P(n)$. Ordinary induction is a special case of complete induction; on the other hand, complete induction can be easily deduced from ordinary induction (Exercise ()).

Complete induction is useful where there is no simple relationship between $P(n)$ and $P(n+1)$, but there is a relationship between $P(n)$ and $P(k)$ for certain k's less than n. The next simple example is a good illustration.

II.4.2.11. PROPOSITION. Every natural number can be written as a finite product of prime numbers. [We define the product of an empty set of primes to be 1, and a product of one prime p to be p.]

PROOF: For each $n \in \mathbb{N}$, let $P(n)$ be the statement "The integer n can be written as a finite product of prime numbers." Fix $n \in \mathbb{N}$, and assume $P(k)$ is true for all $k < n$; we must prove that $P(n)$ is true. There are three cases:
(1) $n = 1$. Then n is an empty product of prime numbers.
(2) $n > 1$ is prime. Then n is a product of one prime number.
(3) $n > 1$ is not prime. Then $n = n_1 n_2$ for some $n_1, n_2 \in \mathbb{N}$, $n_1, n_2 > 1$. Then $n_1, n_2 < n$, so by assumption n_1 and n_2 can each be written as a finite product of prime numbers. Hence so can $n_1 n_2 = n$. 👍

II.4.2.12. Note that in complete induction, it is not necessary in general to separately prove $P(1)$, since it is included in the general case. In the case $n = 1$, the assumption that $P(k)$ is true for all $k < n$ is automatically satisfied since there are no such k's (in this case, the hypothesis is said to be *vacuous*.) Of course, it is *allowed* to prove $P(1)$ as a special case, as was done in the proof of II.4.2.11..

Definitions by Transfinite Induction or Recursion

Transfinite induction (or, technically, *transfinite recursion*) can also be used to define functions whose domain is a well-ordered set. We will again not be completely precise in defining terms; see, for example, [**?**, p. 249-250] for a careful treatment.

II.4.2.13. Let $(I, \leq)$ be a well-ordered set, and X any set. Suppose, for each $i \in I$, we have a definite procedure or rule R_i for choosing an $f(i) = x_i \in X$, which may depend on the choice of the $f(j)$ for $j < i$. Then

II.4.2.14. THEOREM. There is a unique $f : I \to X$ such that, for each $i \in I$, $f(i)$ is determined by the specified procedure R_i applied to $\{f(j) : j < i\}$.

Roughly speaking, if f can be defined "one step at a time" by the rules R_i, then f can be defined "all at once."

II.4.2.15. We outline an argument "proving" the theorem, overlooking the imprecision in the term "procedure". We first prove uniqueness of f by transfinite induction: if f and g have $f(i)$ determined for each i by applying R_i to $\{f(j) : j < i\}$, and $g(i)$ determined by applying R_i to $\{g(j) : j < i\}$, let J be the set of i for which $f(i) = g(i)$. If $i \in I$ and $j \in J$ (i.e. $f(j) = g(j)$) for all $j < i$, then applying R_i we obtain $f(i) = g(i)$, so $i \in J$. Thus $J = I$ by II.4.2.9..

For existence, let J be the set of all i such that there is a function f_i on $\{k : k \leq i\}$ with $f_i(k)$ determined by R_k applied to $\{f_i(j) : j < k\}$ for each $k \leq i$. If $i_1, i_2 \in J$ and $k \leq i = \min(i_1, i_2)$, then $f_{i_1}(k) = f_{i_2}(k)$ by the uniqueness of the first part of the proof, applied to $\{k : k \leq i\}$. Thus, if $i \in I$ and $j \in J$ for all $j < i$, there is a (unique) function f on $\{j : j < i\}$ with $f(j)$ determined by R_j applied to $\{f(k) : k < j\}$ for all $j < i$. Apply R_i to $\{f(j) : j < i\}$ to define $f(i)$ and extend f to $\{j : j \leq i\}$. Thus $i \in J$. By II.4.2.9., $J = I$.

This "proof" is actually fallacious; see Exercise (). The actual proof (cf. []) is an adaptation of the proof of the following special case, which is the most important one for applications.

II.4.2.16. THEOREM. Let X be a set, and for each $n \in \mathbb{N}$ let f_n be a function from X^n to X. If x_1 is any element of X, then there is a unique sequence (x_n) in X with $x_{n+1} = f_n(x_1, \ldots, x_n)$ for all $n \in \mathbb{N}$.

PROOF: Let S be the smallest subset of $\mathbb{N} \times X$ containing $(1, x_1)$, with the property that if, for any $n \in \mathbb{N}$, whenever $\{(k, y_k) : k \in \mathbb{N}, k \leq n\} \subseteq S$ for $(y_1, \ldots, y_n) \in X^n$, then $(n+1, f_n(y_1, \ldots, y_n)) \in S$. ($S$ is the intersection of all subsets of $\mathbb{N} \times X$ with these two properties; there is at least one such subset, $\mathbb{N} \times X$ itself.) Then let

$$A = \{n \in \mathbb{N} : (k, x_k) \in S \text{ for exactly one } x_k \in X \text{ for all } k \in \mathbb{N}, k \leq n\} .$$

First we show $1 \in A$. Set

$$S' = \{(1, x_1)\} \cup \{(n, x) \in S : n > 1\}$$

and note that $S' \subseteq S$, $(1, x_1) \in S'$, and $\{(k, y_k) : k \in \mathbb{N}, k \leq n\} \subseteq S'$ implies $(n+1, f_n(y_1, \ldots, y_n)) \in S'$; hence $S' = S$ and $1 \in A$. Now suppose $n \in A$, i.e. there is a unique $x_k \in X$ with $(k, x_k) \in S$ for all $k \in \mathbb{N}$, $k \leq n$. Set

$$S'' = \{(k, x_k) : k \in \mathbb{N}, k \leq n\} \cup \{(n+1, f_n(x_1, \ldots, x_n))\} \cup \{(m, x) \in S : m > n+1\}$$

and note that $S'' \subseteq S$, $(1, x_1) \in S''$, and $\{(k, y_k) : k \in \mathbb{N}, k \leq m\} \subseteq S''$ implies $(m+1, f_m(y_1, \ldots, y_m)) \in S''$. Thus $S'' = S$ and $n+1 \in A$. So $A = \mathbb{N}$.

For each $n \in \mathbb{N}$, let x_n be the unique element of X for which $(n, x_n) \in S$. Let $B = \{n \in \mathbb{N} : x_{n+1} = f_n(x_1, \ldots, x_n)\}$. We have $1 \in B$ since $(2, f_1(x_1)) \in S$; and if $n \in B$, then $(n+2, f_{n+1}(x_1, \ldots, x_{n+1})) \in S$, so $n+1 \in B$. Thus $B = \mathbb{N}$.

Note that no form of the Axiom of Choice is needed in this proof; in fact, the AC is not needed even for the transfinite result II.4.2.14..

II.4.2.17. In other words, if a definite rule is specified for each n for determining the $(n+1)$'st term of the sequence from the first n terms, there is a unique sequence generated beginning with any first term. The rule f_n can vary with n, but all the f_n must be specified in advance. The sequence is said to be *defined by induction* (or, more correctly, *defined by recursion*) from x_1 by the functions (f_n).

II.4.2.18. **Caution:** In definitions by transfinite induction, the rule R_i can be quite general, but it must be a definite rule giving a specific element as a function of $\{f(j) : j < i\}$. Thus, for example, a "rule" like "Let $f(i)$ be any element of $X \setminus \{f(j) : j < i\}$" is not permissible, even if it is proved that this set is always nonempty (say, by a cardinality argument.) But if X is well-ordered, it is permissible to let R_i be the rule "Let $f(i)$ be the *first* element of $X \setminus \{f(j) : j < i\}$" (provided this set is always nonempty.) In fact, this rule can be used to give a proof of the "uniqueness up to length" of well orderings:

II.4.2.19. THEOREM. Let $(I, \precsim)$ and $(J, \leq)$ be well-ordered sets. Then there is a unique order-isomorphism from one of I, J onto an initial segment of the other.

PROOF: The idea is that we pair up the first element of I with the first element of J, and each succeeding element of I with the next remaining element of J, until we "run out of elements" in either I or J. This is clearly the only possible way to obtain an order-isomorphism of the desired type, giving uniqueness.

More precisely, we try to define $f : I \to J$ as follows. For each $i \in I$ let $f(i)$ be the first element of $J \setminus \{f(k) : k \prec i\}$. This inductively defines f on the set $A = \{i \in I : Y \setminus \{f(k) : k \prec i\} \neq \emptyset\}$. This set A is either all of I or is of the form $\{k \in I : k \prec i\}$ for some $i \in I$, which is a proper initial segment of I; and f is an order-isomorphism from A onto $f(A)$. In the case $A \neq I$, f maps A bijectively to J. If $A = I$, then either f maps I onto J, or there is a smallest element j of $J \setminus f(I)$, in which case $f(I) = \{k \in J : k < j\}$, which is an initial segment of J, since we can never have $j \leq f(i)$ for any $i \in I$ by the way f is defined. 👍

An alternate proof not using transfinite induction is outlined in Exercises ().

II.4.3. Exercises

II.4.3.1. Consider the partially ordered set $(\mathbb{R}^2, \precsim)$ of II.4.1.4.(iv) (strict ordering).

(a) Show that $(1, 0)$, $(0, 1)$, and $(1, 1)$ are mutually incomparable.

(b) Show that $\{(1, 0), (0, 1)\}$ has no supremum or infimum.

(c) Show that $\{(x_i, y_i) : i \in I\}$ is bounded in $(\mathbb{R}^2, \precsim)$ if and only if $\{x_i\}$ and $\{y_i\}$ are bounded in $\mathbb{R}$. Conclude that $(\mathbb{R}^2, \precsim)$ is directed.

(d) Define an analog of the strict ordering on $\mathbb{R}^X$ for a nonempty set X, and prove that $(\mathbb{R}^X, \precsim)$ is directed but not a lattice (unless X is a singleton.)

II.4.3.2. Consider the partially ordered set $(\mathbb{R}^2, \precsim)$ of II.4.1.4.(v) (lexicographic ordering).

(a) Show that a subset $\{(x_i, y_i) : i \in I\}$ is bounded in $(\mathbb{R}^2, \precsim)$ if and only if $\{x_i\}$ is bounded in $\mathbb{R}$.

(b) Show that $\{(x, y) : 0 < x < 1\}$ and $\{(x, y) : 0 \leq x \leq 1\}$ are bounded but have no supremum or infimum.

II.4.3.3. Consider the partially ordered set $\mathcal{F}(X, Y)$ of II.4.1.4.(vii).

(a) Show that if f and g have a common domain point at which they disagree, then $\{f, g\}$ has no upper bound. Hence $\mathcal{F}(X, Y)$ is not (upward) directed.

(b) If $A = \{f_i : i \in I\}$ is a nonempty subset of $\mathcal{F}(X, Y)$ which is bounded above by a function g, let h be the restriction of g to the union of the domains of the f_i. Then $h = \sup(A)$.

(c) Every subset of $\mathcal{F}(X, Y)$ is bounded below; in fact, $\mathcal{F}(X, Y)$ has a smallest element, the function whose domain is $\emptyset$.

(d) If $A = \{f_i : i \in I\}$ is a nonempty subset of $\mathcal{F}(X, Y)$, set

$$Z = \{x \in X : x \in Dom(f_i) \text{ for all } i \in I,\ f_i(x) = f_j(x) \text{ for all } i, j \in I\}$$

and let g be the restriction of any f_i to Z. Then $g = \inf(A)$. Thus $\mathcal{F}(X, Y)$ is complete.

II.4.3.4. Suppose $\{X_i : i \in I\}$ is an indexed collection of partially ordered sets, and I also has a well ordering. We will put a lexicographic ordering on $\prod_{i\in I} X_i$. Use the symbol $\precsim$ for all the partial orders on I and the X_i's, and write $x \prec y$ if $x \precsim y$ and $x \neq y$.

(a) If $x = (x_i)$ and $y = (y_i)$ are elements of $\prod X_i$, and $x \neq y$, then $\{i \in I : x_i \neq y_i\}$ is nonempty, hence has a smallest element i_0. Let $x \leq y$ if $x_{i_0} \precsim y_{i_0}$ Show that $\leq$ is a partial order on $\prod X_i$.

(b) $\prod X_i$ is totally ordered if and only if each X_i is totally ordered.

(c) If each X_i is well-ordered, is $\prod X_i$ well-ordered?

(d) Discuss the difficulties in trying to define a lexicographic ordering if I is not well-ordered.

II.4.3.5. Explain why ordinary induction is a special case of complete induction.

The next three problems give an alternate proof of Theorem II.4.2.19. not using transfinite induction.

II.4.3.6. Let $(I, \leq)$ be a well-ordered set, and $f : I \to I$ an injective order-preserving function. Prove directly from the definition of a well-ordered set that $f(i) \geq i$ for all $i \in I$. [If not, let i_0 be the smallest i for which $f(i) < i$, and set $j = f(i_0)$. Show that $f(j) < j$.]

II.4.3.7. (a) Let $(I, \leq)$ be a well-ordered set. Then the only order-isomorphism from I onto an initial segment of I is the identity map on I. [If f is such an order-isomorphism, then $f(i) \geq i$ for all i by II.8.4.2.. Let i_0 be the smallest i with $f(i) > i$, and let $j \in I$ with $f(j) = i_0$. Show $j < i_0$ and $f(j) > j$.]

(b) Let $(I, \precsim)$ and $(J, \leq)$ be well-ordered sets. Then there is at most one order-isomorphism from I to J.

II.4.3.8. Let $(I, \precsim)$ and $(J, \leq)$ be well-ordered sets. If $u \in I$, write $I_u = \{i \in I : i < u\}$. Also, if X and Y are partially ordered sets, write $X \sim Y$ if X and Y are order-isomorphic. Define

$$\Gamma = \{(u, v) \in I \times J : I_u \sim J_v\}.$$

Show that Γ (or Γ^{-1}) is the graph of an order-isomorphism from an initial segment of I onto J (or an order-isomorphism from an initial segment of J onto I), as follows:

(a) Let

$$I_0 = \{u \in I : \exists v \in J \text{ with } (u, v) \in \Gamma\}, J_0 = \{v \in J : \exists u \in I \text{ with } (u, v) \in \Gamma\}$$

(i.e. I_0 and J_0 are the domain and range of the relation Γ.) Show that I_0 and J_0 are initial segments of I and J respectively. [If $(u_0, v_0) \in \Gamma$ and $f : I_{u_0} \to J_{v_0}$ is an order-isomorphism, and $u \leq u_0$, $v \leq v_0$, restrict f and f^{-1} to I_u and J_v.]

(b) Show that Γ is the graph of a one-to-one function ϕ from I_0 to J_0. [Suppose $u \in I$, $v_1, v_2 \in J$, $v_1 \leq v_2$, and $(u, v_1), (u, v_2) \in \Gamma$. Let $f : I_u \to J_{v_1}$ and $g : I_u \to J_{v_2}$ be order-isomorphisms. Then $f \circ g^{-1}$ is an order-isomorphism from the well-ordered set J_{v_2} onto its initial segment J_{v_1}, so $v_i = v_2$ (and $f = g$) by II.4.3.7..]

(c) Show that ϕ is an order-isomorphism from I_0 onto J_0. [Suppose (u_1, v_1) and (u_2, v_2) are in Γ and $u_1 < u_2$, and $f : I_{u_1} \to J_{v_1}$ and $g : I_{u_2} \to J_{v_2}$ are order-isomorphisms. If $v_1 > v_2$, then $h = g \circ f^{-1}$ is an injective order-preserving function from J_{v_1} to $J_{v_2} \subseteq J_{v_1}$ with $h(v_2) < v_2$, contradicting II.8.4.2..]

(d) Show that either $I_0 = I$ or $J_0 = J$. [If not, let u_0 be the smallest element of $I \setminus I_0$ and v_0 the smallest element of $J \setminus J_0$. Then $I_0 = I_{u_0}$, $J_0 = J_{v_0}$, and ϕ gives an order-isomorphism from I_{u_0} to J_{v_0}, so $(u_0, v_0) \in \Gamma$.]

(e) The function ϕ gives the *unique* order-isomorphism from I_0 to J_0 by II.4.3.7..

II.4.4. Directed Sets

There is another kind of "ordered" set which is important, especially in topology. In many instances, one only needs to have a notion of "sufficiently large" elements of an "ordered" set. This notion is made precise in a useful way in directed sets.

II.4.4.1. DEFINITION. Let X be a nonempty set. A *direction* on X is a transitive relation (preorder) $\leq$ with the property that any two elements of X have an upper bound, i.e. for any $a, b \in X$ there is a $c \in X$ with $a \leq c$ and $b \leq c$. A *directed set* is a pair $(X, \leq)$ where X is a nonempty set and $\leq$ is a direction on X. A direction $\leq$ on a set X is *proper* if for every $a \in X$ there is a $c \in X$ with $a \leq c$ and $c \not\leq a$.

II.4.4.2. Note that a direction on a set does not have to be a partial order, i.e. a direction does not have to be antisymmetric; however, many important directions are partial orders, and there is no essential loss of generality in requiring that the preorder be a partial order (II.4.4.11.), although there is a loss of utility in applications.

A direction is also often required to be reflexive, but this is unnecessary (and can be easily arranged anyway; see Exercise II.4.5.1.). In the definition, a and b do not have to be distinct, i.e. in a directed set X, for any $a \in X$ there is a $c \in X$ with $a \leq c$.

A directed set in the sense of II.4.4.1. is sometimes called *upward directed*. Downward directed sets can also be defined, but we will not work with them in this book.

II.4.4.3. EXAMPLES. (i) Any totally ordered set is a directed set (provided the "largest" element a, if there is one, satisfies $a \leq a$). The ordering is a proper direction if and only if there is no largest element. In particular, $\mathbb{N}$ is properly directed in the usual order.

(ii) More generally, any lattice is a directed set. In particular, the power set of any set X is directed by inclusion, as is the collection of finite subsets of X; the latter is a properly directed set if X is infinite.

(iii) Any partially ordered set with a largest element is directed but not properly directed. These sets are not very interesting as directed sets; the interesting directed sets are ones with no maximal elements, since the most important feature of a directed set is its "tails." A partially ordered and directed set is properly directed if and only if it has no maximal elements.

(iv) A set with the full relation ($x \leq y$ for all x and y) is a directed set which is not partially ordered.

(v) The equality relation on a set with more than one element is a partial order which is not a direction.

(vi) The partially ordered set of () is not directed.

(vii) Here is a very important example from topology. Let X be a topological space and $p \in X$. Let $\mathfrak{N}_p$ be the set of all neighborhoods of p. If $U, V \in \mathfrak{N}_p$, let $V \leq U$ if $U \subseteq V$. Then $\mathfrak{N}_p$ is a directed set with this ordering. $\mathfrak{N}_p$ can be replaced here by the set of open neighborhoods of p, or by any local base for the topology at p. There are other variations, such as setting $V \leq U$ if the closure of U is contained in V (if X is a regular space).

(viii) As a variation of (vii), let X be a set and $\mathcal{F}$ be a filter () of subsets of X. If $A, B \in \mathcal{F}$, set $A \leq B$ if $B \subseteq A$. Then $(\mathcal{F}, \leq)$ is a directed set.

(ix) If $(X, \leq)$ is a directed set, a subset Y of X is *cofinal* in X if for every $x \in X$ there is a $y \in Y$ with $x \leq y$. A cofinal subset of a directed set is directed in the induced preorder.

II.4.4.4. PROPOSITION. Let $(X, \leq)$ be a directed set. Then any finite subset of X has an upper bound, i.e. for any finite subset $\{a_1, \ldots, a_n\}$ of X, there is a $c \in X$ with $a_k \leq c$ for all $1 \leq k \leq n$.

PROOF: The c (or, more accurately, a c that works – the c will not be unique in general) can be found iteratively: let c_2 be an upper bound for a_1 and a_2, c_3 an upper bound for c_2 and a_3, ..., c an upper bound for c_{n-1} and a_n. (This argument should really be done more carefully by induction.) 👍

II.4.4.5. PROPOSITION. Let $\leq$ be a direction on a set X. Then $\leq$ is proper if and only if there is no largest element (an element a is a *largest element* if $x \leq a$ for all $x \in X$).

PROOF: If there is a largest element, then $\leq$ is obviously not proper. Suppose there is no largest element. If $a \in X$, then there is a $b \in X$ such that $b \not\leq a$. Let c be an upper bound for $\{a, b\}$. Then $a \leq c$. If $c \leq a$, then since $b \leq c$ we get $b \leq a$ by transitivity, a contradiction. Thus $c \not\leq a$ and $\leq$ is proper. 👍

We also need the notion of a direction-preserving function between directed sets:

II.4.4.6. DEFINITION. Let $(X, \leq)$ and $(Y, \precsim)$ be directed sets. A function $f : X \to Y$ is *directed* if, for every $b \in Y$, there is an $a \in X$ such that $b \precsim f(x)$ for all $x \in X$ with $a \leq x$.

The condition in the definition is sometimes called the *cofinal condition*.

II.4.4.7. Note that for f to be directed, it is neither necessary nor sufficient that f be order-preserving. For example, the function from $\mathbb{N}$ to $\mathbb{N}$ which interchanges $2n-1$ and $2n$ for each n is directed but not order-preserving. In fact, any bijection from $\mathbb{N}$ to $\mathbb{N}$ is directed (II.4.5.2.). But a constant function from $\mathbb{N}$ to $\mathbb{N}$ is not directed, even though it is order-preserving. There are even injective order-preserving functions which are not directed: the inclusion of $\mathbb{N}$ into $\mathbb{N} \cup \{+\infty\}$ is an example. Directed functions are "almost order-preserving" (Exercise II.4.5.3.).

Directed functions preserve "tails" or "cofinality" and thus are the appropriate maps between directed sets. In particular, the inclusion of a subset Y of X into X is a directed function if and only if Y is cofinal (II.4.4.3.(ix)) in X.

The next result is easily verified:

II.4.4.8. PROPOSITION. A composition of directed functions is directed.

II.4.4.9. DEFINITION. Let $(X, \leq)$ and $(Y, \precsim)$ be directed sets.

A function $f : X \to Y$ is a *directed equivalence* if f is a bijection and f and f^{-1} are directed.

$(X, \leq)$ and $(Y, \precsim)$ are *directed equivalent* if there is a directed equivalence from X to Y.

Two directions $\leq$ and $\precsim$ on a set X are *equivalent* if the identity map from $(X, \leq)$ to $(X, \precsim)$ is a directed equivalence.

II.4.4.10. The directed sets and directed functions form a category (), and the directed equivalences are just the isomorphisms in this category.

A largest element in a partially ordered set, if there is one, is unique; but a largest element in a directed set is not necessarily unique (if X has the full relation, then every element of X is a largest element.) The directions equivalent to partially ordered directions are precisely the ones for which this does not happen:

II.4.4.11. THEOREM. Let $\leq$ be a direction on a set X. Then there is an equivalent direction $\leqslant$ on X which is a reflexive partial order if and only if there is at most one largest element for $\leq$. In particular, if $\leq$ is proper, there is an equivalent proper direction which is a reflexive partial order.

PROOF: Suppose b and c are largest elements of X, and suppose there is a direction $\leqslant$ equivalent to $\leq$ which is a partial order. Let $a \in X$ such that $a \leq x$ implies $b \leqslant x$. Applying this with $x = c$, we obtain $b \leqslant c$. A similar argument shows that $c \leqslant b$, so $b = c$ since $\leqslant$ is a partial order. Thus $(X, \leq)$ cannot have more than one largest element.

Suppose $\leq$ has exactly one largest element c. Define $\leqslant$ by $x \leqslant y$ if $x = y$ or $y = c$. Then $\leqslant$ is a direction on X which is a reflexive partial order. To show that $\leqslant$ is equivalent to $\leq$, we need that for every $b \in X$ there is an $a \in X$ such that $a \leq x$ implies $b \leqslant x$, and for any $a \in X$ there is a $b \in X$ such that $b \leqslant x$ implies $a \leq x$. Just take $a = c$ in the first case and $b = c$ in the second.

Now suppose $\leq$ is proper, i.e. there are no largest elements. Define $\leqslant$ by setting $x \leqslant y$ if $x = y$ or if $x \neq y$, $x \leq y$, and $y \not\leq x$. It is obvious that $\leqslant$ is reflexive and antisymmetric. To see that $\leqslant$ is transitive, suppose $x \leqslant y$ and $y \leqslant z$. If $x = z$, then $x \leqslant z$. Suppose $x \neq z$. If $x = y$ or if $y = z$, it is trivial that $x \leqslant z$, so we may assume x, y, z are all distinct. Then $x \leq y$ and $y \leq z$, so $x \leq z$. If $z \leq x$, then $z \leq y$ by transitivity; but $y \neq z$ and $y \leqslant z$, so $z \not\leq y$, a contradiction. Thus $z \not\leq x$ and $x \leqslant z$.

To show that $\leqslant$ is a direction, let $a, b \in X$. Take $c \in X$ with $a \leq c$ and $b \leq c$, and take $d \in X$ with $c \leq d$ and $d \not\leq c$. We have $d \neq a$ and $d \neq b$, and $a \leq d$ and $b \leq d$ by transitivity. If $d \leq a$, we would have $d \leq c$ by transitivity, a contradiction; thus $d \not\leq a$, i.e. $a \leqslant d$. Similarly $b \leqslant d$.

Finally, to show that $\leqslant$ is equivalent to $\leq$, let $b \in X$. We must find $a \in X$ such that $a \leqslant x$ implies $b \leq x$. Let $a \in X$ with $b \leq a$ and $a \not\leq b$. If $a \leqslant x$, then either $a = x$ or $a \leq x$; in either case $b \leq x$. Conversely, let $a \in X$. We must find $b \in X$ such that $b \leq x$ implies $a \leqslant x$. Take $b \in X$ with $a \leq b$ and $b \not\leq a$. If $x \in X$ and $b \leq x$, then $x \neq a$ and $a \leq x$ by transitivity. If $x \leq a$, then $b \leq a$ by transitivity, a contradiction, so $x \not\leq a$ and $a \leqslant x$. 👍

Note that no form of Choice is needed for this result (rephrase the last two paragraphs as proofs by contradiction; a contradiction can be obtained by making only finitely many choices).

II.4.5. Exercises

II.4.5.1. Let $\precsim$ be a direction on a set X. Define $\leq$ on X by $a \leq b$ if $a \precsim b$ or $a = b$. Show that $\leq$ is a reflexive direction on X which is equivalent to $\precsim$.

II.4.5.2. (a) Show that a function $f : \mathbb{N} \to \mathbb{N}$ is directed (for the usual order) if and only if $f^{-1}(\{n\})$ is finite for every $n \in \mathbb{N}$.

(b) Show that every bijection from $\mathbb{N}$ to $\mathbb{N}$ is a directed equivalence.

II.4.5.3. Let $(X, \leq)$ and $(Y, \precsim)$ be directed sets, and $f : X \to Y$ a directed function. Show that there is a direction $\leqslant$ on X which is equivalent to $\leq$, for which f is order-preserving. [Set $a \leqslant b$ if $a \leq b$ and $f(a) \precsim f(b)$.]

II.4.5.4. Let $(X, \leq)$ and $(Y, \precsim)$ be directed sets. On $X \times Y$, define $(x_1, y_1) \leqslant (x_2, y_2)$ if $x_1 \leq x_2$ and $y_1 \precsim y_2$. Show that $\leqslant$ is a direction on $X \times Y$, and that the coordinate projections are directed functions. Generalize to arbitrary Cartesian products of directed sets (the Axiom of Choice is needed in the general case).

II.4.5.5. There are interesting countable directed sets other than $\mathbb{N}$ with the usual ordering. Define a relation $\precsim$ on $\mathbb{N}$ by setting $n \precsim m$ if $n|m$.

(a) Show that $\precsim$ is a partial order and a proper direction on $\mathbb{N}$.

(b) Show that $(n!)$ is a cofinal subsequence of $(\mathbb{N}, \precsim)$.

(c) Show that the sequence of prime numbers is not cofinal in $(\mathbb{N}, \precsim)$; in fact, the relative relation is not a direction. The sequence of powers of 2 is directed (even totally ordered) in the relative relation, but is not cofinal.

(d) Show that $\precsim$ is not equivalent to the ordinary order $\leq$ on $\mathbb{N}$.

(e) Can $(\mathbb{N}, \precsim)$ be written as an infinite product (cf. II.4.5.4.) of copies of $(\mathbb{N}, \leq)$?

II.5. Cardinality of Sets

> "The infinite! No other question has ever moved so profoundly the spirit of man; no other idea has so fruitfully stimulated his intellect; yet no other concept stands in greater need of clarification than that of the infinite."
>
> *David Hilbert*[6]

> "The rather colorless idea of a collection of elements that had lurked in the background of mathematical thought since prehistory might have remained there to this day if Cantor had not had the audacity to assume that they could be infinite. This was the bold and revolutionary hypothesis that launched modern mathematics; it should be seen as nothing less."
>
> *Penelope Maddy*[7]

One of the great mathematical advances of the late nineteenth century was CANTOR'S discovery, as part of his pioneering systematic study of infinite sets, that there are different "sizes" of infinite sets. CANTOR'S ideas were only gradually accepted by mathematicians, but eventually became one of the foundational cornerstones of modern analysis.

II.5.1. Cardinality

The term *cardinality* is used to denote the "size" of sets. This notion is most interesting and important for infinite sets; for finite sets it reduces to the "number of elements" in the set. While it is not essential for an analyst to have a detailed knowledge of the theory of cardinality, it is crucial to understand the most basic aspects of the theory: the distinction between finite and infinite sets, and between countable and uncountable sets. In this section, we will develop only enough of the general theory to establish what we need in analysis; a more complete treatment of the theory is in ().

The underlying idea is quite straightforward: two sets are the "same size" if there is a one-to-one correspondence (bijective function) between them.

II.5.1.1. DEFINITION. Let X and Y be sets. X and Y have the *same cardinality*, or are *equipotent* (or *equipollent*), written $card(X) = card(Y)$ or $|X| = |Y|$, if there is a bijective function from X to Y. We say $card(X) \leq card(Y)$ if X is equipotent with a subset of Y, i.e. if there is an injective function from X to Y. Write $card(X) < card(Y)$ if $card(X) \leq card(Y)$ and $card(X) \neq card(Y)$.

Note that we do not actually say what $card(X)$ *is*; we only use the term in comparing two sets. A definition of a cardinality (cardinal number) is given in (). Later in this section we will use particular symbols to denote the cardinality of certain standard sets, but even then no intrinsic meaning need be assigned to the cardinalities involved.

> "Mathematicians are not interested in the nature of their objects ... or in their metaphysical character. What matters are the relations and operations applying to the objects ... just as chess players are not interested in what a bishop 'is' but in how it performs."
>
> *A. Fraenkel*[8]

[6] [**?**]; English translation in [**?**].
[7] "Believing the Axioms I," *J. Symbolic Logic* **53** (1988), p. 486.
[8] [Fra66, p. 11].

II.5.2. Finite and Infinite Sets

"The notion of infinity is our greatest friend; it is also the greatest enemy of our peace of mind."

James Pierpont

"The infinite in mathematics is always unruly unless it is properly treated."

Edward Kasner and James Newman[9]

Infinite sets have some unusual properties which could be considered paradoxical, but with sufficient familiarity one can work with them comfortably; an analyst must develop a comfort level with them. One important feature of infinite sets is that they contain proper subsets which are as "large" as the whole set. For example, $\mathbb{N}$ can be written as the disjoint union of the even and odd natural numbers. Each of these sets is intuitively only "half" of $\mathbb{N}$, but on the other hand they are both actually equipotent with $\mathbb{N}$: the functions $f(n) = 2n$ and $g(n) = 2n - 1$ give bijections from $\mathbb{N}$ onto the even and odd numbers, respectively. (The oldest surviving printed record of this observation is due to Galileo, who actually used the obvious bijection $n \leftrightarrow n^2$ of $\mathbb{N}$ with the square numbers rather than even and odd ones.)

This phenomenon cannot occur in a finite set. Intuitively, a finite set is one which can be "counted." There are thus two standard ways of defining finite sets:

II.5.2.1. Definition. Let X be a set.

(i) X is *finite* if $X = \emptyset$ or $card(X) = card(\{m \in \mathbb{N} : 1 \leq m \leq n\})$ for some $n \in \mathbb{N}$. In the latter case, we say $card(X) = n$; and $card(\emptyset) = 0$.

(ii) X is *Dedekind-finite* if it is not equipotent with any proper subset of itself.

II.5.2.2. We can show (II.5.2.3.) that the two conditions are equivalent (technically, we must assume some version of the Axiom of Choice; the Countable AC suffices). A set which is not finite (in sense (i)) is called *infinite*. A set which is equipotent with a finite [infinite] set is finite [infinite]. It is intuitively obvious that any subset of a finite set is finite (see Exercise () for a proof); and hence any set which has an infinite subset is infinite.

We now carefully state the equivalence between the two definitions of a finite set. It is convenient to phrase the equivalence for infinite sets rather than for finite sets; the statement can be easily converted into the equivalence for finite sets.

II.5.2.3. Proposition. Let X be a set. Then the following are equivalent:

(i) X is infinite.

(ii) There is an injective function from X to X which is not surjective (i.e. X is not Dedekind-finite).

(iii) There is a surjective function from X to X which is not injective.

The proof, which uses II.5.2.5. below, is left to the exercises. The implication (ii) $\Rightarrow$ (i) is Exercise ()(c), (iii) $\Rightarrow$ (i) is Exercise ()(e), and (i) $\Rightarrow$ (ii) and (i) $\Rightarrow$ (iii) are Exercise ().

[9] *Mathematics and the Imagination*, p. 210.

II.5.2.4. The most obvious and important example of an infinite set is $\mathbb{N}$. (It is a fundamental property of $\mathbb{N}$ that it is not equipotent with $\{1, \ldots, n\}$ for any n. Although this seems obvious, it must in fact be proved; see Exercise ().) $\mathbb{N}$ is the "smallest" infinite set in the following sense:

II.5.2.5. PROPOSITION. Let X be a set. Then X is infinite if and only if X has a subset equipotent with $\mathbb{N}$, i.e. if and only if there is an injective function from $\mathbb{N}$ to X (or, in other words, a sequence of distinct elements in X.)

PROOF: If there is an injective function from $\mathbb{N}$ to X, then X is infinite. Conversely, suppose X is infinite. We will use an inductive procedure to define an injective function $f : \mathbb{N} \to X$. Let $f(1)$ be any element of X ($X \neq \emptyset$ since X is infinite.) Suppose $f(1), \ldots, f(n)$ have been defined to all be distinct. $X \setminus \{f(1), \ldots, f(n)\} \neq \emptyset$, since otherwise X would be equipotent with $\{1, \ldots, n\}$ and thus finite. So choose $f(n+1)$ to be any element of $X \setminus \{f(1), \ldots, f(n)\}$ (the Axiom of Choice, or at least the Axiom of Dependent Choice, is needed to do this properly; in fact, the result depends on the Countable AC – see Exercises II.5.6.14. and II.6.4.9..) 👍

Another Approach

We describe an alternate approach to finite sets which does not presuppose the availability of $\mathbb{N}$. This approach is perhaps less satisfying since it does not really specify what finite sets *are*, but it avoids certain circularities in definitions. It can be shown (II.5.6.16.) that the sets which are finite according to this definition are precisely the sets which are finite according to Definition II.5.2.1.. This approach was first taken by SIERPIŃSKI [].

The following definition is quite reminiscent of PEANO'S axiom P5 (III.2.2.1.).

II.5.2.6. DEFINITION. If X is a set, the *finite subsets* of X are the elements of the smallest subset $\mathcal{F}(X)$ of $\mathcal{P}(X)$ satisfying:

(i) $\emptyset \in \mathcal{F}(X)$.

(ii) $\{x\} \in \mathcal{F}(X)$ for all $x \in X$.

(iii) If $A, B \in \mathcal{F}(X)$, then $A \cup B \in \mathcal{F}(X)$.

X is a *finite set* if $X \in \mathcal{F}(X)$.

For any X, there is such a smallest subset $\mathcal{F}(X)$ of $\mathcal{P}(X)$, the intersection of all subsets of $\mathcal{P}(X)$ satisfying (i)–(iii) (this is well defined within ZF set theory).

II.5.2.7. PROPOSITION. Let $Y \subseteq X$. Then $\mathcal{F}(Y) = \{A \cap Y : A \in \mathcal{F}(X)\}$.

PROOF: Let $\mathcal{S} = \{A \cap Y : A \in \mathcal{F}(X)\} \subseteq \mathcal{P}(Y)$. Then it is easily checked that $\mathcal{S}$ satisfies (i)–(iii) (with X replaced by Y), and hence $\mathcal{F}(Y) \subseteq \mathcal{S}$. On the other hand, set

$$\mathcal{T} = \{A \cup B : A \in \mathcal{F}(Y), B \in \mathcal{F}(X \setminus Y)\} \subseteq \mathcal{P}(X) \ .$$

It is again easily checked that $\mathcal{T}$ satisfies (i)–(iii), and hence $\mathcal{F}(X) \subseteq \mathcal{T}$. If $C \in \mathcal{P}(X)$, then C can be uniquely written as $A \cup B$ with $A \in \mathcal{P}(Y)$ and $B \in \mathcal{P}(X \setminus Y)$, namely $A = C \cap Y$ and $B = C \cap (X \setminus Y)$; if $C \in \mathcal{F}(X)$, then $C \in \mathcal{T}$, and hence $A = C \cap Y \in \mathcal{F}(Y)$. Thus $\mathcal{S} \subseteq \mathcal{F}(Y)$. 👍

II.5.2.8. PROPOSITION. If $A \in \mathcal{F}(X)$ and $B \subseteq A$, then $B \in \mathcal{F}(X)$. In particular, every subset of a finite set is finite.

PROOF: Let $\mathcal{S} = \{A \in \mathcal{P}(X) : B \in \mathcal{F}(X) \text{ for all } B \subseteq A\}$. Then $\mathcal{S} \subseteq \mathcal{F}(X)$. It is easily checked that $\mathcal{S}$ satisfies (i)–(iii), so $\mathcal{F}(X) \subseteq \mathcal{S}$. 👍

II.5.2.9. COROLLARY. If $Y \subseteq X$, then $\mathcal{F}(Y) \subseteq \mathcal{F}(X)$, i.e. every subset of Y which is a finite subset of Y is a finite subset of X.

PROOF: Let $B \in \mathcal{F}(Y)$. Then $B = Y \cap A$ for some $A \in \mathcal{F}(X)$. Thus $B \subseteq A$, so $B \in \mathcal{F}(X)$. 👍

II.5.2.10. COROLLARY. If $Y \subseteq X$, then $\mathcal{F}(X) = \{A \cup B : A \in \mathcal{F}(Y), B \in \mathcal{F}(X \setminus Y)\}$.

PROOF: Set $\mathcal{T} = \{A \cup B : A \in \mathcal{F}(Y), B \in \mathcal{F}(X \setminus Y)\}$. The proof of II.5.2.7. shows that $\mathcal{F}(X) \subseteq \mathcal{T}$. Conversely, $\mathcal{F}(Y)$ and $\mathcal{F}(X \setminus Y)$ are contained in $\mathcal{F}(X)$ by II.5.2.9., and hence $\mathcal{T} \subseteq \mathcal{F}(X)$ by property (iii). 👍

II.5.2.11. PROPOSITION. Let X and Y be sets and $f : X \to Y$ a function. If $A \in \mathcal{F}(X)$, then $f(A) \in \mathcal{F}(Y)$, i.e. the image of a finite set is finite.

PROOF: Let $\mathcal{S} = \{B \subseteq X : f(B) \in \mathcal{F}(Y)\}$. Then it is easily checked that $\mathcal{S}$ satisfies (i)–(iii); hence $\mathcal{F}(X) \subseteq \mathcal{S}$. 👍

II.5.2.12. PROPOSITION. Let X be a totally ordered set. Then every nonempty finite subset of X has a largest element and a smallest element.

PROOF: Let $\mathcal{S}$ be the collection of subsets of X consisting of $\emptyset$ and all nonempty subsets of X which have a largest element and a smallest element. It is easily checked that $\mathcal{S}$ satisfies (i)–(iii). Thus $\mathcal{F}(X) \subseteq \mathcal{S}$. 👍

II.5.3. Countable and Uncountable Sets

Until the work of CANTOR, it was commonly believed by mathematicians that all infinite sets (if such sets exist at all; some mathematicians, both before and after CANTOR, rejected infinite sets entirely) were essentially the same (although even in ancient Greece a distinction was made between the "infinite" and the "infinitesimal"). But in fact there are sets which are not only infinite, but "very infinite" in the sense of having cardinality greater than $card(\mathbb{N})$. We explain this surprising fact in this section.

II.5.3.1. Definition. A set X is *countable* if $card(X) \leq card(\mathbb{N})$. X is *countably infinite* if $card(X) = card(\mathbb{N})$. A (necessarily infinite) set X is *uncountable* if $card(X) > card(\mathbb{N})$.

A set X is uncountable if there are "too many" elements in X to list in a sequence, or in other words if any sequence in X necessarily leaves out some (indeed, "almost all") elements of X. Any subset of a countable set is countable; thus if a set X has an uncountable subset, then X itself is uncountable. (If the Countable AC is not assumed, there could be sets which are neither countable nor uncountable by this definition.)

II.5.3.2. The symbol $\aleph_0$, pronounced "aleph zero", "aleph naught", or "aleph null", is commonly used to denote the cardinality of $\mathbb{N}$. ($\aleph$, "aleph", is the first letter of the Hebrew alphabet.) Thus X is countable if X is finite or $card(X) = \aleph_0$ ()

It is not obvious that uncountable sets exist, but the following famous theorem of Cantor shows that they do exist in abundance:

II.5.3.3. Theorem. Let X be any set. Then $card(\mathcal{P}(X)) > card(X)$.

Proof: It is easy to see that $card(X) \leq card(\mathcal{P}(X))$, since there is an obvious one-one correspondence between X and the set of singleton subsets of X. So we need to show that X and $\mathcal{P}(X)$ are not equipotent. Suppose f is a function from X to $\mathcal{P}(X)$; we will show that f cannot be surjective, which will prove the theorem. If $x \in X$, then $f(x)$ is a subset of X, so it makes sense to ask whether $x \in f(x)$. Define

$$Y = \{x \in X : x \notin f(x)\} \subseteq X.$$

We claim there can be no $y \in X$ with $Y = f(y)$. Suppose there is; is $y \in Y$? If $y \in Y$, then $y \in f(y)$ since $Y = f(y)$ by assumption; but then $y \notin Y$ by definition of Y, a contradiction. Similarly, if $y \notin Y$, then $y \in f(y)$, so $y \in Y$, another contradiction. Thus the assumption that $Y = f(y)$ for some $y \in X$ necessarily leads to a contradiction, so there is no such y. 👍

II.5.3.4. This proof was adapted by Bertrand Russell to establish his Paradox (II.5.6.13.) (Zermelo actually discovered this paradox before Russell, using a similar argument). A simple version of this paradox based on II.5.3.3. is the following: if there is a set X of all sets, then every subset of X is a set, hence a member of X; so $\mathcal{P}(X) \subseteq X$. But it is then easy to construct a surjective function from X to $\mathcal{P}(X)$ (cf. II.5.4.1.), contradicting II.5.3.3..

A similar, but more sophisticated, type of "self-referential" argument was used by Kurt Gödel in his famous Incompleteness Theorem (see [**?**] or [Ruc82] for mathematically accurate popular accounts of this theorem; for the full story see [Smu92].)

II.5.3.5. Corollary. $\mathcal{P}(\mathbb{N})$ is an uncountable set.

Since $card(\mathcal{P}(\mathbb{N})) = card(\{0,1\}^{\mathbb{N}})$ (), we obtain that there are uncountably many sequences of 0's and 1's. Because of this, the symbol $2^{\aleph_0}$ is often used for the cardinality of $\mathcal{P}(\mathbb{N})$ (the notation is explained in ().) From this we can obtain a proof (see Exercise ()) of the following theorem, which is one of the most important facts in analysis. We instead give a somewhat different proof, using a "diagonalization argument" of Cantor (the proof is not really different; see II.5.6.9.. For another version of the same proof, see II.5.6.10..)

II.5.3.6. THEOREM. $\mathbb{R}$ is uncountable. In fact, every nonempty open interval in $\mathbb{R}$ is uncountable.

PROOF: It suffices to show that $[0,1)$ is uncountable, since every open interval contains a half-open subinterval, which is equipotent with $[0,1)$ via a linear function. Every number in $[0,1)$ has a unique decimal expansion

$$.a_1a_2a_3\cdots$$

where $0 \leq a_n \leq 9$ for each n, which does not end in an infinite string of 9's.

Suppose we consider a sequence (x_n) in $[0,1)$. We will construct an $x \in [0,1)$ which is not equal to any x_n; this will prove that $[0,1)$ is uncountable since no sequence can exhaust the set. Write x_n in its decimal expansion

$$x_n = .a_{n1}a_{n2}a_{n3}\cdots$$

not ending in a string of 9's, where a_{nm} is the m'th decimal digit of x_n. Now for each n let b_n be a digit which is not 9 and which is different from a_{nn} (to be precise, we can let $b_n = 0$ if $a_{nn} \neq 0$ and $b_n = 1$ if $a_{nn} = 0$), and set $x = .b_1b_2b_3\cdots$. Then, for any n, $x \neq x_n$ since it differs from x_n in at least the n'th decimal place. 👍

II.5.3.7. The cardinality of $\mathbb{R}$ is often called the *cardinality of the continuum*, denoted $\mathfrak{c}$. In fact, $card(\mathbb{R}) = card(\mathcal{P}(\mathbb{N}))$ (), so $\mathfrak{c} = 2^{\aleph_0}$. Every nontrivial interval in $\mathbb{R}$ also has cardinality $\mathfrak{c}$ by Exercise ().

II.5.3.8. From Theorem II.5.3.3., it follows that not only are there uncountable sets, but there are uncountable sets of different sizes: for example, $card(\mathcal{P}(\mathbb{R})) = card(\mathcal{P}(\mathcal{P}(\mathbb{N}))) > card(\mathcal{P}(\mathbb{N})) = 2^{\aleph_0}$ (i.e. $\mathcal{P}(\mathbb{R})$ is "very uncountable.") The symbol $2^{\mathfrak{c}}$ or $2^{2^{\aleph_0}}$ is used for $card(\mathcal{P}(\mathbb{R}))$. In fact, there is a never-ending hierarchy of infinite cardinalities:

$$card(\mathbb{N}) < card(\mathbb{R}) < card(\mathcal{P}(\mathbb{R})) < card(\mathcal{P}(\mathcal{P}(\mathbb{R}))) < card(\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{R})))) < \cdots$$

and there is no "largest cardinality." (There are even cardinalities much larger than all the ones on this list!)

II.5.3.9. Note that II.5.3.3. is valid also in the case where X is a finite set. In this case, it amounts to the statement that $2^n > n$ for every $n \geq 0$.

II.5.4. Properties of Countable Sets

We will now analyze countable sets in more detail. We first establish a useful criterion for countability:

II.5.4.1. PROPOSITION. Let X be a nonempty set. Then X is countable if and only if there is a surjective function from $\mathbb{N}$ to X.

PROOF: First assume that X is countable. By definition, there is an injective function $f : X \to \mathbb{N}$. Fix an $x_0 \in X$, and let $S = f(X) \subseteq \mathbb{N}$. Define a function $g : \mathbb{N} \to X$ by $g(n) = x$ if $n \in S$ and $f(x) = n$ (this is well defined since f is injective), and set $g(n) = x_0$ if $n \notin S$. Then g is a surjective function from $\mathbb{N}$ to X.

Conversely, suppose $g : \mathbb{N} \to X$ is a surjective function. For each $x \in X$, let $f(x)$ be the smallest number in $g^{-1}(\{x\})$. Then f is an injective function from X to $\mathbb{N}$, so $card(X) \leq card(\mathbb{N})$. 👍

II.5.4.2. Corollary. If X and Y are sets and there is a surjective function $f : X \to Y$, and X is countable, then Y is countable.

It seems obvious that if X is countable, then either X is finite or $card(X) = card(\mathbb{N})$, and hence the term "countably infinite" means precisely "countable and infinite"; however, this takes a little proof:

II.5.4.3. Proposition. Let X be a countable set. Then either X is finite, or $card(X) = card(\mathbb{N})$.

This would seem to follow from II.5.2.5., which says that $card(\mathbb{N}) \leq card(X)$ if X is infinite, and $card(X) \leq card(\mathbb{N})$ by definition if X is countable. However, it is not obvious that these two inequalities together imply that $card(X) = card(\mathbb{N})$ (that is, that the "ordering" on cardinalities is antisymmetric.) This antisymmetry is true in general by the Schröder-Bernstein Theorem, but we can and will take a simpler approach in the countable case.
Proof: If $X = \emptyset$, the conclusion is true since $\emptyset$ is finite by definition. Assume X is nonempty. By II.5.4.1., there is a surjective function $g : \mathbb{N} \to X$. Inductively define a function $f : X \to \mathbb{N}$ as follows. Let $x_1 = g(1)$, and set $f(x_1) = 1$. If $x_1, \ldots, x_n$ have been defined with $f(x_k) = k$ for $1 \leq k \leq n$, then either $X = \{x_1, ..., x_n\}$, in which case X is finite, or $Y_n = X \setminus \{x_1, \ldots, x_n\} \neq \emptyset$. In the latter case, let k be the smallest element of $\mathbb{N}$ with $g(k) \in Y_n$ (there is such a k since g is surjective.) Set $x_{n+1} = g(k)$ and $f(x_{n+1}) = k$. If this process terminates, then X is finite; and if it never terminates, then f is an injective function from X onto $\mathbb{N}$. 👍

Another useful fact is that $\mathbb{N}^2 = \mathbb{N} \times \mathbb{N}$ is equipotent with $\mathbb{N}$. This seems a little surprising, but is easily proved:

II.5.4.4. Proposition. $\mathbb{N}^2$ is equipotent with $\mathbb{N}$, and hence is countable (countably infinite.)

Proof: The function $(n, m) \mapsto 2^{n-1}(2m - 1)$ is a bijection from $\mathbb{N}^2$ to $\mathbb{N}$, as is easily checked using the unique factorization in $\mathbb{N}$. 👍

II.5.4.5. Corollary. A Cartesian product of a finite number of countable sets is countable.

II.5.4.6. It is not true that a Cartesian product of a countable number of countable sets is countable; for example, $\{0, 1\}^{\mathbb{N}}$ is uncountable ().

There are some other very important consequences of II.5.4.4.:

II.5.4.7. Corollary. $\mathbb{Z}$ and $\mathbb{Q}$ are countable.

Proof: To show that $\mathbb{Z}$ is countable, define a function $f : \mathbb{N}^2 \to \mathbb{Z}$ by $f(n, m) = (-1)^n(m - 1)$; then f is surjective, so $\mathbb{Z}$ is countable by II.5.4.4. and II.5.4.2.. Thus $\mathbb{Z} \times \mathbb{N}$ is countable. Define $g : \mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$ by $g(n, m) = n/m$, and apply II.5.4.2. again to conclude that $\mathbb{Q}$ is countable. 👍

II.5.4.8. Corollary. A countable union of countable sets is countable. More precisely, let $\{X_i : i \in I\}$ be an indexed collection of sets. If each X_i is countable and I is countable, then $\cup_{i\in I} X_i$ is countable.

Proof: We may assume each X_i is nonempty by deleting the empty ones. For each i let f_i be a surjective function from $\mathbb{N}$ onto X_i. Then $(i, n) \mapsto f_i(n)$ is a surjective function from $I \times \mathbb{N}$ to $\cup_{i\in I} X_i$. Apply II.5.4.5. and II.5.4.2.. (This argument requires the Countable Axiom of Choice.) 👍

There is also a version of these facts for the continuum:

II.5.4.9. Proposition.

(i) If X and Y have cardinality $\leq 2^{\aleph_0}$, then $card(X \times Y) \leq 2^{\aleph_0}$.

(ii) Let $\{X_i : i \in I\}$ be an indexed collection of sets. If $card(X_i) \leq 2^{\aleph_0}$ for all i and $card(I) \leq 2^{\aleph_0}$, then $card(\cup_{i\in I} X_i) \leq 2^{\aleph_0}$.

Proof: It suffices to prove (i), since then (ii) follows as in the proof of II.5.4.8. (the full Axiom of Choice is needed). To prove (i), we need only show that $card(\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})) = 2^{\aleph_0}$. But there is a bijective function from $\mathcal{P}(\mathbb{N}) \times \mathcal{P}(\mathbb{N})$ to $\mathcal{P}(\mathbb{N})$ given by $(A, B) \mapsto f(A) \cup g(B)$, where $f(n) = 2n$ and $g(n) = 2n-1$ are bijections from $\mathbb{N}$ onto the even and odd numbers. 👍

This result is actually true for any infinite cardinal (II.9.9.3.) (but depends on the Axiom of Choice).

Countable Ordered Sets

II.5.4.10. The ordered set $\mathbb{Q}$ can be characterized by a few simple properties. $\mathbb{Q}$ satisfies:

(i) $\mathbb{Q}$ is countable.

(ii) $\mathbb{Q}$ is totally ordered.

(iii) $\mathbb{Q}$ has no largest or smallest element.

(iv) $\mathbb{Q}$ has no consecutive elements: if $a, b \in \mathbb{Q}$, $a < b$, then there is a $c \in \mathbb{Q}$ with $a < c < b$.

Property (iv) is sometimes phrased by saying that the ordering on $\mathbb{Q}$ is *dense*. This condition implies, by repeated application, that between any two elements of $\mathbb{Q}$ there are actually infinitely many other elements.

It turns out that these conditions completely characterize $\mathbb{Q}$ as an ordered set:

II.5.4.11. THEOREM. Let $(X, \precsim)$ be a countable totally ordered set with no consecutive elements and no first or last element. Then $(X, \precsim)$ is order-isomorphic to $\mathbb{Q}$.

PROOF: Let $(x_n : n \in \mathbb{N})$ be an enumeration of X, and $(y_n : n \in \mathbb{N})$ an enumeration of $\mathbb{Q}$. We will inductively construct an order-isomorphism f of X onto $\mathbb{Q}$. Begin by defining $f(x_1) = y_1$. Now suppose f has been defined on $x_1, \ldots, x_n$ and possibly finitely many other x_i, and $y_1, \ldots, y_n$ are in the range of the part of f already defined. We will define f on x_{n+1} and possibly on another x_i, so that y_{n+1} is in the range. Let X_n be the set of x_i for which f has already been defined, and $Y_n = f(X_n)$.

First define f on x_{n+1}. If $x_{n+1} \in X_n$, $f(x_{n+1})$ is already defined and we skip this step. Otherwise, note that either $x_{n+1} \prec x_i$ for all $x_i \in X_n$, or there is a largest $x_i \in X_n$ with $x_i \prec x_{n+1}$; let a be this x_i. Similarly, either $x_j \prec x_{n+1}$ for all $x_j \in X_n$, or there is a smallest x_j with $x_{n+1} \prec x_j$; let b be this x_j. Let k be the smallest index for which y_k satisfies $f(a) < y_k < f(b)$ (or $y_k < f(b)$ if a is undefined, $f(a) < y_k$ if b is undefined.) We automatically have $y_k \notin Y_n$. Set $f(x_{n+1}) = y_k$.

Then work backwards in a similar way for the other part of the inductive step. If y_{n+1} is already in the range of f, skip this step. Otherwise, let c and d be the largest and smallest elements of the range of f (as defined so far) with $c < y_{n+1} < d$ (either c or d may be undefined.) Then $c = f(x_i)$ and $d = f(x_j)$ for some i, j. Let k be the smallest index for which $x_i \prec x_k \prec x_j$; then $f(x_k)$ is undefined and so we can define $f(x_k) = y_{n+1}$.

It is easy to check that the f thus defined is a bijection and an order-isomorphism. 👍

II.5.4.12. A similar argument shows that if $(X, \precsim)$ is a countable totally ordered set with no consecutive elements, but with a first or last element, then $(X, \precsim)$ is order-isomorphic to $\mathbb{Q} \cap [0, 1)$, $\mathbb{Q} \cap (0, 1]$, or $\mathbb{Q} \cap [0, 1]$.

II.5.4.13. COROLLARY. Every countable totally ordered set is order-isomorphic to a closed subset of $\mathbb{Q}$.

PROOF: Let $(X, \precsim)$ be a countable totally ordered set. If $x \in X$, define $Y_x \subseteq \mathbb{Q}$ by

$Y_x = \{0\}$ if x has no immediate predecessor or successor, but is not the largest or smallest element of X.

$Y_x = \mathbb{Q} \cap (-\infty, 0]$ if x has an immediate predecessor or is the smallest element of X, and a successor but not an immediate successor.

$Y_x = \mathbb{Q} \cap [0, \infty)$ if x has an immediate successor or is the largest element of X, and a predecessor but not an immediate predecessor.

$Y_x = \mathbb{Q}$ if x has an immediate predecessor or is the smallest element of X, and an immediate successor or is the largest element of X.

Let $Z = \{(x, y) : y \in Y_x\} \subseteq X \times \mathbb{Q}$. Give $X \times \mathbb{Q}$ the lexicographic ordering, and give Z the relative ordering. It is easily checked that Z with this ordering is countable, totally ordered, and has no consecutive elements or largest or smallest element. Thus Z is order-isomorphic to $\mathbb{Q}$. $(X, \precsim)$ is order-isomorphic to the subset $X \times \{0\}$ of Z. It is easy to check that if (x_n) is an increasing sequence in X with supremum x, then $(x, 0)$ is the supremum of $((x_n, 0))$ in Z and conversely, and similarly for decreasing sequences, so $X \times \{0\}$ is closed in $Z \cong \mathbb{Q}$ by ().

👍

II.5.5. Additional Facts about Cardinality

There are three rather natural questions about cardinality which turn out to have interesting answers:

II.5.5.1. QUESTIONS.

(1) If X and Y are sets with $card(X) \leq card(Y)$ and $card(Y) \leq card(X)$, is $card(X) = card(Y)$, i.e. if each of X and Y can be put in bijective correspondence with a subset of the other, are X and Y equipotent? (Is the "ordering" on cardinalities a true partial order, i.e. antisymmetric?)

(2) If X and Y are sets, is it always true that $card(X) \leq card(Y)$ or $card(Y) \leq card(X)$, i.e. are any two sets comparable in size? (Is the partial ordering on cardinalities a total ordering?)

(3) Is $\mathfrak{c} = 2^{\aleph_0}$ the smallest cardinality larger than $\aleph_0$, i.e. if $X \subseteq \mathbb{R}$, is it true that either X is countable or $card(X) = card(\mathbb{R})$? (The i.e. statement is only equivalent to the first statement if the answer to question (2) is yes for subsets of $\mathbb{R}$.)

It turns out that only question 1 has a definitive answer, given by the *Schröder-Bernstein Theorem*:

II.5.5.2. THEOREM. Let X and Y be sets. If $card(X) \leq card(Y)$ and $card(Y) \leq card(X)$, then $card(X) = card(Y)$.

A proof is outlined in Exercise II.5.6.11.. Actually, the commonly-used name "Schröder-Bernstein Theorem" is not really accurate, since the proof published by SCHRÖDER in 1898 was incorrect; the proof published by F. BERNSTEIN the same year (he was 19 years old when he did it) was the first and only correct proof at that time and is the one still commonly used. The theorem is sometimes called the "Cantor-Bernstein Theorem" since CANTOR stated it in his book without proof; apparently CANTOR tried rather hard to prove it, without success. There are indications that DEDEKIND found a proof in the 1880's, but never published it.

II.5.5.3. Question 2 turns out to be dependent on the Axiom of Choice, which is independent of the other (ZF) axioms of set theory. If the Axiom of Choice is assumed, as is usual in analysis, then a positive answer to question 2 can be obtained; see II.9.9.1..

II.5.5.4. The conclusion of Question 3, that $\mathfrak{c}$ is the smallest uncountable cardinality, is called the *Continuum Hypothesis*. This is also known to be independent of the other axioms of set theory, including the Axiom of Choice (ZFC), through work of KURT GÖDEL and PAUL COHEN. There is also a *Generalized Continuum Hypothesis* which is independent of ZFC.

II.5.6. Exercises

II.5.6.1. Let f be an injective function from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$. Inductively define a function g from an initial segment of $\mathbb{N}$ to $\{1, \ldots, n\}$ as follows. Let k_1 be the smallest element of $S = f(\{1, \ldots, n\})$, and set $g(1) = f^{-1}(k_1)$. Inductively let k_{r+1} be the smallest element of $S \setminus \{k_1, \ldots, k_r\}$ if this set is nonempty, and $g(r+1) = f^{-1}(k_{r+1})$.

(a) Show that g is defined precisely on $\{1,\dots,n\}$, and that g maps $\{1,\dots,n\}$ bijectively to itself.

(b) Let $h = f \circ g$. Show that h is an injective function from $\{1,\dots,n\}$ onto S, and prove inductively that h is increasing and $h(k) \geq k$ for every $k \in \{1,\dots,n\}$.

(c) On the other hand, $h(n) \leq n$ since $S \subseteq \{1,\dots,n\}$. Conclude that $h(k) = k$ for every k, and thus that h is surjective, i.e. that f is surjective. [Show inductively that if $h(k_0) > k_0$ for some k_0, then $h(k) > k$ for all $k > k_0$.] Use this to show that an injective function from a finite set to itself is surjective.

(d) Suppose g is an injective function from $\mathbb{N}$ to $\{1,\dots,n\}$. If h is the restriction of g to $\{1,\dots,n+1\}$, then h is an injective function from $\{1,\dots,n+1\}$ to $\{1,\dots,n+1\}$ whose range is contained in $\{1,\dots,n\}$, contradicting (c). Thus such a g cannot exist.

(e) Let h be a surjective function from $\{1,\dots,n\}$ to $\{1,\dots,n\}$. For each $k \in \{1,\dots,n\}$, let $g(k)$ be the smallest element of $h^{-1}(\{k\})$. Then g is an injective function from $\{1,\dots,n\}$ to $\{1,\dots,n\}$, and hence is surjective by (c). Conclude that h is injective. From this show that a surjective function from a finite set to itself is injective.

II.5.6.2. Let S be a subset of $\{1,\dots,n\}$. Define a function f from S to $\mathbb{N}$ inductively by letting k_1 be the smallest element of S and setting $f(k_1) = 1$, and inductively letting k_{r+1} being the smallest element of $S \setminus \{k_1,\dots,k_r\}$ if this set is nonempty, and $f(k_{r+1}) = r+1$.

(a) Prove by induction that $f(k) \leq k$ for every $k \in S$.

(b) Conclude that $f(S) = \{1,\dots,m\}$ for some $m \leq n$, and hence that S is finite.

(c) Show that a subset of a finite set is finite.

(d) Show that if X and Y are finite sets with $card(X) = n$ and $card(Y) = m$ for some $n, m \in \mathbb{N} \cup \{0\}$, then $card(Y) \leq card(X)$ if and only if $m \leq n$, so the "ordering" on cardinalities of finite sets is consistent with the ordering on $\mathbb{N} \cup \{0\}$. In particular, if $Y = X$, then $m = n$, i.e. the cardinality of a finite set is a well defined nonnegative integer.

II.5.6.3. Prove (i) $\Rightarrow$ (ii) and (i) $\Rightarrow$ (iii) of II.5.2.3.. [Hint: first prove that the statements (ii) and (iii) are true if $X = \mathbb{N}$. For a general infinite set X, use II.5.2.5..]

II.5.6.4. Let X be an infinite set, and Y a countable set. Prove that $card(X \cup Y) = card(X)$. [Use II.5.2.5..]

II.5.6.5. Show, without the Countable AC, that the set of sequences in $\mathbb{Q}$ with only finitely many nonzero terms is countable. [Give an explicit enumeration.]

II.5.6.6. (a) Prove that the set of polynomials with rational coefficients is countable. [Use Exercise II.5.6.5..]

(b) A real or complex number is *algebraic* if it is a root of a nonzero polynomial with rational coefficients (which can be assumed to be integers by clearing denominators.) Show that the set of algebraic numbers is countable. (Do not use the Countable AC; note that $\mathbb{C}$ can be totally ordered, e.g. with the lexicographic ordering.)

(c) A real or complex number is *transcendental* if it is not algebraic. Prove that there exists a transcendental real number (and, in fact, the cardinality of the set of transcendental real numbers is $\mathfrak{c}$ by Exercise II.5.6.4..)

II.5.6.7. (a) Prove that any two bounded open intervals (a, b) and (c, d) in $\mathbb{R}$ are equipotent by finding a bijection [consider linear functions.]

(b) Use the tangent function to show that $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ is equipotent with $\mathbb{R}$.

(c) Give two proofs that $card([0,1]) = card([0,1)) = card((0,1]) = card((0,1))$, one using the Schröder-Bernstein Theorem and the other by directly finding bijections [use II.5.2.5..]

(d) Show that all intervals in $\mathbb{R}$ have the same cardinality.

II.5.6.8. (a) Define a function $f : \{0,1\}^{\mathbb{N}} \to [0,1)$ by $f(a_1, a_2, a_3, \cdots) = .a_1a_2a_3\cdots$, where the last expression is the decimal expansion of a number in $[0,1)$. Show that f is injective, and hence $card([0,1)) \geq card(\{0,1\}^{\mathbb{N}}) = 2^{\aleph_0}$, so $[0,1)$ is uncountable.

(b) Define a function $g : [0,1) \to \{0,1\}^{\mathbb{N}}$ by $g(x) = (a_1, a_2, \cdots)$, where $x = .a_1a_2\cdots$ is the *binary* expansion of x not ending in a string of 1's. Show that g is injective, and hence $card([0,1)) \leq card(\{0,1\}^{\mathbb{N}}) = 2^{\aleph_0}$.

(c) Apply the Schröder-Bernstein Theorem to conclude that $card([0,1)) = 2^{\aleph_0}$.

(d) As an alternative to (c), $\{0,1\}^{\mathbb{N}} \setminus g([0,1))$ consists precisely of all sequences which are eventually 1. Show that this set is countable (see Exercise II.5.6.5.), and hence $card(\{0,1\}^{\mathbb{N}}) = card(g([0,1)) = card([0,1)) = card(\mathbb{R})$ by Exercises II.5.6.4. and II.5.6.7..

II.5.6.9. Show that the proof of II.5.3.3. is really a diagonalization argument along the lines of the proof of II.5.3.6., as follows.

(a) Let X be a set, and f a function from X to $\mathcal{P}(X)$. Define a subset Y of X as follows: for $x \in X$, let $x \in Y$ if and only if $x \notin f(x)$.

(b) Show that Y is not in the range of f, i.e. Y is not in the "list" $\{f(x) : x \in X\}$ of subsets of X defined by f, since for each $x \in X$ the set Y differs from $f(x)$ at least with regard to membership of x (i.e. for each $x \in X$, x belongs to the symmetric difference $Y \triangle f(x)$).

II.5.6.10. This exercise outlines an apparently different proof that $\mathbb{R}$ is uncountable, based on the Nested Intervals Theorem (). This was essentially Cantor's original proof. We show that $[0,1]$ is uncountable.

(a) Let (x_n) be a sequence in $[0,1]$. We will show that there is an $x \in [0,1]$ with $x \neq x_n$ for all n.

(b) Let $I_0 = [0,1]$. Inductively define I_n as follows. Suppose I_n has been defined. Subdivide I_n into three closed subintervals $J_{n,1}$, $J_{n,2}$, $J_{n,3}$ from left to right, of equal length 3^{-n-1}, meeting only at endpoints. Then there is at least one k with $x_{n+1} \notin J_{n,k}$; let I_{n+1} be the leftmost such $J_{n,k}$.

(c) Set $S = \cap_{n=1}^{\infty} I_n$. Then $S \neq \emptyset$ by the Nested Intervals Theorem. If $x \in S$, then $x \neq x_n$ for all n. (Actually $S = \{x\}$.)

(d) Show that this proof is essentially the diagonalization argument, using base 3 decimals.

II.5.6.11. This exercise outlines a proof of the Schröder-Bernstein Theorem. Suppose X and Y are sets, and there are injective functions $f : X \to Y$ and $g : Y \to X$. We will construct a bijection h from X to Y.

Let $A_1 = X \setminus g(Y)$ and $B_1 = f(A_1)$, and inductively define $A_{n+1} = g(B_n)$ and $B_{n+1} = f(A_{n+1})$.

(a) Prove by induction on n that the A_n are disjoint subsets of X and the B_n are disjoint subsets of Y.

(b) Let $A = \cup_{n=1}^{\infty} A_n$ and $B = \cup_{n=1}^{\infty} B_n$. Show that f maps A onto B and g maps $Y \setminus B$ onto $X \setminus A$. [Since g maps B onto $\cup_{n=2}^{\infty} A_n$, g maps $Y \setminus B$ onto $g(Y) \setminus \cup_{n=2}^{\infty} A_n = (X \setminus A_1) \setminus \cup_{n=2}^{\infty} A_n = X \setminus A$.]

(c) Define $h : X \to Y$ as follows. If $x \in A$, let $h(x) = f(x)$. If $x \in X \setminus A$, let $h(x) = g^{-1}(x)$. Show that h is a bijection from X to Y.

II.5.6.12. Let X be a set of cardinality κ. Assume that $\kappa^2 = \kappa \geq \aleph_0$ and that X can be totally ordered. Fix a total order on X and a one-one correspondence $\phi : X \to X^2$, and a one-to-one map $\psi : \mathbb{N} \to X$.

(a) Show that for $n > 2$, iteration of ϕ defines a one-one correspondence ϕ_n from X to X^n. Let $\phi_2 = \phi$ and $\phi_1 : X \to X$ be the identity.

(b) Show that the ϕ_n define a one-one correspondence between $\mathbb{N} \times X$ and the disjoint union of the X^n.

(c) Use the fixed ordering on X to identify the set of n-element subsets of X with a subset of X^n.

(d) From (b) and (c) obtain a one-one correspondence between the set Φ of finite subsets of X and a subset of $\mathbb{N} \times X$. The function ψ then gives a one-one correspondence between Φ and a subset of X^2. Conclude that $|\Phi| \leq |X^2| = |X|$, so that $|\Phi| = |X|$ by the Schröder-Bernstein Theorem.

II.5.6.13. **(Russell's Paradox)** This argument shows that there can be no such thing as the "set of all sets."

(a) Say that a set A is *ordinary* if A is not a member of A. (Almost every set normally encountered in mathematics is an ordinary set by this definition.)

(b) Suppose X is the set of all sets. Let Y be the collection of all ordinary sets. Then Y is a subset of X, hence a set. Is Y an ordinary set?

II.5.6.14. Show without using any form of the Axiom of Choice that a set is Dedekind-finite if and only if it does not have any countably infinite subset. [One direction follows from II.5.2.3.. Conversely, if there is an $f : X \to X$ which is injective but not surjective, fix $x \in X$ not in the range of f, and consider $x, f(x), f(f(x)), \ldots$. See Exercise II.6.4.9..]

II.5.6.15. (a) Show (without any form of AC) that if X is any set, then there is a set z with $z \notin X$. [Use II.5.3.3. to show that $X \cap \mathcal{P}(X)$ is a proper subset of $\mathcal{P}(X)$.]

(b) Use (a) to give an alternate proof that there is no "set of all sets."

(c) Show (without any form of AC) that if X and Y are sets, then there is a set Y' which is equipotent with Y and disjoint from X. [Let z be a set which is not in $\mathcal{C}(X)$, and let

$$Y' = \{(z, y) : y \in Y\} = \{\{\{z\}, \{z, y\}\} : y \in Y\} .$$

If $x \in Y'$, then $z \in \mathcal{C}(x)$, but then $x \notin X$ by II.2.1.4..]

II.5.6.16. Show that a nonempty set X is finite in the sense of II.5.2.6. if and only if X is equipotent to $\{1, \ldots, n\}$ for some $n \in \mathbb{N}$. [Show that the set $\mathcal{S}$ of $\emptyset$ and all subsets of X equipotent to $\{1, \ldots, n\}$ for some n satisfies (i)–(iii). For the converse, see III.1.4.11..]

II.5.6.17. Let X be an infinite set. Show that $\mathcal{P}(\mathcal{P}(X))$ is Dedekind-infinite. [Consider the set of n-element subsets of X for each $n \in \mathbb{N}$.]

II.5.6.18. ***T*-Finite Sets.**

DEFINITION. A set X is *T-finite* if every nonempty collection $\mathcal{S}$ of subsets of X has a $\subseteq$-minimal element, i.e. a set $Y \in \mathcal{S}$ such that no proper subset of Y is in $\mathcal{S}$.

(a) Show that X is T-finite if and only if every nonempty collection of subsets has a $\subseteq$-maximal element. [Consider complements.]

(b) Show that every finite set is T-finite.

(c) Show that every T-finite set is Dedekind-finite. [If X is Dedekind-infinite, let $\mathcal{S}$ be the set of all Dedekind-infinite subsets of X. If Y is a Dedekind-infinite subset of X and $p \in Y$, show that $Y \backslash \{p\}$ is Dedekind-infinite.]

(d) Show that every T-finite set is finite. [Consider the set of all finite subsets of X, and use (a).]

The T stands for TARSKI, who introduced this definition and examined its properties [].

II.5.6.19. Here is another paradox showing there cannot be a "set of all sets." This one was noticed by CANTOR, but not as quickly as it could have been.

(a) Show that if X is a set and Y a nonempty subset of X, then there is a surjective function from X to Y. [Adapt the proof of II.5.4.1..]

(b) Suppose X is the set of all sets. Then every element of $\mathcal{P}(X)$ is a subset of X, hence a set, so $\emptyset \neq \mathcal{P}(X) \subseteq X$.

(c) By (a), there is a surjective function from X to $\mathcal{P}(X)$, contradicting Cantor's Theorem (II.5.3.3.).

II.6. The Axiom of Choice and Zorn's Lemma

If $\{X_i : i \in I\}$ is an indexed collection of nonempty sets, it would seem obvious that we could pick an element out of each set; after all, isn't this what is meant by saying that each X_i is nonempty? However, with more thought, although we can "obviously" choose an element out of any individual X_i (but see II.6.1.6.), it becomes less obvious that we could *simultaneously* choose an element out of each X_i (E. Cheng [**?**] calls this "decision fatigue"). In fact, it can be neither proved nor disproved in ordinary (ZF) set theory that this can always be done![10] Since it is very useful in analysis (and elsewhere in mathematics) to be able to make such selections, an additional axiom is usually added to set theory, called the *Axiom of Choice*; the expanded theory is called *ZFC* set theory.

II.6.1. The Axiom of Choice

II.6.1.1. Definition. Let $\{X_i : i \in I\}$ be an indexed collection of (nonempty) sets. A *choice function* for $\{X_i : i \in I\}$ is a function $c : I \to \cup_{i\in I} X_i$ such that $c(i) \in X_i$ for each i.

If there is a choice function for $\{X_i : i \in I\}$, then each X_i is obviously nonempty. The Axiom of Choice asserts the converse:

II.6.1.2. **Axiom of Choice:** (Indexed Form) If $\{X_i : i \in I\}$ is an indexed collection (set) of nonempty sets, then there is a choice function for $\{X_i : i \in I\}$.

(Actually one must be more careful stating this version, cf. II.7.5.8.; we will gloss over this technicality, which has no practical effect for the applications in this book.)

The Axiom of Choice is often stated in a slightly different but equivalent form:

II.6.1.3. **Axiom of Choice:** (Power Set Form) If X is a nonempty set, there is a function (choice function) $c : \mathcal{P}(X) \setminus \{\emptyset\} \to X$ with $c(A) \in A$ for every nonempty $A \subseteq X$.

II.6.1.4. To obtain the Power Set Form from the Indexed Form, set $I = \mathcal{P}(X) \setminus \{\emptyset\}$, and set $X_A = A$. Conversely, to obtain the Indexed Form from the Power Set Form, let $\{X_i : i \in I\}$ be an indexed collection of nonempty sets, and set $X = \cup_{i\in I} X_i$. Let $c : \mathcal{P}(X) \setminus \{\emptyset\} \to X$ be a choice function for X, and let $h : I \to \mathcal{P}(X) \setminus \{\emptyset\}$ be defined by $h(i) = X_i$. Then $c \circ h$ is a choice function for $\{X_i : i \in I\}$.

II.6.1.5. The Axiom of Choice (abbreviated AC) is often slightly misunderstood, and it is important to understand what it does *not* say. The AC is frequently stated somewhat informally as:

> "If $\{X_i\}$ is a collection of nonempty sets, you can choose (or simultaneously choose) an element out of each of the sets."

This statement of the AC is not incorrect if properly interpreted, but it can be misleading. What the AC actually asserts is simply the *existence* of a choice function; it does not give any procedure for finding or describing one, or even assert that there is an explicit description of one (in fact, the AC is typically invoked precisely in situations where an explicit description of a choice function is not possible.)

[10] This statement is only true if ZF set theory is assumed to be consistent; consistency of ZF set theory remains unproved and is perhaps unprovable (it is unprovable within ZF.)

II.6.1.6. The distinction comes into sharpest focus when just one nonempty set X is considered. One can ask two very different questions about X:

1. Does there exist an element (equivalently, a singleton subset) of X?
2. Can one find or identify an explicit element of X?

The answer to Question 1 is trivially yes: this is the very definition of the word "nonempty." The Axiom of Choice in this setting simply asserts a positive answer to Question 1; thus the AC is trivially true for a collection consisting of one nonempty set. (It is also true, only slightly less trivially, for a collection of a finite number of nonempty sets.) But Question 2 is definitely nontrivial, and is more of a philosophical than a mathematical question. For example, suppose one proves that $\mathbb{R}$ is uncountable (II.5.3.6.) and that the set of algebraic numbers is countable (II.5.6.6.). These results together prove that the set of transcendental numbers is nonempty (and, in fact, uncountable). However, just knowing these results one would be very hard pressed to exhibit a transcendental number. (Actually, in this instance, the standard proofs of the assertions could in principle be used to give an extremely cumbersome construction of an explicit transcendental number. There are, of course, much better ways of identifying explicit transcendental numbers, e.g. III.8.2.35., III.9.3.1., III.9.3.2..) For a cleaner analogy, let X be the set of choice functions for a specific indexed collection of nonempty sets: the AC asserts a positive answer to Question 1 (i.e. that X is nonempty), not Question 2.

As pointed out by J. HADAMARD [Moo82,], the situation is very similar to the question of whether we allow consideration of functions from $\mathbb{R}$ to $\mathbb{R}$ for which we have no formula or effective way of calculating the values. Indeed, until RIEMANN in the mid-19th century, the term "function" was restricted to functions with a closed-form formula, or at least a power-series representation (i.e. analytic functions). The modern definition of function as simply a set of ordered pairs without necessarily a specified rule or method for calculation of its values would have been much too broad to be acceptable 200 years ago.

The Axiom of Choice is a pure existence assertion, not a constructive one.

II.6.2. Equivalent Formulations

One simple reformulation of the AC is to require the sets to be pairwise disjoint:

II.6.2.1. **Axiom of Choice:** (Disjoint Indexed Form) If $\{X_i : i \in I\}$ is an indexed collection (set) of pairwise disjoint nonempty sets, then there is a choice function for $\{X_i : i \in I\}$.

II.6.2.2. It is obvious that the Indexed Form of the AC implies the Disjoint Indexed Form. Conversely, if $\{X_i : i \in I\}$ is an indexed collection (set) of nonempty sets set $X = \cup_{i \in I} X_i$ and

$$\tilde{X}_i = \{(x, i) : x \in X_i\} \subseteq X \times I$$

for each $i \in I$. Then $\{\tilde{X}_i : i \in I\}$ is an indexed set of pairwise disjoint nonempty sets, and a choice function for the $\tilde{X}_i$ gives a choice function for the X_i in the obvious way.

Equivalently, this version can be rephrased:

II.6.2.3. **Axiom of Choice:** (Equivalence Relation Form) Let X be a set, and $\sim$ an equivalence relation on X. Then there is a subset of X containing exactly one element from each $\sim$-equivalence class.

The Axiom of Choice has many other equivalent reformulations, some of which seem intuitively plausible and others intuitively implausible. This situation reflects the unexpectedly subtle nature of the Axiom of Choice; it must be kept in mind that it is meaningless (within ZF) to ask whether any of these versions of the Axiom are "true".

II.6.2.4. An obvious reformulation which seems plausible concerns Cartesian products. A choice function for $\{X_i : i \in I\}$ is just an element of $\prod_{i\in I} X_i$. Thus the Axiom of Choice is equivalent to:

Multiplicative Axiom: If $\{X_i : i \in I\}$ is an indexed collection of nonempty sets, then $\prod_{i\in I} X_i$ is not empty.

Bertrand Russell, who first formulated the Multiplicative Axiom version (he phrased it somewhat differently since infinite Cartesian products were not defined at the time), regularly called the Axiom of Choice by this name.

II.6.2.5. One potential way of generating a choice function is by transfinite induction (recursion). Thus the Axiom of Choice appears to be related to well ordering. In fact, the Axiom of Choice is equivalent to the Well-Ordering Principle:

Well-Ordering Principle: Every set can be well ordered.

II.6.2.6. To see that the Well-Ordering Principle implies the Axiom of Choice, suppose $\{X_i : i \in I\}$ is an indexed collection of nonempty sets, and fix a well ordering $\leq$ on I and a well ordering $\precsim$ on $X = \cup_{i\in I} X_i$. For $i \in I$, let R_i be the rule "Let $c(i)$ be the first element of $(X, \precsim)$ which is in X_i." Then a choice function c is defined by transfinite induction (recursion) ().

The proof that the Axiom of Choice implies the Well-Ordering Principle uses the notion of ordinals, and will be discussed in (). In fact, the Axiom of Choice was first formulated by E. Zermelo specifically to prove the Well-Ordering Principle.

Zorn's Lemma and the Hausdorff Maximal Principle

There are two similar statements about partially ordered sets which are also equivalent to the Axiom of Choice. To describe them simply, we need a definition:

II.6.2.7. Definition. Let $(X, \precsim)$ be a partially ordered set. A *chain* in $(X, \precsim)$ is a subset of X which is totally ordered in the induced ordering. A *maximal chain* is a chain which is not properly contained in any larger chain.

II.6.2.8. **Hausdorff Maximal Principle:** Let $(X, \precsim)$ be a partially ordered set. Then $(X, \precsim)$ contains a maximal chain.

II.6.2.9. **Zorn's Lemma:** Let $(X, \precsim)$ be a partially ordered set. If each chain in $(X, \precsim)$ has an upper bound in X, then X contains a maximal element.

Zorn's Lemma is called a "Lemma" for historical reasons, but it is in fact equivalent to the Axiom of Choice. A name like *Zorn's Principle* might be more suitable from a logical standpoint. The Hausdorff Maximal Principle is also sometimes called *Kuratowski's Lemma* (HAUSDORFF and KURATOWSKI independently studied essentially this condition, with nuances of differences; the Hausdorff Maximal Principle was originally stated only for sets ordered by inclusion.)

II.6.2.10. It is fairly easy to see that the Hausdorff Maximal Principle and Zorn's Lemma are equivalent. To see that the Hausdorff Maximal Principle implies Zorn's Lemma, let $(X, \precsim)$ be a partially ordered set in which every chain has an upper bound. Assume the Hausdorff Maximal Principle and let C be a maximal chain in X. Then C has an upper bound x, and we must have $x \in C$ since otherwise $C \cup \{x\}$ would be a chain properly containing C. Thus x is the largest element of C. Then x must be a maximal element of X, since if there were a $y \in X$ with $x \prec y$, then $C \cup \{y\}$ would be a chain properly containing C.

II.6.2.11. The proof that Zorn's Lemma implies the Hausdorff Maximal Principle is a good example of how Zorn's Lemma is used in practice. Let $(X, \precsim)$ be a partially ordered set. Let $\mathcal{C}$ be the set of all chains in $(X, \precsim)$; then $\mathcal{C}$ is partially ordered by inclusion. If $\{C_i : i \in I\}$ is a chain in $(\mathcal{C}, \subseteq)$, i.e. each C_i is a chain in $(X, \precsim)$ and for any $i, j \in I$, one of C_i and C_j is contained in the other, then it is easy to verify that $\cup_{i\in I} C_i$ is a chain in X: if $x, y \in \cup_{i\in I} C_i$, then $x \in C_i$, $y \in C_j$ for some $i, j \in I$; but either $C_i \subseteq C_j$ or vice versa, so both x and y are in either C_i or C_j, so they are comparable since both C_i and C_j are chains. Also, $\cup_{i\in I} C_i$ contains each of the C_i; so $\cup_{i\in I} C_i$ is an upper bound for $\{C_i : i \in I\}$ in $(\mathcal{C}, \subseteq)$. Thus by Zorn's Lemma $(\mathcal{C}, \subseteq)$ has a maximal element C, which is a maximal chain in $(X, \precsim)$.

II.6.2.12. In applications of Zorn's Lemma, as here, the Lemma is frequently applied to a judiciously constructed new partially ordered set. Another relevant example is the proof that Zorn's Lemma implies the Well-Ordering Principle, which is left as an Exercise (Exercise ().) See Exercises (), () for other applications of Zorn's Lemma.

II.6.2.13. The proof that the Axiom of Choice, the Well-Ordering Principle, the Hausdorff Maximal Principle, and Zorn's Lemma are equivalent is completed by showing that the Axiom of Choice implies Zorn's Lemma. This argument uses ordinals, and will be discussed in ().

Tukey's Lemma

Another variation of AC similar to Zorn's Lemma and the Hausdorff Maximal Principle is Tukey's Lemma, often (and properly) called the *Teichmüller-Tukey Lemma* since it was independently formulated by J. TUKEY and O. TEICHMÜLLER; but I don't feel guilty slighting TEICHMÜLLER in the name since he was a Nazi.

II.6.2.14. DEFINITION. A collection $\mathcal{F}$ of sets is of *finite character* if a set X is in $\mathcal{F}$ if and only if every finite subset of X is in $\mathcal{F}$.

A good example of a collection of finite character is the collection of all linearly independent subsets of a vector space V, since a set of vectors in V is linearly independent if and only if every finite subset is linearly independent (since a linear combination involves only finitely many vectors).

II.6.2.15. Tukey's Lemma: If $\mathcal{F}$ is a collection of sets of finite character, then $\mathcal{F}$ has a maximal element under inclusion, i.e. there is an $X \in \mathcal{F}$ not properly contained in any other set in $\mathcal{F}$.

II.6.2.16. Tukey's Lemma is an easy consequence of Zorn's Lemma, since if $\mathcal{F}$ is a collection of sets of finite character, regarded as a partially ordered set under inclusion, and $\mathcal{C} = \{C_i : i \in I\}$ is a chain in $\mathcal{F}$, then $C = \cup_{i \in I} C_i$ is in $\mathcal{F}$ since every finite subset of C is contained in some C_i, and C is an upper bound for $\mathcal{C}$.

II.6.2.17. It is also not difficult to show that Tukey's Lemma implies the Axiom of Choice (Exercise II.6.4.11.).

II.6.2.18. The Axiom of Choice is also equivalent to Tikhonov's Theorem: an arbitrary product of compact topological spaces is compact in the product topology. This theorem will be discussed in ().

II.6.2.19. It must be pointed out that the Axiom of Choice is not universally accepted by mathematicians, even by analysts. This was particularly true in the early 20th century when the foundations of mathematics were being axiomatized; see [Sie65] and [Moo82] for interesting discussions of the points of view of leading mathematicians of the time. Today most analysts are more comfortable with the Axiom of Choice. But there is a branch of analysis called *Constructive Analysis*, where the full Axiom of Choice is not and cannot be assumed. There are results of both theoretical and practical importance in Constructive Analysis, but this theory is a separate subject from the one treated in this book. See e.g. [BB85] and [BV06]. There is also an axiom called the *Axiom of Determinacy* which is sometimes considered, which is inconsistent with the Axiom of Choice. A weaker version of this axiom, called the *Axiom of Projective Determinacy*, not inconsistent with the Axiom of Choice, plays a role in Descriptive Set Theory ().

II.6.3. Other Forms of Choice

II.6.3.1. Weaker versions of the Axiom of Choice, which are somewhat more palatable to some mathematicians (see [Pot04, p. 164-165, 240] for some relevant philosophical discussion), are sometimes assumed. One is the *Countable Axiom of Choice* (or Axiom of Countable Choice), written *Countable AC* or sometimes AC_ω, abbreviated CC, which is also not provable or disprovable in ZF:

Countable Axiom of Choice: (Indexed Form) If $\{X_n : n \in \mathbb{N}\}$ is a sequence of nonempty sets, then there is a choice function for $\{X_n : n \in \mathbb{N}\}$.

Note that it is the index set, not the X_n, which is required to be countable. It is a quite different matter to weaken the AC by requiring the X_i to be countable, or finite, with no restriction on the index set; see II.6.3.13.. There is no implication either way between the CC and these forms of AC.

II.6.3.2. Analysis in its usual form is impossible without at least the CC. The following statements, which can be readily proved using CC, cannot be proved in ZF [Jec73]:

(i) A countable union of countable sets is countable (Exercise II.6.4.8.).

(ii) $\mathbb{R}$ is not a countable union of countable sets.

(iii) Every infinite set contains a countably infinite subset (Exercise II.6.4.9.).

(iv) In $\mathbb{R}$, or in metric spaces, the sequential criterion for limits and continuity is equivalent to the $\epsilon - \delta$ definition (). [But see II.6.4.6. and VI.2.4.2..]

(v) In $\mathbb{R}$, or in metric spaces, a point p is in the closure of a subset A if and only if there is a sequence of points of A converging to p ().

(vi) Every second countable topological space (in particular, every subset of $\mathbb{R}$) is separable ().

II.6.3.3. It turns out that a slightly stronger axiom called the *Axiom of Dependent Choice* is often more useful:

Axiom of Dependent Choice: Let R be a relation on a set X. If, for every $x \in X$, there is a $y \in X$ with xRy, then there is a sequence (x_n) in X with $x_n R x_{n+1}$ for all n.

Roughly speaking, the Axiom of Dependent Choice (abbreviated DC) asserts that a sequence of choices can be made, each depending on the previous choice(s).

II.6.3.4. The Axiom of Choice implies the Axiom of Dependent Choice, which implies the Countable Axiom of Choice (Exercise II.6.4.3.); neither of the implications can be reversed. See [**?**] for a discussion of the Axiom of Dependent Choice. In some of the applications of the Axiom of Choice in this book, one can get away with the Axiom of Dependent Choice or even the Countable AC; we will try to identify most of these (see Exercise () for one example.)

II.6.3.5. Note that the Axiom of Dependent Choice is not needed (in fact no form of Choice is needed) to prove the existence of sequences defined by recursion (II.4.2.16.).

The Boolean Prime Ideal Theorem

Another AC-type statement which has several important and useful formulations is the Boolean Prime Ideal Theorem, abbreviated BPI. We give several equivalent formulations of this statement, without, however, explaining all the terms involved (see, for example, [Jec73]). For part of the proof, see XIII.11.9.19..

II.6.3.6. Boolean Prime Ideal Theorem (BPI): Every ideal in a Boolean algebra is contained in a prime ideal.

II.6.3.7. Compactness Theorem of Model Theory: Let Σ be a set of sentences in a first-order theory. If every finite subset of Σ has a model, then Σ has a model.

II.6.3.8. **Ultrafilter Property:** Every filter on a set is contained in an ultrafilter (XVI.1.8.15.).

II.6.3.9. **Tikhonov's Theorem for Hausdorff Spaces:** An arbitrary product of compact Hausdorff spaces is compact in the product topology (XIII.11.6.6.).

II.6.3.10. **Stone-Čech Compactification Theorem:** Every completely regular topological space has a Stone-Čech compactification (XIII.11.9.13.).

II.6.3.11. THEOREM. The following are equivalent in ZF set theory:

(i) The Boolean Prime Ideal Theorem.

(ii) The Compactness Theorem of Model Theory.

(iii) The Ultrafilter Property.

(iv) Tikhonov's Theorem for Hausdorff spaces.

(v) The Stone-Čech Compactification Theorem.

II.6.3.12. It is easy to deduce the Ultrafilter Property from Zorn's Lemma (it is also not difficult to obtain the BPI directly from Zorn's Lemma, and it is trivial to obtain the Hausdorff Tikhonov Theorem from the general one), and hence the AC implies the BPI. The converse is false, i.e. the BPI is a strictly weaker assumption than AC. In fact:

II.6.3.13. THEOREM. The following statements are successively strictly logically weaker in ZF set theory, i.e. each can be proved from the previous one in ZF, and none follows in ZF from the succeeding ones (and even (v) is not a theorem of ZF for $n \geq 2$):

(i) The Axiom of Choice.

(ii) The Boolean Prime Ideal Theorem.

(iii) The *Ordering Principle*: Every set can be totally ordered.

(iv) The *Axiom of Choice for Finite Sets*: Every collection of nonempty finite sets has a choice function.

(v) For each fixed $n \in \mathbb{N}$, the Axiom of Choice for n-element sets.

See Exercise II.6.4.10. for (iii) $\Rightarrow$ (iv), and [Jec73] for a complete discussion including the failure of the converse implications, and other variations of AC. Here are some selected additional facts, discussed in [Jec73]:

II.6.3.14. DC does not imply (v), hence neither DC nor CC implies any of the stronger statements. There is no implication the other way between any of (ii)–(v) and either the CC or DC. (This last question is not discussed in [Jec73], but see [HR98].)

II.6.3.15. The statement that there exists a free ultrafilter on every infinite set is strictly weaker than the Ultrafilter Property. The existence of a free ultrafilter on $\mathbb{N}$ (an assumption underlying nonstandard analysis) is weaker still.

II.6.3.16. There is another pair of equivalent statements which are important in analysis, which are strictly weaker than the BPI: the Hahn-Banach Theorem () and the Boolean Measure Theorem (). These statements are independent of CC and DC and of II.6.3.13.(iv)–(v), and apparently independent of II.6.3.13.(iii).

Axiom of Global Choice

II.6.3.17. The AC is a statement about sets. In ZF set theory, there are nothing but sets. However, in NBG set theory there are also proper classes, and one could ask more generally whether Choice extends to proper classes:

II.6.3.18. **Axiom of Global Choice:** if C is a proper class of pairwise disjoint nonempty sets, then there is a proper class containing exactly one element from every set in C.

II.6.3.19. The Axiom of Global Choice (abbreviated AGC) obviously implies the AC, since a set is a proper class. (It is strictly stronger: if NBG is consistent, there are models of NBG in which the AC holds but the AGC does not.) Thus the AGC can be called "the Axiom of Choice on steroids."

What Gödel actually proved is that if ZF set theory is consistent, hence NBG set theory is also consistent, then NBG+(AGC) is consistent.

II.6.3.20. Just as the AC implies the Well-Ordering Principle, the AGC implies the *Global Well-Ordering Principle:* every proper class can be well ordered.

See [RR85] for a compilation of many other statements equivalent to the AC or other versions of Choice, and [HR98] for an enormous variety of Choice statements of mostly strictly varying logical strength. The fact that there are so many models of set theory in which some of these statements are true and others false tends to seriously undercut the credibility of set theory in its present form as a satisfactory foundation for all of mathematics.

The Axiom of Choice has been controversial, and remains somewhat so; however, most modern mathematicians, especially analysts, have no serious qualms about accepting and using it, often without comment. The attitude of R. Boas [Boa96, p. xi] is widespread:

> "[A]fter Gödel's results, the assumption of the axiom of choice can do no mathematical harm that has not already been done."

In this book we will feel free to use the AC, but we will try to be careful to point out what version of Choice is used and really needed whenever it is employed. Analysis as we will develop it would not be the same, or even very similar, without it.

> "Rarely have the practitioners of mathematics, a discipline known for the certainty of its conclusions, differed so vehemently over one of its central premises as they have done over the Axiom of Choice. Yet without the Axiom, mathematics today would be quite different. The very nature of modern mathematics would be altered and, if the Axiom's most severe constructivist critics

prevailed, mathematics would be reduced to a collection of algorithms. Indeed, the Axiom epitomizes the fundamental changes – mathematical, philosophical, and psychological – that took place when mathematicians seriously began to study infinite collections of sets."

Gregory H. Moore[11]

MOORE'S book [Moo82] gives a thorough treatment of the history of the Axiom of Choice and its place in modern mathematics.

"The infinite comes into the proof theory as soon as we introduce the ideas 'all' and 'there exists.' The former is equivalent to an infinite 'and,' and the latter to an infinite 'or.' ... But if we thoughtlessly apply to infinite totalities a procedure which is permissible in the finite case, then we open the gate and door to errors. It is the same source of error that we see in analysis; in the latter field the carrying over, to infinite sums and products, of theorems which are valid for the corresponding finite sums and products is permissible only if special convergence conditions, etc., are satisfied. Similarly we cannot treat the infinite sums and products in the same manner as the finite, unless our proof theory reveals such treatment to be justified."

D. Hilbert[12]

II.6.4. Exercises

II.6.4.1. Show that a totally ordered set is well ordered if and only if it contains no strictly decreasing sequences. (The Axiom of Dependent Choice is needed for one direction.)

II.6.4.2. Let X be a set. Let $\mathcal{W}$ be the set of all pairs $(A, \leq)$, where $A \subseteq X$ and $\leq$ is a well ordering on A. Define a partial order $\precsim$ on $\mathcal{W}$ by setting $(A, \leq_A) \precsim (B, \leq_B)$ if $A \subseteq B$, A is an initial segment of B under $\leq_B$, and the restriction of $\leq_B$ to A is $\leq_A$. (Note that $\mathcal{W} \neq \emptyset$ if $X \neq \emptyset$ since every singleton subset of X has an obvious well ordering.)

(a) Show that if $\{(A_i, \leq_i) : i \in I\}$ is a chain in $\mathcal{W}$, then $(A, \leq)$ is an upper bound, where $A = \cup_i A_i$ and $\leq = \leq_i$ on A_i. [It must be shown that $\leq$ is well defined on A and is a well ordering, and that each A_i is an initial segment.]

(b) If $(A, \leq)$ is any element of $\mathcal{W}$, and $x \in X \setminus A$, extend $\leq$ to $A \cup \{x\}$ by setting $a \leq x$ for all $a \in A$. Show that $(A \cup \{x\}, \leq) \in \mathcal{W}$.

(c) If $(A, \leq)$ is a maximal element of $\mathcal{W}$, then $A = X$.

(d) Conclude that Zorn's Lemma implies the Well-Ordering Principle.

II.6.4.3. (a) Prove that the Axiom of Choice implies the Axiom of Dependent Choice. [If R is a relation on X, let $c : \mathcal{P}(X) \setminus \{\emptyset\} \to X$ be a choice function, and set $x_{n+1} = c(\{y \in X : x_n R y\})$.]

(b) Prove that the Axiom of Dependent Choice implies the Countable Axiom of Choice. [If $\{X_n : n \in \mathbb{N}\}$ is an indexed collection of nonempty sets, let Z be the set of all ordered pairs (n, x), where $n \in \mathbb{N}$ and $x \in X_n$. Let $(n, x) R (m, y)$ if and only if $m = n + 1$.]

[11] [Moo82, p. 1]
[12] *Die logischen Grundlagen der Mathematik*, p. 154-155.

II.6.4.4. Using the Axiom of Dependent Choice, prove that every infinite set X contains a sequence (x_n) of distinct elements (), i.e. there is an injective function from $\mathbb{N}$ to X. [Let Z be the set of ordered pairs (n, A), where $n \in \mathbb{N}$ and A is an infinite subset of X. Define $(n, A)R(m, B)$ if and only if $m = n+1$, $B \subseteq A$, and $A \setminus B$ contains exactly one element.] See II.6.4.9. for a slightly better statement.

II.6.4.5. Identify how the proof of II.5.4.8. uses the Countable Axiom of Choice. [In fact, II.5.4.8. is not a theorem of ZF [], i.e. some type of Choice Axiom is necessary.]

II.6.4.6. [Her06, 3.15] Show that even without CC a function from $\mathbb{R}$ to $\mathbb{R}$ is sequentially continuous (i.e. sequentially continuous at every point) if and only if it is $(\epsilon$–$\delta)$-continuous. [If $x \in \mathbb{R}$, show that sequential continuity and $(\epsilon$–$\delta)$-continuity are the same on $\mathbb{Q} \cup \{x\}$.] Contrast this with the fact that the CC is needed to prove that the two types of continuity at a point are equivalent. See VI.2.4.2..

II.6.4.7. ([Bla77]; cf. [Oxt71, p. 95]) Show that the Axiom of Dependent Choice is equivalent to the statement that if (X_n) is a sequence of nonempty sets, then $\prod_n X_n$ is a Baire space, where each X_n is given the discrete topology, as follows.

(a) Show that the product topology on $\prod_n X_n$ is given by a complete metric (cf. XIII.10.1.8., but note that the CC is not needed in this case).

(b) Let R be a relation on a nonempty set X, such that for every $x \in X$ there is a $y \in X$ with xRy. For each n let

$$U_n = \{f \in X^{\mathbb{N}} : f(n)Rf(m) \text{ for some } m > n\} .$$

Show that U_n is a dense open subset of $X^{\mathbb{N}}$. [Write

$$U_n = \bigcup_{m>n} \bigcup_{(a,b)\in R} \{f \in X^{\mathbb{N}} : f(n) = a, f(m) = b\} .]$$

(c) Let $g \in \cap_{n=1}^{\infty} U_n$. Let $k_1 = 1$, $x_1 = g(1)$, and inductively let k_{n+1} be the smallest $m > k_n$ for which $g(k_n)Rg(m)$, and $x_{n+1} = g(k_{n+1})$.

(d) Conclude that the Baire Category Theorem (BC) implies the Axiom of Dependent Choice (DC). Since the proof of (BC) uses only (DC) (along with ZF), ZF+(BC) is logically equivalent to ZF+(DC).

(e) Show that under the hypotheses of DC, the sequence (x_n) can be chosen with x_1 arbitrary. [$\cap_{n=1}^{\infty} A_n$ is dense in $X^{\mathbb{N}}$ by (BC).]

(f) Show that the Countable Axiom of Choice is equivalent to the statement that $\prod_n X_n$ is nonempty for any sequence (X_n) of nonempty sets.

II.6.4.8. Show how the CC is used in the proof of ().

II.6.4.9. Using CC, show that every infinite set contains a countably infinite subset. [A set X is infinite if and only if, for each $n \in \mathbb{N}$, the collection $\mathcal{P}_n(X)$ of n-element subsets of X is nonempty. Choose such a subset for each n and consider their union.]

II.6.4.10. Show that the Ordering Principle implies the Axiom of Choice for Finite Sets. [If X is totally ordered, every finite subset of X has a smallest element.]

II.6.4.11. Let $\{X_i : i \in I\}$ be an indexed collection of nonempty sets, and $\mathcal{F}$ the set of all choice functions for subcollections of the form $\{X_i : i \in J\}$ for $J \subseteq I$. Each element of $\mathcal{F}$ is a set of ordered pairs of the form (i, x), where $i \in I$ and $x \in X_i$. Regard $\mathcal{F}$ as a partially ordered collection of sets under inclusion.

(a) Show that $\mathcal{F}$ is nonempty. [Consider choice functions on finite subsets of I.]

(b) Show that $\mathcal{F}$ is a collection of sets of finite character.

(c) Show that if f is a maximal element of $\mathcal{F}$, then the domain of f is all of I. [Show that a choice function defined on a proper subset of I can be extended to one more index.]

(d) If Tukey's Lemma is assumed, conclude that $\{X_i : i \in I\}$ has a choice function, i.e. that AC holds.

II.6.4.12. (cf. III.3.7.4.) Let R be a unital ring (III.3.6.3.). A nonempty subset I of R is a *(two-sided) ideal* if I is closed under addition and $x \in I$, $y \in R$ imply $xy, yx \in I$. An ideal is *proper* if it is a proper subset of R. A *maximal ideal* is a proper ideal which is not contained in any strictly larger proper ideal.

(a) Show that $\{0\}$ is a (proper) ideal in R. More generally, if $S \subseteq R$, then the set of finite sums of elements of R of the form xzy, where $z \in S$ and $x, y \in R$, is an ideal of R (not necessarily proper), and is the smallest ideal of R containing S, called the *ideal generated by* S, usually denoted (S).

(b) Show that if I is an ideal in R, then $0 \in I$, and $x \in I$ implies $-x \in I$. [Use $-1 \in R$.]

(c) Show that an ideal I in R is proper if and only if $1 \notin I$.

(d) Show that any intersection of ideals in R is an ideal in R.

(e) Show that a union of ideals is not an ideal in general. But the union of a *chain* of ideals (under inclusion) is an ideal. If I and J are ideals, the ideal generated by $I \cup J$ is

$$I + J = \{x + y : x \in I, y \in J\} .$$

(f) Use Zorn's Lemma to prove that every proper ideal in a unital ring R is contained in a maximal ideal.

(g) Show that if I and J are ideals in R, then the set IJ of finite sums of elements of the form xy, where $x \in I$ and $y \in J$, is an ideal of R contained in $I \cap J$.

(h) A proper ideal P of R is *prime* if, whenever I and J are ideals of R with $IJ \subseteq P$, then $I \subseteq P$ or $J \subseteq P$. Show that every maximal ideal is a prime ideal. [If M is a maximal ideal and I and J are ideals with $IJ \subseteq M$ and $I \not\subseteq M$, then $I + M = R$, so $1 \in I + M$.]

Ideals can be defined also in nonunital rings (although (b) is not automatic and must be included as part of the definition of ideal; cf. III.3.7.4.). However, a nonunital ring need not contain any maximal ideals, e.g. there can be an increasing sequence of proper ideals whose union is the whole ring.

The term "ideal" arose since ideals were first considered in rings of algebraic numbers (). Every element x in a ring R generates an ideal $(x) = (\{x\})$; such ideals are called *principal ideals*. In many rings such as $\mathbb{Z}$ or a polynomial ring in one variable over a field, every ideal is principal, so there is no real distinction between ideals and elements (or equivalence classes of elements obtained by multiplication by units, e.g. $(n) = (-n)$ in $\mathbb{Z}$). But in some rings of algebraic integers, not every ideal is principal, and in these rings arithmetic is more complicated. KUMMER had the idea to add "ideal numbers" (i.e. nonprincipal ideals), and to consider the arithmetic of ideals, which turns out to be considerably simpler than arithmetic of numbers, e.g. every ideal in a ring of algebraic integers factors uniquely as a product of prime ideals. See books on Abstract Algebra or Algebraic Number Theory for details.

II.6.4.13. Use the AC and Cantor's diagonalization argument to prove the following:

THEOREM. Let X and Y be sets with $card(X) = card(Y) = \kappa$. Then there is a set $\mathcal{F}$ of functions from X to Y with $card(\mathcal{F}) > \kappa$ such that any two functions in $\mathcal{F}$ agree only on a set of cardinality less than κ.

(a) Use Zorn's Lemma to obtain a maximal set $\mathcal{F}$ with the given property.

(b) Suppose $\mathcal{F}$ is equipotent with a subset Z of X. Index $\mathcal{F}$ by Z, i.e. $\mathcal{F} = \{f_z : z \in Z\}$.

(c) Well order X by the first ordinal of cardinality κ. Define $f : X \to Y$ as follows. For each $x \in X$, the cardinality of $\{z \in Z : z < x\}$ is less than κ, so there is a $y \in Y$ with $y \neq f_z(x)$ for all $z < x$. Choose such a y and call it $f(x)$.

(d) If $z \in Z$, then $f(x) \neq f_z(x)$ if $x > z$, so $\{x : f(x) = f_z(x)\} \subseteq \{x : x \leq z\}$ has cardinality less than κ. Thus $\mathcal{F} \cup \{f\}$ has the given property, contradicting maximality of $\mathcal{F}$.

(e) Show that $\mathcal{F}$ can be chosen to contain all constant functions. So the nonconstant functions in $\mathcal{F}$ take each value in Y only on a set of cardinality less than κ, and themselves form a set of cardinality greater than κ.

(f) Can $\mathcal{F}$ be chosen so that each $f \in \mathcal{F}$ is injective? Bijective? Can $\mathcal{F}$ be chosen to have cardinality $\kappa^\kappa = 2^\kappa$ (cf. II.9.9.3.(e))?

(g) Here is an alternate argument in the case $\kappa = \aleph_0$. Identify X with $\mathbb{N}$ and Y with $\mathbb{Q}$. For each real number $r \in \mathbb{R}$, choose a sequence of (distinct) rational numbers converging to r. One obtains $2^{\aleph_0}$ sequences of rational numbers, i.e. (injective) functions from $\mathbb{N}$ to $\mathbb{Q}$, and any two of the sequences can have only finitely many terms in common since they converge to different limits. Can this argument be adapted to general κ?

II.6.4.14. The Multiplicative Axiom Revisited. Let $\{X_i : i \in I\}$ be an indexed family of nonempty sets, and $X = \prod_{i \in I} X_i$.

(a) If $X \neq \emptyset$, show (in ZF, i.e. without any form of Choice) that $X_J \neq \emptyset$ for any $J \subseteq I$, where

$$X_J = \prod_{i \in J} X_i \ .$$

[Use the natural one-one correspondence $X \cong X_J \times X_{I \setminus J}$.] In fact, $X \neq \emptyset$ if and only if both X_J and $X_{I \setminus J}$ are nonempty.

(b) If $X \neq \emptyset$ and $J \subseteq I$, show that the natural projection $\pi_J : X \to X_J$ is surjective. In particular, if $i \in I$, the coordinate projection $\pi_i : X \to X_i$ is surjective.

(c) If there is any choice function for $\{X_i : i \in I\}$, there is one with any finite number of values chosen arbitrarily. Thus if $X \neq \emptyset$, it is fairly large (infinite if infinitely many X_i are not singletons).

(d) Suppose there is an indexed collection $\{Y_i : i \in I\}$ of pairwise disjoint sets with each Y_i nonempty but $\prod_i Y_i = \emptyset$ (there will be such a collection in the absence of AC). Let p be an object (set) such that neither p nor $\{p\}$ is in any Y_i, e.g. $p = \cup_i Y_i$, and for each i let $q_i = (p, i) = \{\{p\}, \{p, i\}\}$. Then $q_i \notin Y_j$ for any i and j. Set $X_i = Y_i \cup \{q_i\}$ for each i. Then the X_i are pairwise disjoint and not singletons, and $X = \prod_{i \in I} X_i \neq \emptyset$, since $f(i) = q_i$ $(i \in I)$ is a choice function. But if g is any choice function (element of X), then $g(i) = q_i$ for all but finitely many i. So X is "not very large."

II.6.4.15. Say that an $(n+1)$-ary relation R on a set X (i.e. $R \subseteq X^{n+1}$) is *total* if for every $x_1, \ldots, x_n \in X$ there is a $y \in X$ such that $(x_1, \ldots, x_n, y) \in R$. Consider the following axiom:

Relational Axiom of Choice (RAC): For every X, every $n \in \mathbb{N}$, and every total $(n+1)$-ary relation R on X, there is an n-ary function $f : X^n \to X$ such that, for all $x_1, \ldots, x_n \in X$, $(x_1, \ldots, x_n, f(x_1, \ldots, x_n)) \in R$.

For $n \in \mathbb{N}$ define (RAC_n) to be the same statement for n fixed.

(a) Show that (RAC) $\Rightarrow$ $(\mathrm{RAC}_{n+1}) \Rightarrow (\mathrm{RAC}_n)$ for any $n \in \mathbb{N}$.

(b) Show that (AC) $\Rightarrow$ (RAC). [If X is a nonempty set, let c be a choice function for $\mathcal{P}(X) \setminus \{\emptyset\}$, and set $f(x_1, \ldots, x_n) = c(\{y : (x_1, \ldots, x_n, y) \in R\})$.]

(c) Show that $(\mathrm{RAC}_1) \Rightarrow$ (AC) as follows. Let X be a nonempty set, and fix $x_0 \in X$. Let $Y = \mathcal{P}(X) \setminus \{\emptyset\}$. Define a binary relation R on $X \times Y$ by

$$(x_1, A)R(x_2, B) \text{ [i.e. } ((x_1, A), (x_2, B)) \in R] \text{ if and only if } x_2 \in A \ .$$

Let $f : X \times Y \to X \times Y$ be a function such that $(x, A)R(f(x, A))$ for all $x \in X$ and $A \in Y$. For $A \in Y$ set $c(A) = \pi_X(f(x_0, A))$, where $\pi_X : X \times Y \to X$ is projection onto the first coordinate.

(d) Conclude that (RAC) and (RAC_n) for every n are all equivalent to (AC).

II.7. Axiomatic Set Theory

> "[A]xiomatic set theory is often viewed as a *foundation of mathematics*: it is alleged that all mathematical objects are sets, and their properties can be derived from the relatively few and elegant axioms about sets. Nothing so simple-minded can be quite true, but there is little doubt that in standard, current mathematical practice, 'making a notion precise' is essentially synonymous with 'defining it in set theory.' "
>
> *Y. Moschovakis*[13]

The first set of axioms for modern set theory was proposed by Ernst Zermelo in 1908, although closely related work had been done previously by Cantor, Russell, and several others. The axioms were refined by Abraham Fraenkel (in part using ideas of Thoralf Skolem), and are today known as the *Zermelo-Fraenkel Axioms* and the theory is called *ZF set theory*. An additional axiom proposed by John von Neumann (the Axiom of Regularity) is usually included. Most of Zermelo's axioms were not controversial, but one of them was: the Axiom of Choice, discussed in () (the Axiom of Regularity is also somewhat controversial); the Axiom of Choice is customarily not included among the ZF axioms, and the theory with the Axiom of Choice included is called *ZFC set theory*. Other axiom systems, most notably the ones developed by von Neumann, Paul Bernays, and Kurt Gödel, and Nicolas Bourbaki[14], have been proposed, but the principal ones turn out to be equivalent to ZF theory in the sense that exactly the same theorems about sets can be proved. The von Neumann-Bernays-Gödel (NBG) axioms are generally similar to the ZF axioms except that classes which are not sets appear explicitly.

The framework of set theory is two undefined concepts: *set* and *membership*. All objects of set theory are sets (or sometimes *classes* which are too large to be sets, and *atoms* which are not sets are sometimes allowed); the elements of a set (or class) are also sets (or atoms). Thus, to encode the rest of mathematics into set theory, other mathematical objects must be defined as sets of some kind. For example, an ordered pair (x, y) is defined to be $\{\{x\}, \{x, y\}\}$; thus relations and functions can be defined as sets of ordered pairs, i.e. sets of sets, and therefore as sets. In II.7.1.35., we describe how to define the natural numbers as sets, and then $\mathbb{Z}$ and $\mathbb{Q}$ can be described as equivalence classes of ordered pairs of natural numbers. A description of real numbers as sets of rational numbers is given in III.5..

The basic axioms of set theory assert that the usual operations give sets. The Axiom of Infinity asserts that the natural numbers (as constructed in II.7.1.35.) are sets and that the collection of all natural numbers is a set. The axioms of Union and Power Set assert that the union of the elements of a set x is a set, written $\cup x$ (remember that the elements of a set are themselves sets, so this union makes sense), and that $\mathcal{P}(x)$ is a set. Other axioms assert that any subset of a set is a set, and that the range of a function is a set. The axioms are inspired by the following somewhat imprecise principle, first espoused by Cantor:

Basic Principle of Set Theory: Every collection of objects [sets] is regarded as a set unless there is some logical contradiction in allowing it.

Thus, for example, Russell's Paradox disqualifies allowing "the set of all sets."

[13] [**?**, p. vii].

[14] The Nicolas Bourbaki of mathematics was not an actual person, but the *nom de plume* of a gradually changing consortium of leading French mathematicians whose goal has been to systematize all of mathematics. See [**?**], [**?**], or [**?**] for a history of this interesting project.

II.7.1. The ZF Axioms

The Zermelo-Fraenkel axioms appear in the literature in various equivalent forms. We will give one convenient version, and not worry too much about minimality of the set of axioms (some axioms can be deduced from the others; see e.g. II.7.1.42.). Most of the axioms are no-brainers. Some of the axioms have standard names which are of somewhat obscure meaning. We will state the axioms in somewhat casual form in words, and also give a careful statement in the object language, which is the language of predicate calculus () with one additional symbol $\in$.

Since the only objects considered are sets, there is no reason to distinguish notationally between sets and their members (which are also sets). Thus we will use lower-case letters to denote sets.

II.7.1.1. Axiom of Extensionality: If x and y have the same elements, then $x = y$. Formally:

$$\forall x \forall y [(z \in x \leftrightarrow z \in y) \rightarrow x = y] \ .$$

In other words, a set is uniquely determined by its elements. This is a no-brainer, although the name "extensionality" is not.

II.7.1.2. Axiom of Pairing: If x and y are sets, then $\{x, y\}$ is a set. Formally:

$$\forall x \forall y \exists z [(w \in z) \leftrightarrow (w = x \vee w = y)] \ .$$

The set z (which is unique by extensionality) is denoted $\{x, y\}$.

If x is a set, applying this with $y = x$ we obtain that $\{x\} = \{x, x\}$ is a set. Thus existence of singleton sets follows from the Axiom of Pairing (provided any sets exist!)

II.7.1.3. Axiom of Union: If x is a set, then $\cup x$ is a set. Formally:

$$\forall x \exists u [(z \in u) \leftrightarrow \exists y (y \in x \wedge z \in y)] \ .$$

The set u (which is unique by extensionality) is denoted $\cup x$.

The x should be thought of as the "index set" of a collection of sets (the elements of x). Thus, as is often written, if $x = \{y_i : i \in I\}$, then $\cup x = \cup_{i \in I} y_i$. If x and y are sets, write $x \cup y = \cup\{x, y\}$.

We define a new symbol $\subseteq$ ("subset"):

II.7.1.4. Definition.

$$\forall x \forall y [y \subseteq x \leftrightarrow (z \in y \rightarrow z \in x)] \ .$$

II.7.1.5. We have extended the language of ZF set theory by adding a new symbol. But this is not an essential extension (it is technically a *conservative extension*) since we have only added an abbreviation: $y \subseteq x$ can be replaced wherever it occurs by $\forall z (z \in y \rightarrow z \in x)$. Mathematics is full of such abbreviations; cf. (). We will do this again with the language of classes ().

Even citing a previous result in a proof can be considered a mere abbreviation: it could be avoided by inserting the proof of the previous result in its place. But such abbreviations are crucial, since without them even moderately complicated proofs would quickly become incomprehensible.

II.7.1.6. Axiom of Power Set: The power set of any set exists. Formally:

$$\forall x \exists p [y \in p \leftrightarrow y \subseteq x] \ .$$

The set p (which is unique by extensionality) is denoted $\mathcal{P}(x)$.

The axioms of pairing, union, and power set are clearly obvious types of constructions we would want to do with sets, so make sense to assume (and their names are also obvious). There is another somewhat ill-defined construction we would like to be able to do: make a set of all things with a certain specified property. This construction is called "separation" (a somewhat obscure name, presumably meaning it is possible to "separate" things with the specified property from things which do not).

We first have to say what is meant by a "property." An appropriate way to do this is:

II.7.1.7. Let ϕ be a sentence with $n+1$ free variables $x_0, \ldots, x_n$ (n can be 0), and let $p_1, \ldots, p_n$ be specified sets (called *parameters*). If y is a set, write $\phi(y; p_1, \ldots, p_n)$ if the statement with x_0 replaced by y and $x_1, \ldots, x_n$ replaced by $p_1, \ldots, p_n$ holds.

II.7.1.8. The unrestricted principle of separation or comprehension is that, for any $\phi, p_1, \ldots, p_n$,

$$\{y : \phi(y, p_1, \ldots, p_n)\}$$

is a set. This unrestricted principle leads to a contradiction, however: the set

$$\{y : y \notin y\}$$

cannot exist by Russell's Paradox. We must be a little less ambitious: we can require that all elements of a set with a certain property form a subset (this is an example of what is described in [Mad88] as *one step back from disaster*, a widely-used rule of thumb in Axiomatic Set Theory). Formally:

II.7.1.9. Axiom Schema of Separation: If ϕ is a sentence with $n+1$ free variables $x_0, \ldots, x_n$ (n can be 0), then

$$\forall x \forall p_1 \cdots \forall p_n \exists z [y \in z \leftrightarrow (y \in x \wedge \phi(y, p_1, \ldots, p_n)] \ .$$

This is called an *Axiom Schema* since there is actually a separate axiom for each ϕ (separate axioms for different ϕ are needed since quantification over sentences is not allowed in a first-order theory). Some references call this schema the *Axiom Schema of Subset Selection* or *Axiom Schema of Specification*, which may be more comprehensible and descriptive names.

The set z, which we can abbreviate $\{y \in x : \phi(y, p_1, \ldots, p_n)\}$, is said to be *constructible* from $x, \phi, p_1, \ldots, p_n$. The notion of a constructible, or "definable," set is more versatile than one might at first think (cf. the proof of ()), since sets which exist can be used as parameters to construct other sets.

Here is one important application:

II.7.1.10. PROPOSITION. For every nonempty set x, the intersection $\cap x$ exists. Formally,

$$\forall x \exists u[(z \in u) \leftrightarrow \forall y(y \in x \rightarrow z \in y)] \ .$$

The set u (which is unique by extensionality) is denoted $\cap x$.

If x and y are sets, write $x \cap y = \cap\{x, y\}$. Note that we do not define $\cap\emptyset$: if the definition were applied, every set would be an element of $\cap\emptyset$, which is impossible by Russell's Paradox.

There is one glaring inadequacy in the axioms so far: it cannot be proved from them that there exists a set! Thus we need to add:

II.7.1.11. **Axiom of Existence:** There exists a set.

II.7.1.12. If x is a set, then by separation we have

$$\exists z[y \in z \leftrightarrow (y \in x \wedge y \notin x)] \ .$$

This z (which is unique by extensionality) contains no elements, and is denoted $\emptyset$. Thus if any set exists, then the empty set exists. The converse is obvious, so the Axiom of Existence is equivalent to:

II.7.1.13. **Axiom of Empty Set:** The empty set exists. Formally,

$$\exists z[\forall x(x \notin z)] \ .$$

This z is unique by extensionality and is denoted $\emptyset$.

II.7.1.14. In some treatments of set theory, the symbol $\emptyset$ (or 0) is included in the language as a constant. This takes care of existence, although it must be specified in some way that this symbol does indeed represent the empty set.

Ordered Pairs

Given the Axiom of Pairing, we can define ordered pairs:

II.7.1.15. DEFINITION. If x and y are sets, the *ordered pair* (x, y) is the set

$$(x, y) = \{\{x\}, \{x, y\}\} \ .$$

As with all definitions, this can just be regarded as a notation which is a convenient abbreviation. Note that if x is a set, then $(x, x) = \{\{x\}\}$.

II.7.1.16. DEFINITION. If x and y are sets, the *Cartesian product* $x \times y$ is

$$x \times y = \{(z, w) : z \in x \wedge w \in y\} .$$

This can more carefully but clumsily be written

$$x \times y = \{z : z \in \mathcal{P}(\mathcal{P}(x \cup y)) \wedge \exists v \exists w (v \in x \wedge w \in y \wedge z = \{\{x\}, \{x, y\}\})$$

which shows that $x \times y$ is a set by separation.

II.7.1.17. One can similarly define ordered triples:

$$(x, y, z) := ((x, y), z)$$

and, more generally, ordered n-tuples, and Cartesian products $x \times y \times z$ and $x_1 \times \cdots \times x_n$ for any n. Write x^n for $x \times \cdots \times x$ (n times).

II.7.1.18. There is an alternate approach once one has $\emptyset$: define

$$(x, y) = \{\{x, \emptyset\}, \{y, \{\emptyset\}\}\}$$

(note that $\{\emptyset\} \neq \emptyset$!). If we define $0 = \emptyset$ and $1 = \{\emptyset\}$, as is common in set theory (), we have $(x, y) = \{\{x, 0\}, \{y, 1\}\}$. This is a good definition for ordered pairs. It does not, however, work for ordered n-tuples for $n \geq 3$: if we let 2 be a set different from 0 or 1 (e.g. $2 = \{\emptyset, \{\emptyset\}\}$), and try to define

$$(x, y, z) = \{\{x, 0\}, \{y, 1\}, \{z, 2\}\}$$

then what ordered triple does the set $\{\{0, 1\}, \{0, 2\}, \{1, 2\}\}$ denote? Is it $(1, 2, 0)$ or $(2, 0, 1)$? See also II.7.9.1..

Relations and Functions

II.7.1.19. DEFINITION. A *(binary) relation* from x to y is a subset of $x \times y$. An *n-ary relation* on $x_1, \ldots, x_n$ is a subset of $x_1 \times \cdots \times x_n$. An n-ary relation on x is a subset of x^n.
A *definable relation* from x to y is a constructible (II.7.1.9.) subset of $x \times y$, i.e. a subset

$$\{(z, w) \subseteq x \times y : \phi(z, w, p_1, \ldots, p_n)\}$$

for some $\phi, p_1, \ldots, p_n$.
If r is a relation from x to y, the *domain* and *range* of r are

$$dom(r) = \{z : z \in x \wedge \exists w (w \in y \wedge (z, w) \in r)\}$$

$$ran(r) = \{w : w \in y \wedge \exists z (z \in x \wedge (z, w) \in r)\}$$

(which are sets by separation if r is definable).

The definition of relation will be extended in II.7.1.36.. In considering n-ary relations, it suffices to consider relations on a single set, since an n-ary relation on $x_1 \times \cdots \times x_n$ can be considered an n-ary relation on $x = x_1 \cup \cdots \cup x_n$. This applies in particular to (binary) relations (2-ary relations).

If r is a relation, we usually write xry instead of $(x, y) \in r$.

II.7.1.20. Definition. A *function* from x to y is a relation from x to y in which no two distinct ordered pairs have the same first coordinate.

The definition of function will be extended in II.7.1.36.. If f is a function, we usually write $f(x) = y$ instead of xfy or $(x, y) \in f$.

II.7.1.21. If f is a function from x to y, we do not assume the domain of f is all of x, and we especially do not assume the range of f is y. However, if we use the common arrow notation $f : x \to y$ to denote a function from x to y, without explicit statement to the contrary we conventionally assume the domain is all of x (but that the range is not necessarily all of y).

Infinite Sets

It is consistent with the axioms so far that all sets are finite. An axiom is needed to guarantee the existence of an infinite set. So we want to add:

II.7.1.22. **Axiom of Infinity:** (Informal) There exists an infinite set.

II.7.1.23. Unfortunately, this is too vague since "infinite set" is not well defined. There are several equivalent definitions which can be used, such as () and (). Zermelo and many subsequent authors take a more restrictive approach, which requires existence of the empty set. The best form, usually used today, is:

II.7.1.24. Definition. A set x is *inductive* if (i) $\emptyset \in x$ and (ii) if $y \in x$, then $y \cup \{y\} \in x$.

II.7.1.25. **Axiom of Infinity:** (Inductive form) There exists an inductive set.

II.7.1.26. It can be shown () that if there is an inductive set, then there is a smallest inductive set, and it is infinite by any of the definitions (). Thus the inductive form of the Axiom of Infinity implies the informal form.

II.7.1.27. Some authors such as [Jec03] are a little sloppy in their approach, giving the Axiom of Infinity in inductive form and claiming that it implies the Existence Axiom, so Existence is unnecessary to include. But the inductive form of the Axiom of Infinity cannot even be formulated unless existence of $\emptyset$ is known. One can instead use the informal form of the Axiom of Infinity, specifying that "infinite set" means "T-infinite set" (II.5.6.18.), whose definition does not presuppose the existence of $\emptyset$. This axiom does indeed supersede the Existence Axiom and thus guarantees the existence of $\emptyset$. But this form of the Axiom of Infinity is weaker than the inductive form and does not guarantee the existence of an inductive set (see [Pot04] for an extensive discussion), although it does if the Axiom Schema of Replacement is added (II.8.9.12.).

Classes

II.7.1.28. There is no mention of "classes" in the ZF(C) axioms, so classes which are not sets do not "exist" in this theory. However, we can use the language of classes to give a convenient abbreviation for things which can be properly expressed in the ZF language in a more cumbersome way.

II.7.1.29. Informally, a class is a collection of objects (sets). To be precise, we will usually only consider "definable" classes, and later admissible classes (II.7.1.33.). If ϕ is a sentence with $n+1$ free variables, and $p_1, \ldots, p_n$ are sets, we can consider the class

$$C = \{x : \phi(x, p_1, \ldots, p_n)\}$$

which is *definable (over $p_1, \ldots, p_n$)* by ϕ. For example,

$$V = \{x : x = x\}$$

is a perfectly good class, the "universal class" of all sets. The rough philosophy is that every formula defines a class, but not every formula defines a set, and that sets are classes which are "not too large."

II.7.1.30. Every set can be regarded as a class. If x is a set, then x can be identified with the class

$$X = \{y : y \in x\} .$$

Not every class is (identified with) a set: the universal class V cannot be a set by Russell's Paradox. A class which is not a set is called a *proper class*. We use capital letters to denote classes to distinguish them from sets.

II.7.1.31. We can do "set operations" on classes. If C and D are classes, defined by sentences ϕ and ψ respectively, then we can define $C \cup D$, $C \cap D$, and $C \setminus D$ to be the classes defined by $\phi \vee \psi$, $\phi \wedge \psi$, and $\phi \wedge (\neg\psi)$ respectively. As a special case, we may take $\neg C = V \setminus C$ to be the class defined by $\neg\phi$. We may also define $C \subseteq D$ to mean $\phi \to \psi$. If C and D are definable classes, then $C \cup D$, $C \cap D$, $C \setminus D$, and $\neg C$ are definable classes.

Arbitrary unions and intersections are problematic because classes cannot be elements of other classes, and defining sentences must be finite. But see II.7.1.34. and **??**.

II.7.1.32. If x is a set, identified with the class X, and C is a class defined (over $p_1, \ldots, p_n$) by ϕ, then the class $Y = X \cap C$ is exactly the class corresponding to

$$y = \{z : z \in x \wedge \phi(z, p_1, \ldots, p_n)\}$$

which is a set by Separation. So the intersection of a definable class with a set is a set, and it follows that a definable class which is a subset of a set is a set.

We can extend the notion of definable class in a useful way:

II.7.1.33. Definition. A class C is *admissible* if $C \cap X$ is a set for every set x.

By II.7.1.32., every definable class is an admissible class. Note that this is not a real definition in the language of ZF: it is really only a definition which is applicable in models of ZF, and there are models in which every class is admissible, and other models with classes which are not admissible (cf. II.7.5.7.).

II.7.1.34. We can extend intersection of sets to definable classes. If C is a nonempty class, defined over $p_1, \ldots, p_n$ by ϕ, we can define

$$\cap C = \{z : \forall x(x \in C \to z \in x)\} = \{z : \forall x(\phi(x, p_1, \ldots, p_n) \to z \in x)\} \ .$$

This is a set since it is a subset of any $x \in C$. (We do not define the intersection of the "empty class"; we could take this intersection to be V.)

The Natural Numbers

Using the language of classes, plus the previous axioms, we can show that there is a set with all the properties of the natural numbers. We mostly follow [Jec03].

II.7.1.35. (i) Let S be the class of all inductive sets (II.7.1.24.). Then S is a nonempty definable class. Let $N = \cap S$. Then N is a set and is the smallest inductive set. [We do not follow here our convention of denoting sets by lower-case letters since we want to reserve n for a (variable) element of N.]
(ii) If x is inductive, then $\{y \in x : y \subseteq x\}$ is also inductive. Thus $n \subseteq N$ for every $n \in N$ (i.e. N is transitive (II.8.9.5.)).
(ii) If x is inductive, then $\{y \in x : y$ is transitive$\}$ is also inductive. Thus every $n \in N$ is transitive. So if $n \in N$, then $n = \{m \in N : m \in n\}$ and if $m \in n$, then $m \subseteq n$.
(iv) If x is inductive, then $\{y : y \in x \wedge y \notin y\}$ is also inductive. Thus if $n \in N$, then $n \notin n$.
(v) If x is inductive, then

$$\{y \in x : y \text{ is transitive and every nonempty } z \subseteq y \text{ has an } \in\text{-minimal element}\}$$

is also inductive.
(vi) Every nonempty $x \subseteq N$ has an $\in$-minimal element. [Let $n \in x$ and consider $n \cap x$.]
(vii) If x is inductive, then

$$\{y \in x : y = \emptyset \vee \exists z(z \in x \wedge y = z \cup \{z\})\}$$

is also inductive. So if $n \in N$, then $n = \emptyset$ or $n = m \cup \{m\}$ for some $m \in N$.
(viii) If x is inductive, then

$$\{y \in x : \forall z \in y(z \cup \{z\} \in y \vee z \cup \{z\} = y\}$$

is also inductive. Thus for every $n \in N$ and every $m \in N$, either $m \cup \{m\} \in n$ or $m \cup \{m\} = n$.
(ix) If x is an inductive set with the property in (viii), and $y \in x$, then

$$\{z \in x : z \in y \vee z = y \vee y \in z\}$$

is also inductive. So if $m, n \in N$, then either $m \in n$, $m = n$, or $n \in m$.
(x) If $n \in N$, define $n + 1 = n \cup \{n\} \in N$, and if $m, n \in N$, write $m < n$ if $m \in n$. Then $<$ is transitive and

(a) $n < n + 1$.

(b) $n \neq n + 1$ by (iv).

(c) $n + 1 \neq \emptyset$. If $n \neq \emptyset$, then $n = m + 1$ for some $m \in N$ by (vii).

(d) If $n + 1 = m + 1$, then $n = m$ by (iv) and extensionality.

(e) Either $n = m$, $m < n$, or $n < m$ by (ix). The possibilities are mutually exclusive.

(f) N is Dedekind-infinite [take $f(n) = n + 1$], hence T-infinite.

General Relations and Functions

II.7.1.36. DEFINITION. An *n-ary relation* is a class whose elements are ordered n-tuples (i.e. a subclass of V^n). A *function* is a binary relation in which no two distinct elements have the same first coordinate. A function or relation is *definable* if it is a definable class.

II.7.1.37. To be careful, since classes are not technically part of the language, a (definable) relation R is defined by a sentence ϕ with $n+2$ free variables (n can be 0) and parameters $p_1, \ldots, p_n$, i.e.

$$R = \{(x, y) : \phi(x, y, p_1, \ldots, p_n\}$$

and a function is similarly defined by such a sentence with the property that

$$[\phi(x, y, p_1, \ldots, p_n) \leftrightarrow \phi(z, y, p_1, \ldots, p_n)] \rightarrow x = z \ .$$

The definition of an n-ary relation is analogous.

To distinguish between relations [functions] in the sense of II.7.1.36. and ones in the sense of II.7.1.19., we will call the former *class-relations* [*class-functions*], and say a class-relation [class-function] is a *set-relation* [*set-function*] if it is (identified with) a set. Every relation [function] in the sense of II.7.1.19. is a set-relation [set-function]. The converse is not obvious, but follows from II.7.1.50..

Examples of definable class-relations and class-functions are:

$$\{(x, y) : y \in x\}$$

$$\{(x, y) : y \subseteq x\}$$

$$U = \{(x, y) : y = \cup x\} \text{ (function, i.e. } U(x) = \cup x)$$

$$P = \{(x, y) : y = \mathcal{P}(x)\} \text{ (function, i.e. } P(x) = \mathcal{P}(x))$$

(exercise: write these in defined form).

II.7.1.38. If R is a class-relation defined by ϕ, we define its domain and range by

$$Dom(R) = \{x : \exists y((x, y) \in R)\} = \{x : \exists y(\phi(x, y, p_1, \ldots, p_n))\}$$

$$Ran(R) = \{y : \exists x((x, y) \in R)\} = \{y : \exists x(\phi(x, y, p_1, \ldots, p_n))\}$$

which are definable classes. If R corresponds to a set-relation r, then $Dom(R)$ and $Ran(R)$ correspond to the sets $dom(r)$ and $ran(r)$ as defined in II.7.1.19..

If F is a function and C a class (not necessarily a subclass of $Dom(F)$), define the image

$$F(C) = \{y : \exists x[x \in C \wedge (x, y) \in F]\}$$

which is a class: specifically, if C is defined by $\phi(x, p_1, \ldots, p_n)$ and F is defined by $\psi(x, y, q_1, \ldots, q_m)$, then $F(C)$ is defined by

$$\{y : \exists x[\phi(x, p_1, \ldots, p_n)] \wedge [\psi(x, y, q_1, \ldots, q_m)]\}$$

with parameters $p_1, \ldots, p_n, q_1, \ldots, q_m$. This is a definable class. If X corresponds to a set x, then we can write $F(X)$ for the image of x under F. Note that $F(X)$ is not the same as $F(x)$ which is the "value" of F at x, the second coordinate of the ordered pair in F, if any, with first coordinate x. Thus $F(x)$ is undefined if $x \notin Dom(F)$, but $F(X)$ is always defined although it can be $\emptyset$. In fact, $F(X) = \{F(z) : z \in x\}$. To avoid confusion, $F(X)$ is often written $F \restriction x$. We can define preimages of classes similarly, and images and preimages for more general (binary) relations.

Replacement

II.7.1.39. FRAENKEL added to ZERMELO'S axioms of set theory the following strengthening of the Axiom Schema of Separation, also formulated independently by SKOLEM (this was FRAENKEL'S main contribution to the axioms of "Zermelo-Fraenkel" set theory):

II.7.1.40. **Axiom Schema of Replacement:** If F is a definable function and X is (identified with) a set, then $F(X)$ is also (identified with) a set. Formally: if ϕ is a sentence with $n+2$ free variables (n can be 0), then

$$\forall p_1 \cdots \forall p_n[([\phi(x, y, p_1, \ldots, p_n) \leftrightarrow \phi(z, y, p_1, \ldots, p_n)] \rightarrow x = z) \rightarrow \forall x \exists w(y \in w \leftrightarrow \exists z[z \in x \wedge \phi(z, y, p_1, \ldots, p_n)])] \; .$$

This schema is called "replacement" (or sometimes "substitution") since it says that the elements of a set can be "replaced" by their images under any (definable) function to obtain a new set. The rough idea of Replacement, along with II.7.1.32., is that a class is a set if and only if it is "small enough."

II.7.1.41. An immediate consequence is: If F is a definable function and $Dom(F)$ is a set, then $Ran(F)$ is also a set. This statement is actually equivalent to the Axiom Scheme of Replacement.

II.7.1.42. PROPOSITION. The Axiom Schema of Replacement implies the Axiom Schema of Separation.

PROOF: If ϕ is a sentence with $n+1$ free variables, x is a set, and $p_1, \ldots, p_n$ are parameters, we must show that $y = \{z \in x : \phi(z, p_1, \ldots, p_n)\}$ is a set. Set

$$\psi(z, w, p_1, \ldots, p_n) = z = w \wedge \phi(z, p_1, \ldots, p_n) \; .$$

then ψ defines a function F, and $F(X)$ is exactly the class identified with y. 👍

II.7.1.43. The converse is not true: the Axiom Schema of Separation does not imply the Axiom Schema of Replacement, even in the presence of all the other axioms we have considered (plus Regularity and Choice). In fact, one cannot add finitely many axioms to these to deduce the Axiom Schema of Replacement [].

II.7.1.44. The Axiom Schema of Replacement, along with the Existence and Power Set axioms, also implies the Axiom of Pairing: $\emptyset$ exists, so $\{\emptyset\} = \mathcal{P}(\emptyset)$ exists, and thus $\mathcal{P}(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\}$ exists (i.e. is a set). If x and y are sets, define a function F to be the collection $\{(\emptyset, x), (\{\emptyset\}, y)\}$ (to avoid circularity, this can be expressed by a formula not referring to ordered pairs). $Dom(F)$ is the set $\{\emptyset, \{\emptyset\}\}$, so $Ran(F) = \{x, y\}$ is also a set. Thus Replacement is a very powerful assumption (some set theorists consider it "too powerful" and omit it).

II.7.1.45. The Axiom Schema of Replacement requires that the relation be a function; it is false for more general relations since there may be "too many" y's related to one x. The most extreme example is to take $\phi(x, y) = x \subseteq y$, i.e. $R = \{(x, y) : x \subseteq y\}$. Then $\{\emptyset\}$ is a set x, but $R(X) = V$, which is not a set.

Indexed Unions and Intersections

II.7.1.46. As a sample application of the Axiom Schema of Replacement, if s is a set and for each $j \in s$ we have a set x_j, then the collection $\{(j, x_j) : j \in s\}$ is a function F whose domain is the set s, hence by replacement

$$Ran(F) = \{x_j : j \in s\}$$

is a set which can be regarded as a set of sets indexed by s. We also have that the union of this set is a set, usually denoted $\cup_{j \in s} x_j$.

In particular, if we have a set x_n for each $n \in N$, then $\{x_n : n \in N\}$ and $\cup_{n \in N} x_n$ are sets. Functions with domain N can be defined recursively: for example, if x is any set let $x_\emptyset = x$ and $x_{n+1} = \mathcal{P}(x_n)$ recursively. One obtains sets

$$\{x, \mathcal{P}(x), \mathcal{P}(\mathcal{P}(x)), \mathcal{P}(\mathcal{P}(\mathcal{P}(x))), \dots\}$$

and $\cup_{n \in N} \mathcal{P}^n(x)$. Or we can take $x_\emptyset = x$ and $x_{n+1} = \cup x_n$ for each n. Then $\cup_{n \in N} x_n$ is a transitive set, called the *transitive closure* of x and denoted $tc(x)$, which is the smallest transitive set containing x (it can also be obtained as the intersection of the class of all transitive sets containing x).

II.7.1.47. We may also define indexed intersections of sets. If s is a set and for each $j \in s$ we have a set x_j, let $x = \cup_j x_j$, and for each j let $y_j = x \setminus x_j$. Then $y = \cup y_j$ is a set which is a subset of x, and we let $\cap_j x_j = x \setminus y$.

II.7.1.48. There is a subtlety in these constructions. To apply Replacement, the function $F = \{(j, x_j) : j \in s\}$ must be definable, i.e. the sets x_j must have some definable relation with each other, e.g. are definable by recursion (more generally, it suffices that F is a set, cf. II.7.1.50.). Even the union or intersection of a sequence of unrelated sets is problematic; cf. II.7.5.7.. This point is glossed over in most applications, but most indexed collections of sets have sufficient coherence that the difficulty can be overcome with care.

Extending Replacement

II.7.1.49. The Axiom Schema of Replacement only directly applies to definable functions. Not all functions in the sense of II.7.1.20., i.e. functions which are sets of ordered pairs, are necessarily definable in the sense of II.7.1.36.. However, we have the following version of Replacement for general set-functions:

II.7.1.50. Proposition. Let r be a set-relation (e.g. a set-function), not necessarily definable in the sense of II.7.1.36.. Then

$$Dom(r) = \{x : \exists y((x, y) \in r)\}$$

$$Ran(r) = \{y : \exists x((x, y) \in r)\}$$

are sets.

Proof: Define functions (in the class sense) DOM and RAN by

$$DOM = \{((x, y), x)\}$$

$$RAN = \{((x, y), y)\}$$

i.e. DOM and RAN assign to each ordered pair its first and second coordinate respectively. Then DOM and RAN are definable functions in the sense of II.7.1.36., and $Dom(r)$ and $Ran(r)$ are the images of the set r under DOM and RAN respectively, hence are sets by Replacement. 👍

II.7.1.51. Corollary. Let f be a set-function. For every set x, $f \restriction x$ is a set.

Proof: The set $g = \{(y,z) \in f : y \in x\}$ is a subset of f by Separation, hence a set-function, and $Ran(g) = f \restriction x$. 👍

II.7.1.52. Note that II.7.1.50. is a generalization of the condition in II.7.1.19. (in fact, it shows that every set-relation is of the form in II.7.1.19.), but is is not really an extension of Replacement. It is not true (provable) that if F is a function and $Dom(F)$ a set, then $Ran(F)$ is necessarily a set (II.7.5.7.). In fact, the statement that the image of a set under an arbitrary function is a set is equivalent to Morse-Kelley (MK) set theory ().

We may make the following definition:

II.7.1.53. Definition. A class-function F is an *admissible function* if it is an admissible class and $F \restriction x$ is a set for every set x.

By II.7.1.51., every set-function is an admissible function.

II.7.1.54. There is potential language ambiguity here: an admissible class-function is an admissible class, i.e. its intersection with any set is a set-function, but the converse is not obvious. A class-function F is admissible if and only if $F \cap (x \times V)$ is a set for every set x (being an admissible class is equivalent to the condition that $F \cap (x \times y)$ is a set for all sets x and y).

II.7.2. The Axiom of Regularity

II.7.2.1. Von Neumann's Axiom of Regularity had its origin in Russell's Paradox, where one of the main problems was the existence of sets which are not ordinary (i.e. a set x with $x \in x$.) A similar problem arises if there is a finite loop $x_1, \ldots, x_n$ of sets with $x_{k+1} \in x_k$ for $1 \leq k < n$ and $x_1 \in x_n$, so this possibility should also be eliminated. Zermelo's axioms allowed such situations, although they (apparently) do not create a paradox in his system. But it is undesirable to allow such sets, and von Neumann's solution was to add the following axiom, which is also sometimes called the "Axiom of Foundation":

II.7.2.2. **Axiom of Regularity:** For every nonempty set x there is a $y \in x$ with $y \cap x = \emptyset$.

Such a y is called a *minimal element* of x (the terminology is explained in ().) The Axiom of Regularity prevents the existence of loops of the type described above; in fact, it is equivalent to the following statement:

II.7.2.3. Descending Chain Condition: There is no infinite sequence (x_n) of sets (not necessarily distinct) with $x_{n+1} \in x_n$ for all n.

To see that Regularity implies the Descending Chain Condition, suppose (x_n) is a sequence with $x_{n+1} \in x_n$ for all n, and let $z = \{x_n : n \in N\}$. Then z is a set by II.7.1.46.. Thus by Regularity there is a $y \in z$ with $y \cap z = \emptyset$. Then $y = x_n$ for some n. But then $x_{n+1} \in y \cap z$, a contradiction. [The Axiom of Choice is needed to prove the other direction of the equivalence.]

II.7.2.4. The Axiom of Regularity seems reasonable if one thinks of sets as being constructed in "stages", for then the elements of a set constructed at one stage must have been constructed at a previous stage; an element constructed at the earliest possible stage would be a minimal element. This can be made precise, and it turns out that the Axiom of Regularity limits the universe just to the sets which can be constructed out of $\emptyset$ by repeated union, power set, and subset (see II.8.8.4.); thus it seems to be in conflict with the Basic Principle above.

II.7.2.5. PROPOSITION. If C is any nonempty definable class, then there is a $y \in C$ with $Y \cap C = \emptyset$. Such a y is an $\in$-minimal element of C, i.e. $z \notin C$ for all $z \in Y$. In particular, if x is any set, then there is an $\in$-minimal element y of x, namely any $y \in x$ with $y \cap x = \emptyset$.

PROOF: Let $x \in C$. If $X \cap C = \emptyset$, we are done. Otherwise, let $Z = T \cap C$, where $t = tc(x)$. Then Z corresponds to a set z, which is nonempty since $X \subseteq T$. Let $y \in z$ with $y \cap z = \emptyset$. Then $Y \cap C = \emptyset$; otherwise, if $w \in Y \cap C$, then $w \in t$ since t is transitive, so $w \in z$, a contradiction since $w \in y$. 👍

Well-Founded Relations

II.7.3. Controversy

II.7.3.1. The fact that the Axiom of Regularity is a *restrictive* axiom is the source of the discomfort some mathematicians have with it. On the other hand, set theory with the Axiom of Regularity is large enough to encode all of standard mathematics in the usual way, and there are no important results which seem to be obtainable with a weakened version of the Axiom but not with the full version, so there is (at present) no practical reason to weaken it.

II.7.3.2. The situation with the Axiom of Choice is the opposite: it is an *expansive* axiom, allowing a certain infinite operation. Some mathematicians are uncomfortable with it because it seems to allow too much freedom in manipulating sets (although it could be argued that it is more in line with the Basic Principle than the Axiom of Regularity is.) There continue to be parts of mathematics where the Axiom of Choice is not assumed, and sometimes a contradictory axiom is substituted. One of the root causes of the controversy over the Axiom of Choice was identified by B. ROTMAN and G. T. KNEEBONE:

> "[T]he axiom of choice has a curious ambivalence, since it can be regarded with equal justification as a *mathematical* axiom, which guarantees the existence of a set in certain stated circumstances, and as a *logical* principle, needed for the authentication of a pattern of inference that is indispensible in many branches of mathematics."[15]

[15] B. Rotman and G. T. Kneebone, *The Theory of Sets and Transfinite Numbers*, American Elsevier, 1968, p. 60.

In his very interesting book, *The Banach-Tarski Paradox*, STAN WAGON describes the role of the Axiom of Choice in connection with the Paradox, and in mathematics in general:

> "Despite the Banach-Tarski Paradox and other objections to the Axiom of Choice, often focusing on its nonconstructive nature, the great majority of contemporary mathematicians fully accept the use of AC. It is generally understood that nonmeasurable sets (either of the Vitali type or of the sort that arise in the Banach-Tarski Paradox) lead to curious situations that contradict physical reality, but the mathematics that is brought to bear on physical problems is almost always mathematics that takes place entirely in the domain of measurable sets. And even in this restricted domain of sets, the Axiom of Choice is useful, and ZF + AC provides a more coherent foundation than does ZF alone. ...
>
> Despite the general acceptance of AC, it is recognized that this axiom has a different character than the others. Because of the nonconstructivity it introduces, AC is avoided when possible, and proofs in ZF are considered more basic (and are often more informative) than proofs using AC. A purist might argue that once accepted as an axiom, AC should be accorded the same status as the other axioms: a statement that is proved in ZF + AC is just as mathematically true as one proved in ZF. But there are sometimes beneficial side effects of a discriminating attitude towards the use of AC. For instance, consider the Schröder-Bernstein Theorem for cardinality [()], which is much easier to prove in ZFC than in ZF. ... But the ZF-proof ... is much more valuable because it generalizes in a way that the ZFC-proof does not. ...
>
> The interplay between ZF and ZF + AC works the other way as well. For example, the ZF-proof that a square is not paradoxical ... was found by closely analyzing the original ZFC-proof. If a more restrictive attitude toward AC had prevailed, the original ZFC-proof might not have been discovered so soon, and the proof (in ZF) that squares are not paradoxical might have been delayed."

II.7.4. The ZF Axioms (Recapitulation)

II.7.4.1. A minimal set of axioms giving ZF set theory is thus:

Axiom of Existence.

Axiom of Extensionality.

Axiom of Union.

Axiom of Power Set.

Axiom of Infinity.

Axiom Schema of Replacement.

Axiom of Regularity.

II.7.4.2. A minimal set of axioms for ZFC set theory is thus the above axioms plus

Axiom of Choice.

II.7.5. Models of ZF(C) Set Theory

ZF(C) is intended to axiomatize naive set theory. But, as with any (consistent) first-order theory, it has many models, which turn out to have varying properties (for example, in some models of ZF the AC is true, i.e. it is a model of ZFC, and in others the AC is false). We give only a brief introduction to the model theory of ZF(C), which is a major topic in set theory.

II.7.5.1. A model M of ZF set theory consists of a collection V_M of objects, and a relation $\in_M$ on the collection of objects, such that the ZF axioms hold. The objects of M are called "sets" (and thought of this way), although they need not be actual sets in the naive or metalanguage sense; the name "set" is simply a convenient (and suggestive) term with no intrinsic meaning, similar to calling the objects in a general metric or topological space "points." In the same way, the relation $\in_M$ need not have any connection with actual containment in the naive sense. For clarity, we will call the objects of M *M-sets*, and $\in_M$ *M-containment*: if $y \in_M x$, we call y an *M-element* of x.

II.7.5.2. The ZF Axiom of Union entails that for every M-set x, there is an M-set $\cup_M x$ whose M-elements are exactly the M-elements of M-elements of x. We can also say that an M-set y is an *M-subset* of an M-set x, written $y \subseteq_M x$, if every M-element of y is an M-element of x. The Power Set Axiom then asserts that for every M-set x there is an M-set $\mathcal{P}_M(x)$ whose M-elements are exactly the M-subsets of x. We can also define M-relations as M-sets of ordered pairs (defined in the obvious way in M, i.e. the ordered pair $(x, y)_M$ is the unique M-set whose M-elements are exactly z and w, where z is the unique M-set whose only M-element is x and w the unique M-set whose M-elements are exactly x and y, whose existence is given by the Pairing Axiom and uniqueness by Extensionality), and M-functions as M-relations with the function property.

II.7.5.3. The ZF axioms require that certain M-sets exist, beginning with the M-empty set $\emptyset_M$ (which need not have anything to do with the usual empty set $\emptyset$!), including M-unions, M-power sets, M-sets definable by Separation or Replacement, and a model N_M of N. But M may potentially have many more M-sets than just those entailed by the ZF axioms. Similarly, there may be many more M-relations and M-functions than the definable ones.

II.7.5.4. Actually, it turns out (Mostowski Collapsing Theorem) that every model of ZF is equivalent to a model M in which there is a transitive set (in the naive sense) V_M whose elements are the M-sets, and M-containment is actual containment, i.e. if x and y are M-sets, then $y \in_M x$ if and only if the set y is an actual element of the set x. Such a model is called a *transitive model*, or *standard model*, of ZF.

In a standard model, if x is an M-set, then $\cup_M x$ is necessarily just $\cup x$. However, a subset (in the naive sense) of an M-set x is not necessarily an M-set, and hence $\mathcal{P}_M(x)$ is not necessarily all of $\mathcal{P}(x)$. Not all relations or functions on the M-sets are necessarily M-relations or M-functions. In a transitive model M with universe V_M, the admissible classes are the subsets of V_M whose intersection with every M-set is an M-set.

II.7.5.5. It is in fact not entirely clear that ZF has any models, i.e. that it is consistent. Gödel's Second Theorem implies that it cannot be proved (within ZF) that ZF is consistent. But almost all mathematicians take it as an article of faith that ZF is consistent, especially since no hint of inconsistency has appeared so far. Under this assumption, ZF has many models, even many standard models.

The Skolem Paradox

II.7.5.6. The Löwenheim-Skolem Theorem states that any consistent first-order theory with a countable language has models of every infinite cardinality, and in particular has a countable model. In particular, if ZF(C) is consistent, it has a countable transitive model (CTM) M. V_M contains the M-infinite set N_M, which is infinite and necessarily countable since M is countable. Then by Cantor's Theorem (a theorem of ZF), $\mathcal{P}_M(N_M)$ is uncountable. But $\mathcal{P}_M(N_M) \subseteq V_M$, and V_M is countable. This apparent contradiction is *Skolem's Paradox*.

The resolution is actually rather straightforward: $\mathcal{P}_M(N_M)$ is "externally countable," i.e. there is a function from N onto $\mathcal{P}_M(N_M)$, but not "internally countable," i.e. there is no M-function from N_M onto $\mathcal{P}_M(N_M)$.

II.7.5.7. Existence of a CTM M of ZF(C) also illustrates the limits of Replacement. Let N_M be the model of N in M. Then N_M is a countably infinite M-set. N_M has uncountably many subsets, most of which are not M-sets since there are only countably many M-sets. All elements of N_M are M-sets by construction, so all finite subsets of N_M, and in particular all singleton subsets, are M-sets by Pairing. Thus the union of a sequence of M-sets is not necessarily an M-set. Also, there are uncountably many injective functions from N_M to N_M, only countably many of which are M-functions. The domain of each of these uncountably many functions is N_M, an M-set, but the range of most is not an M-set. Thus for the range of a function to be an M-set, it is not sufficient for the domain to be an M-set.

The Axiom of Choice

II.7.5.8. Existence of countable transitive models shows that some care must be exercised in stating the indexed form of the Axiom of Choice II.6.1.2. (there is no problem with the power set form II.6.1.3.). Even the Countable AC is inconsistent if not stated carefully: if M is a CTM of ZF, as in II.7.5.7. there is a sequence of M-singleton subsets of N_M whose union is not an M-set; the Countable AC, if expressed too broadly, would imply that it is (application of the Countable AC to a sequence of singleton sets is nonstandard and not what it was designed for, but is legitimate). Thus if the Countable AC is not phrased carefully, there cannot be a countable model of ZF+(Countable AC) and therefore ZF+(Countable AC) is inconsistent. So we should phrase the indexed form of the AC as follows (the Countable AC is the same with $s = N$):

Axiom of Choice, Indexed Version: Let f be a set-function with domain s, such that $f(i) \neq \emptyset$ for every $i \in s$. Then there is a set-function c with domain s, such that $c(i) \in f(i)$ for all $i \in s$.

We leave to the reader a rephrasing of this axiom in the language of ZF.

II.7.6. The NBG Axioms

II.7.7. The MK Axioms

II.7.8. Set Theory with Atoms

II.7.8.1. It is natural from naive set theory to allow things which are not sets to be elements of sets. For example, it is natural not to regard numbers as sets (), but as fundamental objects, even though numbers can be modeled in a mathematically but not esthetically satisfactory way within ZF set theory. We can allow such members by expanding the language of ZF to include additional objects which are called *individuals* or

atoms; the German name *urelement* is also frequently used. The only real expansion of the language needed is an additional unary predicate Set: for each object x, $Set(x)$ holds if and only if x is a set. We need to modify the ZF axioms and add at least one:

II.7.8.2. Axiom of Membership: $\forall x([\exists y(y \in x)] \rightarrow Set(x))$.

Informally, individuals have no members.

Extensionality must be modified: if there is an individual x, then x has the same members as $\emptyset$, but $x \neq \emptyset$ since $Set(\emptyset)$ but $\neg Set(x)$. So Extensionality becomes:

II.7.8.3. Axiom of Extensionality with Atoms:

$$\forall x \forall y([Set(x) \wedge Set(y) \wedge \forall z(z \in x \leftrightarrow z \in y)] \rightarrow x = y) \ .$$

II.7.8.4. The collection $A = \{x : \neg Set(x)\}$ of individuals is a definable class. It is unclear whether we should take A to be a set. It is true that for any set x, $\{y \in x : \neg Set(y)\}$, the collection of individuals in x, is a set by Separation, as is $\{y \in x : Set(y)\}$.

II.7.9. Exercises

II.7.9.1. Is it a reasonable definition to define the ordered triple (a, b, c) to be $\{\{a\}, \{a, b\}, \{a, b, c\}\}$? [What if a, b, c are not all distinct?]

II.7.10. The Theory of Set Theories

II.7.10.1. In what sense is "set theory" axiomatized? It is illuminating to compare Axiomatic Set Theory with, say, Group Theory. There are close formal analogies, but large philosophical or conceptual differences.

Group Theory, or the Theory of Groups, is a mathematical theory in the sense of (), including a list of axioms (). The theory was developed and axiomatized to formalize structure and properties shared by many specific mathematical objects, and to provide precision and facilitate proofs of statements about these objects (and any others satisfying the axioms). Similarly, Set Theory was axiomatized to provide precision and facilitate proofs of statements about sets (as well as to avoid the paradoxes discovered in naive set theory).

II.7.10.2. But there is a big conceptual difference between the theories. There are many groups (i.e. many models of Group Theory), even many which were already known, with varying properties, and there is no one structure the theory was meant to uniquely describe. But it is, or was, generally believed there is one thing called "set theory" which was not precisely described previously but which was well determined, and the axiomatization was intended to give a formal description of exactly that thing. The notion that there could be more than one "set theory" was rarely if ever contemplated.

As it turns out, there are many models of, say, ZF or ZFC set theory, also with varying properties (there is the additional complication that, unlike Group Theory, there is not a universally accepted axiomatization of Set Theory). Each of these models should properly be called "a set theory," and one should correctly talk about "the Theory of ZF Set Theories" etc.

II.7.10.3. There is an unfortunate language ambiguity here – the word "theory" is used in two different senses. One use is that *set theory* is used as a single term, analogous to the term *group*. The other use is as a mathematical theory in the sense of (). To attempt to ease the confusion, we may write *set-theoretic universe* for the first and *Axiomatic Set Theory* for the second. Thus we have at least three different things in the discussion:

> **Axiomatic Set Theory:** A mathematical theory codifying sets and their properties. There are several Axiomatic Set Theories: ZF, ZFC, ZFU, NBG, NBGC, . . . , and probably more not yet defined.
>
> **set-theoretic universe:** A model of some Axiomatic Set Theory.
>
> **Set Theory:** The actual "theory of sets" (if such a theory exists at all), only incompletely described at present by naive set theory and the various Axiomatic Set Theories.

II.7.10.4. If one believes in the existence of Set Theory, then one must necessarily believe some statements about Set Theory, including:

1. Either atoms (individuals) are allowed as elements of sets, or only other sets are allowed to be elements of sets.
2. The Axiom of Foundation is either true or false.
3. Proper classes either exist, or they do not.
4. The AC is either true or false. The same can be said about each of the weaker forms of Choice.

5. The CH is either true or false. More precisely, $2^{\aleph_0}$ is either some specific aleph, or it is not an aleph at all.

One need not believe one or the other of these in each case, or even believe the answer is knowable; but *unknowable* does not mean *indeterminate*, and one cannot believe the answer is indeterminate.

II.7.10.5. A mathematician need not believe in the existence of Set Theory, at least as part of mathematics. A Platonist or realist would be drawn to believe it, and a formalist perhaps not. Set theorists seem to tend to believe it, at least in some sense. For example, there has recently been great renewed interest in whether the CH is "true," and some set theorists are now apparently becoming convinced it is "false." This question cannot even be discussed unless one believes that ZF(C) is an imperfect axiomatization of Set Theory, since it is meaningless in ZFC. (Belief that there is a Set Theory and that ZFC is an imperfect axiomatization of it has been explicitly stated by some leading set theorists, e.g. in [**?**], [Bel77, Foreword], [KM78], [Mar76].)

II.7.10.6. Even if one believes the existence of Set Theory, by Gödel's Incompleteness Theorem it must be that either Set Theory is not finitely describable or there are statements in Set Theory which are true but unprovable. In either case, or if Set Theory does not exist, there are "truths" in set theory we cannot know.

II.7.11. Are the Foundations of Mathematics Circular?

II.7.11.1. If one carefully examines the (current) foundations of mathematics, there is an uneasy feeling of circularity. When the notion of a mathematical theory and models is formulated, there is an underlying assumption of some naive version of set theory, including some form of the Axiom of Choice (used e.g. in the Completeness and Compactness Theorems of model theory), and even before that an assumption of basic modes of reasoning and drawing conclusions. Then using the notion of a mathematical theory, Axiomatic Set Theory is formulated, and all the rest of mathematics is then developed within set theory. It is particularly problematic to talk about models of set theory in which the AC does not hold, since the AC is implicitly assumed (informally) in formulating the structure in which set theory is axiomatized in the first place.

If we try to carefully formalize the subfoundation allowing us to formulate the notion of a mathematical theory, we run into the same problem again, and then again, ... As the saying goes:

> "It's turtles all the way down."

(https://en.wikipedia.org/wiki/Turtles_all_the_way_down) This refers to the picture that the earth is a large disk resting of the back of an enormous turtle. What does the turtle stand on? Maybe the back of another turtle. But what does it stand on? Continuing, what does the bottom turtle stand on? Maybe there is no bottom turtle: it's turtles "all the way down." (But then what does "all the way" mean?) One wag described the mathematical version as "It's Gödels all the way down."

II.7.11.2. Of course, one has to start *somewhere*! Set theorists admit there is circularity of a sort, since we have to begin with something. One way the current setup is described (or rationalized) is that the underlying notions of sets and modes of reasoning lie at the level of the metalanguage, while theories, including Axiomatic Set Theory, lie at the level of an object language. While this explanation may not be entirely satisfactory, it's probably the best we can do now.

II.8. Ordinals

Ordinals can be thought of as standard models for well ordered sets. The ordinals we describe are technically *von Neumann ordinals*; there are other equivalent approaches (cf. e.g. II.8.9.6.).

II.8.1. Basic Definitions

The formal theory of ordinals takes place in ZF set theory, where every element of a set is itself a set and there is no reason to make notational distinctions between sets and their elements (so we will frequently use lower-case letters to denote both sets and elements.) Thus the following strange-looking definition makes sense:

II.8.1.1. Definition. A set x is *transitive* if every element of x is a subset of x. A set x is an *ordinal* if x is transitive and every element of x is transitive. Denote the class of all ordinals by Ω.

The term "transitive" is appropriate, since x is transitive if $y \in x$ and $z \in y$ imply $z \in x$.

We will normally use the Greek letters σ, τ, ρ to denote general ordinals, and other lower-case Greek letters to denote specific ordinals.

There is one ordinal to begin with, $\emptyset$ (so $\Omega \neq \emptyset$.) There are many other ordinals which can be manufactured from $\emptyset$ by "transfinite induction," using the next two constructions.

II.8.1.2. Definition. Let σ be an ordinal. Then $S(\sigma) = \sigma \cup \{\sigma\}$ is also an ordinal, called the *successor* of σ.

Note that $\sigma \notin \sigma$ by the Regularity Axiom (), so $S(\sigma) \neq \sigma$. The term "successor" makes sense in reference to a natural ordering on ordinals, described below.

Similarly, the next result is immediate from the definition:

II.8.1.3. Proposition. Let $\{\sigma_i : i \in I\}$ be a set of ordinals. Then $\cup_{i\in I}\sigma_i$ and $\cap_{i\in I}\sigma_i$ are ordinals.

Together, these operations give a very large supply of ordinals. In fact:

II.8.1.4. Proposition. If $X = \{\sigma_i : i \in I\}$ is any set of ordinals, there is an ordinal σ which is not in X.

Proof: Set $\sigma = S(\cup_{i\in I}\sigma_i)$. 👍

II.8.1.5. Corollary. The class Ω of all ordinals is not a set.

This is the *Burali-Forti Paradox*, which like Russell's Paradox implies that there is no "set of all sets."

We will describe specific examples arising from these constructions in the next section.

II.8.2. Ordering on Ordinals

We now discuss the natural ordering on ordinals. First comes a simple observation:

II.8.2.1. PROPOSITION. If σ is an ordinal, then every element of σ is also an ordinal.

Indeed, if $\tau \in \sigma$, then τ is transitive; and since $\tau \subseteq \sigma$, every element of τ is an element of σ, hence transitive.

II.8.2.2. DEFINITION. If σ and τ are ordinals, then $\tau < \sigma$ if $\tau \in \sigma$.

Thus, if σ is an ordinal, then

$$\sigma = \{\tau \in \Omega : \tau < \sigma\}.$$

So $<$ defines an "ordering" on the set σ.

II.8.2.3. PROPOSITION. The relation $<$ is a strict "partial order" on Ω.

[The "relation" $<$ is technically not a relation in the usual sense, since Ω is not a set. However, the meaning of the term "partial order" is still clear. In particular, $<$ is a true partial order on any *set* of ordinals, and in particular on any ordinal itself.]
PROOF: If $\tau < \sigma$ and $\sigma < \rho$, then $\tau \in \sigma$ and $\sigma \subseteq \rho$, so $\tau \in \rho$, i.e. $\tau < \rho$, so $<$ is transitive. Antisymmetry and strictness of the partial order follow from the Regularity Axiom, which implies that $\sigma \notin \sigma$ for any σ. 👍

Some version of the Regularity Axiom is necessary here; otherwise there could be strange sets x with the property that $x = \{x\}$ and any such set would be an ordinal.

The Regularity Axiom also implies:

II.8.2.4. PROPOSITION. Every nonempty set of ordinals contains a minimal element (with respect to $<$.)

In fact, if X is a nonempty set of ordinals, then any minimal element of X in the sense of () is also minimal with respect to $<$, as is easily verified (this is the origin and explanation of the term *minimal* in the Regularity Axiom.)

The existence of minimal elements also applies to classes of ordinals which are not sets: if V is a nonempty class of ordinals, let $\sigma \in V$, and let $X = \{\tau \in V : \tau \leq \sigma\}$. Then X is a nonempty *set* of ordinals, and any minimal element of X is a minimal element of V.

This fact and the next one are probably the most important facts about the ordering on Ω.

II.8.2.5. PROPOSITION. The ordering on Ω is a total ordering, i.e. any two ordinals are comparable.

PROOF: Suppose there are ordinals which are not comparable. Let σ be a minimal ordinal for which there is an ordinal not comparable to σ, and let τ be a minimal ordinal which is not comparable to σ. If $\rho \in \sigma$, then $\rho < \sigma$ so ρ is comparable to any ordinal, and hence to τ. We cannot have $\tau \leq \rho$ since then $\tau < \sigma$, so $\rho < \tau$, i.e. $\rho \in \tau$, so $\sigma \subseteq \tau$. Similarly, if $\rho < \tau$, then ρ is comparable to σ and $\sigma \not\leq \rho$, so $\rho < \sigma$ and $\tau \subseteq \sigma$, i.e. $\tau = \sigma$, a contradiction. So there are no such σ and τ. 👍

Putting together the last two results, we obtain:

II.8.2.6. THEOREM. Every nonempty class of ordinals has a smallest element, i.e. Ω is well ordered (more correctly, every set of ordinals is well ordered in the induced ordering.)

In particular, since every ordinal is itself a set of ordinals, each ordinal is well ordered in its natural ordering. If σ is an ordinal, then $S(\sigma)$ is indeed the smallest ordinal larger than σ, justifying the term "successor".

II.8.2.7. COROLLARY. Let σ and τ be ordinals. If $\tau \subseteq \sigma$, then $\tau \leq \sigma$ (i.e. $\tau = \sigma$ or $\tau \in \sigma$.)

Thus, on Ω, the relation $\leq$ is the same as $\subseteq$.

It follows easily that if $x = \{\sigma_i : i \in I\}$ is a nonempty set of ordinals, then $\cap_{i\in I}\sigma_i$ is the smallest element of x. It is also easily seen that $\cup_{i\in I}\sigma_i$ is the smallest ordinal σ such that $\sigma_i \leq \sigma$ for all $i \in I$; this element is often called $\lim(x)$ [x is well ordered in the induced ordering and can be regarded as a net, and this net converges to $\lim(x)$ in the order topology ()]. Note that $\lim(x)$ is also the supremum of x in Ω, although the notation $\sup(x)$ is often used in ordinal theory to denote $S(\lim(x))$, in conflict with the usual notation for ordered sets.

II.8.2.8. DEFINITION. An ordinal σ is a *successor ordinal* if there is a τ with $\sigma = S(\tau)$. Otherwise σ is a *limit ordinal*.

An ordinal σ is a successor ordinal if and only if it contains a maximal element τ, in which case $\sigma = S(\tau)$; it follows that the τ for which $\sigma = S(\tau)$ is unique.

By this definition, $\emptyset$ is a limit ordinal; but this ordinal is special (it is the smallest or first ordinal) and it is often treated separately from all others.

II.8.3. Transfinite Induction

The Principle of Transfinite Induction applies to the class Ω:

II.8.3.1. THEOREM. [PRINCIPLE OF TRANSFINITE INDUCTION] Suppose $P(\sigma)$ is a statement for each $\sigma \in \Omega$. Suppose, for each σ, the following *inductive implication* holds:

"If $P(\tau)$ is true for all $\tau < \sigma$, then $P(\sigma)$ is true."

Then $P(\sigma)$ is true for all $\sigma \in \Omega$.

The argument is identical to the one in II.4.2.6.; the fact that Ω is not a set causes no difficulty.

Definitions by Transfinite Induction (Transfinite Recursion) also work for Ω:

II.8.3.2. THEOREM. Let X be a set. Suppose, for each $\sigma \in \Omega$, we have a definite procedure or rule R_σ for choosing an $f(\sigma) = x_\sigma \in X$, which may depend on the choice of the $f(\tau)$ for $\tau < \sigma$. Then there is a unique $f : \Omega \to X$ such that, for each $\sigma \in \Omega$, $f(\sigma)$ is determined by the specified procedure R_σ applied to $\{f(\tau) : \tau < \sigma\}$.

[We are still being vague about the precise meaning of "statement", "rule", and "procedure". There is also a technical difficulty in saying what is meant by a "function" whose domain is not a set; but the meaning in this context can be easily clarified in an intuitively acceptable way.] See II.8.7.4..

II.8.3.3. PROPOSITION. If X is a set and $f : \Omega \to X$ is a function defined recursively as in II.8.3.2., then there are ordinals σ and τ with $\sigma \neq \tau$ and $f(\sigma) = f(\tau)$.

This is because there cannot be a one-one correspondence between a set and a class which is not a set. (This follows from the Replacement Axiom of ZF set theory.)

II.8.4. Ordinals as Universal Models

We can now show that ordinals are universal models for well ordered sets. Recall that each ordinal is well ordered in its natural ordering.

II.8.4.1. PROPOSITION. No two distinct ordinals are order-isomorphic.

PROOF: It suffices to show that if $\tau < \sigma$, then there is no order-isomorphism from σ onto τ. Suppose f is such an order-isomorphism. The set $x = \{\rho \in \sigma : f(\rho) < \rho\}$ is nonempty, since $\tau \in \sigma$ and $f(\tau) < \tau$. Let ρ be the smallest element of x. Then, for every $\pi < \rho$, $\pi \leq f(\pi) < f(\rho)$ since f is an order-isomorphism. But ρ is the smallest ordinal larger than all $\pi < \rho$, so $\rho \leq f(\rho)$, a contradiction. 👍

II.8.4.2. THEOREM. Let $(I, \precsim)$ be a well ordered set. Then there is a unique ordinal σ such that $(I, \precsim)$ is order-isomorphic to $(\sigma, \leq)$.

PROOF: First fix an $x \notin I$ (for example, $x = I$.) Now define a function $f : \Omega \to I \cup \{x\}$ by transfinite recursion, by letting $f(\sigma)$ be the first element of $I \setminus \{f(\tau) : \tau < \sigma\}$ if this set is nonempty, and $f(\sigma) = x$ otherwise. Then f is not injective by II.8.3.3.; but $f(\sigma) = f(\tau)$ for $\sigma \neq \tau$ can only happen if $f(\sigma) = f(\tau) = x$ by the way f is defined. Thus there is a smallest σ such that $f(\sigma) = x$, and f is a bijection from $\sigma = \{\tau \in \Omega : \tau < \sigma\}$ onto I. It is easily checked that f is an order-isomorphism. Uniqueness follows from II.8.4.1.. 👍

The device in this argument of adding a new element x and applying II.8.3.3. is commonly used, for example in the proof in II.8.5.2. and in Exercise (). It can also be used to carefully justify arguments such as the proof of (), where a function is defined by transfinite recursion until the definition no longer "works".

II.8.5. The Axiom of Choice and Zorn's Lemma

Using ordinals and an argument like the proof of II.8.4.2., we can finally prove that the Axiom of Choice implies Zorn's Lemma. We begin with a lemma:

II.8.5.1. LEMMA. Let $(X, \precsim)$ be a partially ordered set in which every chain has an upper bound. If X has no maximal elements, then every chain C in X has an upper bound in X which is not in C.

PROOF: Suppose C is a chain. We must show there is an upper bound for C which is not in C. By assumption, C has an upper bound x. If $x \notin C$, we are done. If $x \in C$, then x is the largest element of C. By assumption, x is not a maximal element of X, so there is a $y \in X$ with $x < y$. Then y is an upper bound for C which is not in C. 👍

II.8.5.2. Now assume the Axiom of Choice, and let $(X, \precsim)$ be a partially ordered set in which every chain has an upper bound. Let $d : \mathcal{P}(X) \setminus \{\emptyset\} \to X$ be a choice function for X. Suppose that X has no maximal elements. Then by the lemma, for any chain C, the set U_C of upper bounds for C which are not in C is nonempty. Fix $y \notin X$ (say $y = X$) and define $f : \Omega \to X \cup \{y\}$ by transfinite recursion as follows. Let $f(\sigma) = d(U_{C_\sigma})$ if $C_\sigma = \{f(\tau) : \tau < \sigma\}$ is a chain, and $f(\sigma) = y$ otherwise. It is then easily proved by transfinite recursion that C_σ is a chain for every σ, so $f(\sigma) = d(U_{C_\sigma})$ for all σ; hence $f : \Omega \to X$ is injective, contradicting II.8.3.3.. This contradicts the assumption that X has no maximal element. 👍

II.8.5.3. Of course, since Zorn's Lemma implies the Well-Ordering Principle, which implies the Axiom of Choice, the three are thus equivalent. A direct proof along the same lines as II.8.5.2. (but somewhat simpler) can be given that the Axiom of Choice implies the Well-Ordering Principle (Exercise ().)

II.8.6. Ordinal Arithmetic and Examples

In this section, we explicitly describe some of the standard ordinals, and also define arithmetic operations on the ordinals.

Finite Ordinals

II.8.6.1. The finite ordinals are essentially the natural numbers (along with 0.) In fact, in ZF set theory the natural numbers are defined to be the (nonzero) finite ordinals.

The finite ordinals are just constructed from $\emptyset$ by repeated application of the successor operation:
The first ordinal, $\emptyset$, is usually denoted 0.
The next ordinal, usually denoted 1, is $S(0) = \{0\} = \{\emptyset\}$ (note that $\{\emptyset\} \neq \emptyset$!)
The next ordinal, usually denoted 2, is

$$S(1) = 1 \cup \{1\} = \{0, 1\} = \{\emptyset, \{\emptyset\}\}.$$

The next ordinal, usually denoted 3, is

$$S(2) = 2 \cup \{2\} = \{0, 1, 2\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}.$$

This process continues to give $n = \{0, 1, 2, \ldots, n-1\}$ for each $n \in \mathbb{N}$.

Infinite Ordinals

II.8.6.2. The Axiom of Infinity in ZF set theory says that the class of all finite ordinals is a set, which is obviously an ordinal. This class, which is familiar as $\mathbb{N} \cup \{0\}$, is usually denoted ω when regarded as an ordinal (ω_0 might be better notation.) The ordinal ω is the *first infinite ordinal*, and is the first limit ordinal (other than 0.) The ordinals larger than ω are also called *infinite ordinals*.

One can continue generating infinite ordinals via the successor operation, such as $S(\omega)$, $S(S(\omega))$, etc. These can be much more efficiently described using ordinal arithmetic, so we will defer further discussion of them.

II.8.6.3. The Well-Ordering Principle implies that there is an uncountable well ordered set, and hence an uncountable ordinal by II.8.4.2.. It is perhaps surprising that the existence of an uncountable ordinal can be proved without resorting to the Axiom of Choice (apply II.9.4.6. with $X = \mathbb{N}$). The first uncountable ordinal is usually denoted ω_1 (some authors use Ω, which we reserve for the class of all ordinals.) The ordinal ω_1 is the set of all countable ordinals.

Similarly, the Well-Ordering Principle implies that there are ordinals of arbitrary cardinality. We will discuss this point more completely in the next section.

Ordinal Arithmetic

We can define addition, multiplication, exponentiation, and other algebraic operations on the set of ordinals. We begin with addition.

II.8.6.4. If ρ and σ are ordinals, we informally define $\rho + \sigma$ to be the ordinal of the well-ordered set obtained by putting a copy of the well-ordered set σ at the end of the well-ordered set ρ. More precisely, if ρ is a fixed ordinal, we define $\rho + \sigma$ to be $f(\sigma)$, where f is defined by transfinite recursion according to the following rule:

$f(0) = \rho$.

If $\sigma = S(\tau)$, then $f(\sigma) = S(f(\tau))$.

If σ is a nonzero limit ordinal, then $f(\sigma) = \lim\{f(\tau) : \tau < \sigma\}$

We have that $\rho + 1 = S(\rho)$ for every ρ. In particular, $\{\omega + 1, \omega + 2, \dots\}$ is the strictly increasing sequence of immediate successors of ω. The limit of this set is $\omega + \omega$.

II.8.6.5. Addition of finite ordinals coincides with usual addition of numbers, and has the usual familiar algebraic properties. However, most of the properties break down for general ordinals. For example, $\omega + 1 = S(\omega) \neq \omega$, but it is easily checked that $1 + \omega = \omega$; thus addition of general ordinals is not commutative, and the right cancellation law also fails ($\rho + \tau = \sigma + \tau$ does not imply $\rho = \sigma$.) It can be proved by transfinite induction that the associative law does hold in general: $(\rho + \sigma) + \tau = \rho + (\sigma + \tau)$ for any ρ, σ, τ (Exercise ().)

II.8.6.6. Subtraction of ordinals can be defined. If $\tau < \sigma$, then there is a unique ρ such that $\sigma = \tau + \rho$: ρ has the order type of $\{\pi : \tau \leq \pi < \sigma\}$.

II.8.6.7. We can then define multiplication of ordinals in terms of addition. Informally, the well ordered set $\rho \cdot \sigma$ is obtained by taking the well ordered set σ and replacing each element by a copy of ρ. One can describe this to be the set $\sigma \times \rho$ (note the order!) with lexicographic ordering. There is also an inductive definition. If ρ is a fixed ordinal, we define $\rho \cdot \sigma$ to be $f(\sigma)$, where f is defined by transfinite recursion according to the following rule:

$f(0) = 0$.

If $\sigma = S(\tau)$, then $f(\sigma) = f(\tau) + \rho$.

If σ is a nonzero limit ordinal, then $f(\sigma) = \lim\{f(\tau) : \tau < \sigma\}$

Then, for any ρ, $\rho \cdot 1 = \rho$, $\rho \cdot 2 = \rho + \rho$, etc.

II.8.6.8. As with addition, multiplication of finite ordinals coincides with ordinary multiplication of numbers, with the usual rules. But the commutative and right cancellation laws fail in general: $\omega \cdot 2 = \omega + \omega$, but it is easily checked that $2 \cdot \omega = \omega$. Ordinal multiplication is associative (Exercise().)

II.8.6.9. The next step is exponentiation. If ρ is a fixed nonzero ordinal, we define ρ^σ to be $f(\sigma)$, where f is defined by transfinite recursion according to the following rule:

$f(0) = 1$.

If $\sigma = S(\tau)$, then $f(\sigma) = f(\tau) \cdot \rho$.

If σ is a nonzero limit ordinal, then $f(\sigma) = \lim\{f(\tau) : \tau < \sigma\}$

We define $0^\sigma = 0$ for all $\sigma \neq 0$ (0^0 is not defined.)

As usual, ordinal exponentiation coincides with usual exponentiation on the finite ordinals. In fact, for any ρ, we have $\rho^2 = \rho \cdot \rho$, etc.

This hierarchy can be continued, as will be indicated in the next subsection.

Countable Ordinals

We now give a more complete description of at least some of the countable ordinals.

First of all, note that the set of countable ordinals is closed under operations of finite or countable character:

II.8.6.10. Proposition.

(i) If σ is a countable ordinal, so is $S(\sigma) = \sigma + 1$.

(ii) If $x = \{\sigma_i : i \in I\}$ is a countable set of countable ordinals, then $\lim(x)$ is countable.

(iii) If ρ and σ are countable ordinals, then $\rho + \sigma$, $\rho \cdot \sigma$, and ρ^σ are countable.

Proof: Parts (i) and (ii) follow immediately from () (the Countable AC is needed for (ii)), and (iii) is easily proved by transfinite induction using (i) and (ii). 👍

Countable ordinals can be pictured as subsets of $\mathbb{R}$ (in fact, of $\mathbb{Q}$ or $\mathbb{Q} \cap [0, 1)$):

II.8.6.11. Proposition. Any subset of $\mathbb{R}$ which contains no strictly decreasing sequences is order-isomorphic (with respect to the induced ordering) to a countable ordinal. Every countable ordinal is order-isomorphic to such a set.

Proof: Any such set X is well ordered in the induced ordering, hence is order-isomorphic to an ordinal σ. To see that X is countable, let $g : \sigma \to X$ be an order-isomorphism, and for each $\tau < \sigma$, let $f(\tau)$ be a rational number between $g(\tau)$ and $g(\tau + 1)$ (to make this precise and avoid using the Axiom of Choice, let $f(\tau)$ be the first element of $\mathbb{Q} \cap (g(\tau), g(\tau + 1))$ according to some fixed enumeration of $\mathbb{Q}$.) Then f is an injective function from σ to $\mathbb{Q}$. The last statement follows from II.5.4.13.. 👍

Note, however, that there is no *natural* or *standard* way of picturing a countable ordinal as a subset of $\mathbb{R}$.

II.8.6.12. Some of the countable ordinals described earlier can be pictured in $[0,\infty)$ as follows:

ω as $\{1-\frac{1}{n} : n\in\mathbb{N}\}$ (or as $\mathbb{N}\cup\{0\}$)

$\omega+1$ as $\{1-\frac{1}{n} : n\in\mathbb{N}\}\cup\{1\}$

$\omega+2$ as $\{1-\frac{1}{n} : n\in\mathbb{N}\}\cup\{1,2\}$

$\omega+n$ as $\{1-\frac{1}{n} : n\in\mathbb{N}\}\cup\{1,\dots,n\}$

$\dots$

$\omega\cdot 2=\omega+\omega$ as $\{1-\frac{1}{n},2-\frac{1}{n} : n\in\mathbb{N}\}$ (or as $\{1-\frac{1}{n} : n\in\mathbb{N}\}\cup\mathbb{N}$)

$\omega\cdot 2+1$ as $\{1-\frac{1}{n},2-\frac{1}{n} : n\in\mathbb{N}\}\cup\{2\}$

$\dots$

$\omega^2=\omega\cdot\omega$ as $\{1-\frac{1}{n},1-\frac{1}{n}-\frac{1}{n^2m} : m,n\in\mathbb{N}\}$

$\dots$

$\omega^2+\omega$ as $\{1-\frac{1}{n},2-\frac{1}{n},1-\frac{1}{n}-\frac{1}{n^2m} : m,n\in\mathbb{N}\}$

$\dots$

$\omega^2\cdot 2=\omega^2+\omega^2$ as $\{1-\frac{1}{n},2-\frac{1}{n},1-\frac{1}{n}-\frac{1}{n^2m},2-\frac{1}{n}-\frac{1}{n^2m} : m,n\in\mathbb{N}\}$

$\dots$

$\omega^3=\omega^2\cdot\omega$ as $\{1-\frac{1}{n},1-\frac{1}{n}-\frac{1}{n^2m},1-\frac{1}{n}-\frac{1}{n^2m}-\frac{1}{n^2m^2k} : m,n,k\in\mathbb{N}\}$

The intermediate and succeeding ordinals can be drawn in a similar manner. Note that the blocks between the ordinals shown grow successively longer: between ω^2 and ω^3 are all the ordinals of the form $\omega^2\cdot n+\omega\cdot m+k$, for any $m,n,k\in\mathbb{N}\cup\{0\}$, $n>0$.

II.8.6.13. We can make much larger countable ordinals. For example, $\omega^\omega=\lim\{\omega^n : n\in\mathbb{N}\}$. The elements of ω^ω (i.e. the set of ordinals less than ω^ω) can be pictured as a subset of $\mathbb{R}$ (try it!) There is another way to describe this as an ordered set: they can be identified with the set of all polynomials with coefficients in $\mathbb{N}\cup\{0\}$, thought of as functions from $\mathbb{N}$ to $\mathbb{N}$, with $f<g$ if $f(n)<g(n)$ for all sufficiently large n. (This ordering is called the *Fréchet ordering*.)

II.8.6.14. We can go much farther. We can form ω^{ω^ω}, $\omega^{\omega^{\omega^\omega}}$, etc. [As usual, these are grouped from the top down, e.g. $\omega^{\omega^\omega}=\omega^{(\omega^\omega)}$.] There is a limit to this sequence, i.e. a smallest ordinal larger than all these, which is countable and usually called ϵ_0. See II.8.9.2. for a continuation.

II.8.6.15. The vertical arrow notation of Donald Knuth [Knu76] is useful here: if we define $\rho\uparrow\sigma=\rho^\sigma$ for any ρ and σ, we can define

$$\rho\uparrow\uparrow n=\rho\uparrow(\rho\uparrow(\rho\uparrow\cdots(\rho\uparrow\rho)\cdots))=\rho^{\rho^{\rho^{\cdots^\rho}}}$$

(n levels), and $\rho\uparrow\uparrow\omega=\lim\{\rho\uparrow\uparrow n : n\in\mathbb{N}\}$. More generally, $\rho\uparrow\uparrow\sigma$ for any ρ,σ can be defined inductively by setting $\rho\uparrow\uparrow S(\sigma)=\rho\uparrow(\rho\uparrow\uparrow\sigma)=\rho^{\rho\uparrow\uparrow\sigma}$ and $\rho\uparrow\uparrow(\lim(\sigma_i))=\lim(\rho\uparrow\uparrow\sigma_i)$. (This operation is called *tetration*, and $\rho\uparrow\uparrow\sigma$ is sometimes written ${}^\sigma\rho$.) With this notation, $\epsilon_0=\omega\uparrow\uparrow\omega$.

II.8.6.16. We can go on and define $\rho \uparrow\uparrow\uparrow \sigma$ starting by

$$\rho \uparrow\uparrow\uparrow n = \rho \uparrow\uparrow (\rho \uparrow\uparrow (\rho \uparrow\uparrow \cdots (\rho \uparrow\uparrow \rho) \cdots))$$

(n levels), and then $\rho \uparrow\uparrow\uparrow \sigma$ inductively for all σ. Similarly, for any n we can define $\rho \uparrow\uparrow\uparrow\uparrow \sigma$, $\rho \uparrow^n \sigma = \rho \underbrace{\uparrow \cdots \uparrow}_{n} \sigma$, etc. The sequence

$$\{\omega \uparrow\uparrow \omega, \omega \uparrow\uparrow\uparrow \omega, \omega \uparrow\uparrow\uparrow\uparrow \omega, \dots\}$$

has a limit which is again a countable ordinal.

II.8.6.17. We can still go on. We can define $\rho \uparrow^\sigma \tau$ for any ρ, σ, τ (Exercise ()), and then $\rho \Uparrow \sigma = \rho \uparrow^\sigma \rho$,

$$\rho \Uparrow\Uparrow n = \rho \Uparrow (\rho \Uparrow (\rho \Uparrow \cdots (\rho \Uparrow \rho) \cdots))$$

(n levels), $\rho \Uparrow\Uparrow \sigma$, $\rho \Uparrow\Uparrow\Uparrow \sigma$ etc., etc., ... (see also II.8.9.4..) And if we run out of operations, we can always take a (countable) limit, take the successor of it, and start over again!

Even at that we have barely scratched the surface of the set of all countable ordinals. An internet search of "countable ordinals" will turn up numerous articles, from popular to technical, on naming or describing countable ordinals, some of which go far beyond what we have done in depth and sophistication (but even so fall far short of exhausting the countable ordinals). As Rudy Rucker puts it in his book, *Infinity and the Mind* [Ruc82],

> "In trying to think of bigger and bigger ordinals, one sinks into a kind of endless morass. Any procedure you come up with for naming larger ordinals eventually peters out, and the ordinals keep on coming. Finally, your mind snaps, and maybe you get a momentary glimpse of what the Absolute infinite is about. Then you try to formalize your glimpse, and you end up with a new system for naming ordinals ... which eventually peters out ..."

And this comment was just made in reference to the countable ordinals!

Countable ordinals can be pictured as in Figure II.1 (only the length varies). Try drawing pictures of 2ω, ω^2, ω^3, ω^ω and, if you are brave, ω^{ω^ω}, $\epsilon_0 = {}^\omega\omega$ and any larger ones you can. Identify (or estimate) the ordinal pictured in Figure II.1.

II.8.6.18. The first uncountable ordinal ω_1 cannot be pictured as a subset of $\mathbb{R}$ (it looks like Figure II.1 but is "too long" to fit even on a line of countably infinite length). However, it is order-isomorphic to a set of functions from $\mathbb{N} \cup \{0\}$ to $\mathbb{N} \cup \{0\}$ with Fréchet ordering, although there is no natural way to do it. An order-embedding $f : \sigma \mapsto f_\sigma$ can be defined by transfinite recursion as follows. Let f_0 be the constant function 0. If $\sigma = \tau + 1$, let $f_\sigma(n) = f_\tau(n) + 1$ for all n. Now suppose σ is a limit ordinal. It follows easily from II.8.6.11. that there is a strictly increasing sequence (σ_n) of ordinals with $\sigma = \lim(\sigma_n)$. Choose such a sequence (some version of the Axiom of Choice is needed to do this rigorously). Set $k_1 = 1$, and for $n > 1$ fix $k_n \in \mathbb{N}$ with $f_{\sigma_n}(m) > f_{\sigma_{n-1}}(m)$ for all $m \geq k_n$. Then set $f_\sigma(m) = f_{\sigma_n}(m)$ if $k_n \leq m < k_{n+1}$.

It is easily checked that this f is an order-embedding. The function f can be chosen to use the polynomial embedding for ω^ω, and can be extended in a standard way to the ordinals previously considered using the functions $x \uparrow\uparrow n$, $x \uparrow\uparrow x$, $x \Uparrow\Uparrow n$, etc. [These formulas all give functions from $\mathbb{N}$ to $\mathbb{N}$, although they grow unimaginably fast: for example, even

$$4 \uparrow\uparrow 4 = 4^{4^{4^4}}$$

is a number of more than 10^{154} digits, much larger than a googolplex, and

$$3 \uparrow\uparrow\uparrow 3 = 3 \uparrow\uparrow (3 \uparrow\uparrow 3) = 3 \uparrow\uparrow 3^{27} = 3^{3^{3^{\cdots^{3}}}}$$

(3^{27} = 7,625,597,484,987 levels) is far too large to write using conventional notation.]

II.8.7. Induction and Recursion

Induction and recursion are two of the most important applications of ordinals. Induction is quite straightforward, but recursion is considerably more subtle.

Transfinite Induction

II.8.7.1. Proposition. [Principle of Transfinite Induction] (i) Let α be an ordinal, and $x \subseteq \alpha$. Suppose x has the following properties:

(a) $\emptyset \in x$.

(b) For all $\beta < \alpha$, if $\gamma \in x$ for all $\gamma < \beta$, then $\beta \in x$.

Then $x = \alpha$.
(ii) Let C be a subclass of Ω. Suppose C has the following properties:

(a) $0 \in C$.

(b) For all $\beta \in \Omega$, if $\gamma \in C$ for all $\gamma < \beta$, then $\beta \in C$.

Then $C = \Omega$.

Proof: Both parts are essentially the same. Suppose there is an ordinal σ not in x (or C). The set y of ordinals less than or equal to σ not in x or C is nonempty, hence has a smallest element β. If $\beta = 0$, this contradicts (a), and otherwise any $\gamma < \beta$ is in x or C, contradicting (b). 👍

II.8.7.2. Actually (a) is unnecessary in each case, since if $\beta = 0$, the hypothesis of (b) in each case is vacuous; but (a) is included in the statement for emphasis. We can rephrase (ii) slightly differently ((a) is needed for this version); (i) can be rephrased similarly:

II.8.7.3. Corollary. Let C be a subclass of Ω. Suppose C has the following properties:

(a) $0 \in C$.

(b) For all $\beta \in \Omega$, if $\beta \in C$, then $\beta + 1 \in C$.

(c) If $\{\beta_i : i \in I\}$ is a nonempty set of ordinals, and $\beta_i \in C$ for all $i \in I$, then $\sup_{i \in I} \beta_i \in C$.

Then $C = \Omega$.

Transfinite Recursion

II.8.7.4. THEOREM. [TRANSFINITE RECURSION] Let Φ be an admissible class-function with domain $\Omega \times V$, such that $\Phi(\alpha, x)$ is a set-function with $x \in Dom(\Phi(\alpha, x))$ for each $(\alpha, x) \in \Omega \times V$. Then there is a unique class-function F with domain Ω such that $F(\alpha) = [\Phi(\alpha, F \restriction \alpha)](F \restriction \alpha)$ for each $\alpha \in \Omega$. F is admissible, and definable from Φ.

Informally, if for each (α, x) there is a rule $\Phi(\alpha, x)$ (with a mild compatibility condition on the $\Phi(\alpha, x)$ that $(\alpha, x) \mapsto \Phi(\alpha, x)$ is an admissible function) for computing $F(\alpha)$ from $\{F(\beta) : \beta < \alpha\}$ (which turns out to necessarily be a set), i.e. F can be computed "one step at a time," then the entire F exists (i.e. can be specified "all at once.") The rules $\Phi(\alpha, x)$ can depend on α and even on x as long as the compatibility condition is satisfied. When working in a model M, the function Φ need not be definable – it only has to be an M-admissible function.

PROOF: There is a subtlety in the proof, which is not adequately explained in most references (I am indebted to A. KARAGILA for explaining it to me). If α is an ordinal, say that a function f_α is an *α-approximation* to F if f_α has domain α and

$$f_\alpha(\beta) = [\Phi(\beta, f_\alpha \restriction \beta)](f_\alpha \restriction \beta)$$

for all $\beta < \alpha$. Let C be the class of all ordinals α for which there exists a unique α-approximation to F.

We claim $C = \Omega$ by transfinite induction. In fact, we will show something stronger: not only does f_α exist uniquely for all α, but these functions are also definable and coherent, i.e. if $\gamma < \beta < \alpha$, then $f_\alpha(\gamma) = f_\beta(\gamma)$. First, set $f_0 = 0 = \emptyset$. Then f_0 is a function with domain $\emptyset$, and is a 0-approximation to F (and is the only one and is obviously definable). Now suppose f_β exists and is unique for all $\beta < \alpha$, and that they are definable and coherent. If $\alpha = \sigma + 1$ for some σ (i.e. $\alpha = \{\beta : \beta \leq \sigma\}$), then we may define f_α as follows: if $\beta < \sigma$, set $f_\alpha(\beta) = f_\sigma(\beta)$, and set

$$f_\alpha(\sigma) = [\Phi(\alpha, ran(f_\sigma))](ran(f_\sigma)) \ .$$

Then f_α is a definable function, and the only possible α-approximation to F.

Now suppose α is a limit ordinal (this is the subtle part of the proof). The idea is straightforward: we have a coherent family $s = \{f_\beta : \beta < \alpha\}$ and we want to just put them together to get a function with domain $\alpha = \cup\{\beta : \beta < \alpha\}$. The subtle point is that we must show that s is a set. If we just have a coherent family of definable functions indexed by α, with no relation between the definitions, then the definitions may get more and more complicated as the index increases, so the functions cannot necessarily be combined. Even in recursion this appears to happen since Φ is applied over and over, with no limit on the number of times. But in the recursion case, the f_α are actually defined by a single formula, with ordinals as a parameter, and the complexity of the formula is encoded in the parameter (see ()). Thus, if we set

$$h = \{(\beta, f_\beta) : \beta < \alpha\}$$

we obtain that h is a definable set-function, so $s = ran(h)$ is a set. We define $f_\alpha = \cup s$. The elements of s are functions, and since the functions are coherent f_α is also a function. It is clearly a definable function and is coherent with the f_β for $\beta < \alpha$, and is the only possible α-approximation to F.

We thus have a coherent family $A = \{f_\alpha : \alpha \in \Omega\}$ of definable set-functions. Set $F = \cup A$. Then F is clearly a class-function. If X is a set, then $F \cap X$ is the class corresponding to $f_\alpha \cap x$ for some sufficiently large α, so F is admissible. 👍

Recursion is most commonly applied when there is a single admissible class-function G such that $\Phi(\alpha, x) = G|_{\{x\}}$ for each (α, x), i.e. the rules $\Phi(\alpha, x)$ are "all the same":

II.8.7.5. Corollary. [Transfinite Recursion, Common Case] Let G be an admissible class-function with domain V. Then there is a unique class-function F with domain Ω such that $F(\alpha) = G(F \restriction \alpha)$ for each $\alpha \in \Omega$. F is admissible, and if G is definable, then F is definable.

Proof: For $(\alpha, x) \in \Omega \times V$ set $\Phi(\alpha, x) = G|_{\{x\}} = \{(x, G(x)\}$. Then Φ is an admissible class-function with $Dom(\Phi(\alpha, x)) = \{x\}$ for each (α, x) (). 👍

II.8.8. The Theory of Levels

Using ordinals, we may give a description of the (ZF) set-theoretic universe, which can be used to give an alternate axiomatic approach to the entirety of set theory. The idea is that sets are constructed successively out of objects which are already available, e.g. sets which have already been constructed.

Levels Without Atoms

II.8.8.1. Definition. For each ordinal α, define V_α by transfinite induction as follows.

Set $V_0 = \emptyset$.

Set $V_{\alpha+1} = \mathcal{P}(V_\alpha)$.

If α is a limit ordinal, set $V_\alpha = \bigcup_{\beta<\alpha} V_\beta$.

II.8.8.2. Proposition. Every V_α is a transitive set, and $V_\beta \in V_\alpha$ (in particular, $V_\beta \subseteq V_\alpha$) if $\beta < \alpha$.

Proof: This is proved by transfinite induction, using the Empty Set, Power Set, and Replacement Axioms (cf. II.7.1.46., and using that ordinals are sets) to conclude that each V_α is a set, and II.8.9.5.(b) to obtain that each V_α is transitive. We clearly have $V_\beta \in V_{\beta+1}$, hence $V_\beta \subseteq V_{\beta+1}$, for all β; hence $V_\beta \in V_\alpha$ for all $\alpha > \beta$ by transfinite induction. 👍

II.8.8.3. Proposition. If $x \in V_\alpha$, then $\cup x \in V_{\alpha+1}$.

Proof: If $y \in x$, then $y \in V_\alpha$, so $y \subseteq V_\alpha$. Thus $\cup x \subseteq V_\alpha$. 👍

The main result about levels is:

II.8.8.4. THEOREM. If x is any set, then $x \in V_\alpha$ for some α.

PROOF: Let C be the class of all sets which are not in any V_α (this is a definable class, cf. ()). Suppose C is not empty, and let y be a $\in$-minimal element of C (II.7.2.5.). Then any $z \in y$ is in V_α for some α. For each $z \in y$, let $f(z)$ be the smallest α such that $z \in V_\alpha$. Then f is a function from y to the class of ordinals, and by Replacement the range is a set s of ordinals. If $\beta = \cup s$, then β is an ordinal, and $y \subseteq V_\beta$, so $y \in V_{\beta+1}$, a contradiction. Thus $C = \emptyset$. 👍

II.8.8.5. Thus we may say that the union (in the sense of **??**) of the V_α is the entire universe V. Thus every set is built up from $\emptyset$ by (transfinitely) repeated operations of subset, union, and power set. The sets in V_α are called the *sets defined by the α'th stage*, or the *sets definable in α steps*.

II.8.8.6. The class V is called the *von Neumann universe*. There is an apparently smaller universe L, called the *Gödel constructible universe*, where only "constructible" subsets are added at each stage. The hypothesis $V = L$ is called the *Axiom of Constructibility*.

Levels With Atoms

II.8.8.7. If we allow atoms, we must modify the theory of levels. We take V_0 to be the set of all atoms. (As discussed in (), it is debatable whether we should regard the collection of atoms as a set, but we must to get a theory of levels.) Note that V_0 is not transitive unless it is $\emptyset$. We must then define V_α transfinitely by setting $V_{\alpha+1} = V_\alpha \cup \mathcal{P}(V_\alpha)$, and $V_\alpha = \cup_{\beta<\alpha} V_\beta$ if α is a limit ordinal. The rest of the theory then works as before, and $\cup_\alpha V_\alpha$ is the entire universe V.

II.8.8.8. In this case, we define $V_\alpha^\circ = \emptyset$ and $V_{\alpha+1}^\circ = \mathcal{P}(V_\alpha^\circ)$, $V_\alpha^\circ = \cup_{\beta<\alpha} V_\beta^\circ$ if α is a limit ordinal, and $V^\circ = \cup_\alpha V_\alpha^\circ$. We have $V_\alpha^\circ \subseteq V_\alpha$ for each α and $V^\circ \subseteq V$. V° is called the *class of pure sets*.

The Constructible Universe

Levels as an Alternative Axiomatization of Set Theory

II.8.9. Exercises

II.8.9.1. Give a direct proof along the lines of () that the Axiom of Choice implies the Well-Ordering Principle. [Hint: If X is a set, fix $y \notin X$ and a choice function $c : \mathcal{P}(X) \setminus \{\emptyset\} \to X$; set $f(\sigma) = c(X \setminus \{f(\tau) : \tau < \sigma\})$ if this set is nonempty, and $f(\sigma) = y$ otherwise.]

II.8.9.2. An *ϵ-number* (defined by CANTOR) is an ordinal ϵ such that $\omega^\epsilon = \epsilon$.

(a) Show that ϵ_0 (II.8.6.14.) is the smallest ϵ-number.

(b) Show that the next ϵ-number is

$$\epsilon_1 = \epsilon_0 \uparrow\uparrow \omega = \lim(\epsilon_0^{\epsilon_0}, \epsilon_0^{\epsilon_0^{\epsilon_0}}, \dots) = \lim(\omega^{\epsilon_0+1}, \omega^{\omega^{\epsilon_0+1}}, \dots) \ .$$

(c) Show that the ϵ-numbers, in increasing order, are indexed by ordinals, and that ϵ_α is countable if (and only if) α is countable.

(d) The ordinal $\Gamma_0 = \lim(\epsilon_0, \epsilon_{\epsilon_0}, \epsilon_{\epsilon_{\epsilon_0}}, \dots)$ is the smallest ordinal α such that $\epsilon_\alpha = \alpha$. Γ_0 is countable. Define Γ-numbers, show they are indexed by ordinals, and find the smallest ordinal α such that $\Gamma_\alpha = \alpha$. Show that this ordinal is countable. Continue the process as far as your sanity allows.

II.8.9.3. (a) Give a careful inductive definition of $\rho \uparrow\uparrow \sigma$ as outlined in II.8.6.15..

(b) Give a careful inductive definition of $\rho \uparrow\uparrow\uparrow \sigma$ and $\rho \uparrow^n \sigma$ for each n.

(c) Use transfinite induction on σ to define $\uparrow^\sigma$ for all σ in terms of $\uparrow^\pi$ for all $\pi < \sigma$, i.e. suppose $\rho \uparrow^\pi \tau$ has been defined for all $\pi < \sigma$ and all ρ, τ. In defining $\rho \uparrow^\sigma \tau$, one may also assume that $\rho \uparrow^\sigma \pi$ has been defined for all $\pi < \tau$ (this assumption must be justified; consider lexicographic ordering.)

(d) Give a careful inductive definition of $\rho \Uparrow \sigma$, $\rho \Uparrow\Uparrow \sigma$, $\rho \Uparrow^\sigma \tau$, etc. Continue the hierarchy as far as possible.

II.8.9.4. Define Conway's "chained arrow notation" as follows. All latin letters denote natural numbers.

(i) $a \to b$ denotes a^b.

(ii) $a \cdots x \to y \to 1$ denotes $a \cdots x \to y$.

(iii) $a \cdots x \to y \to (z+1)$ denotes

$a \cdots x$ if $y = 1$

$a \cdots x \to (a \cdots x) \to z$ if $y = 2$

$a \cdots x \to (a \cdots x \to (a \cdots x) \to z) \to z$ if $y = 3$

etc.

(a) Show that $a \to b \to c = a \uparrow^c b$ for any a, b, c.

(b) Show how to define $\alpha \to \beta \to \cdots \to \sigma$ for any ordinals $\alpha, \beta, \dots, \sigma$. Show that these are again countable ordinals if $\alpha, \beta, \dots, \sigma$ are countable.

(c) Define $a \Rightarrow b$ to be $a \to a \to \cdots \to a$ (b terms). Define $a \Rightarrow b \Rightarrow \cdots \Rightarrow x$ inductively as for $a \to b \to \cdots \to x$. Extend this to show how to define $\alpha \Rightarrow \beta \Rightarrow \cdots \Rightarrow \sigma$ for any ordinals $\alpha, \beta, \dots, \sigma$. Show that these are again countable ordinals if $\alpha, \beta, \dots, \sigma$ are countable.

II.8.9.5. **Transitive Sets.** (a) Let x be a set. Show that the following are equivalent:

(i) x is transitive, i.e. $y \in x \Rightarrow y \subseteq x$.

(ii) If $y \in x$ and $z \in y$, then $z \in x$.

(b) Show that the following are equivalent:

(i) x is transitive.

(ii) $\mathcal{P}(x)$ is transitive.

(iii) $x \subseteq \mathcal{P}(x)$.

(c) Show that every nonempty transitive set contains $\emptyset$ as an element. [If x is transitive, consider a minimal element of x.]

(d) If x is any set, show there is a smallest transitive set $tc(x)$ containing x as a subset. [Let $x_1 = x$ and $x_{n+1} = \cup x_n = \cup\{y : y \in x_n\}$; then $tc(x) = \cup\{x_n : n \in \mathbb{N}\}$ is transitive, and is the smallest transitive set containing x.] The set $tc(x)$ is called the *transitive closure* of x.

(e) If x is a transitive set, show that not all subsets of x are elements of x. [Use Cantor's Theorem II.5.3.3..]

II.8.9.6. Ordinals without Regularity. The theory of ordinals can be developed without use of the Axiom of Regularity, as follows.

(a) Prove the following without using the Axiom of Regularity:

PROPOSITION. Let x be a transitive set. The following are equivalent:

(i) Every element of x is transitive.

(ii) Membership is a transitive relation on the elements of x.

(iii) If $n \in \mathbb{N}$ and $x_n \in x_{n-1} \in \cdots \in x_1 \in x$, then x_n is transitive and $x_n \in x$.

[Prove (i) $\Rightarrow$ (iii) by induction on n, and (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i).]

(b) Make the following definition:

DEFINITION. A set x is an *ordinal* if x is transitive and membership is a strict well ordering on the elements of x.

Show that an ordinal by this definition is an ordinal by Definition II.8.1.1., and that the converse holds under the Axiom of Regularity. Show that subsets of ordinals (by this definition) automatically satisfy the Axiom of Regularity.

(c) Show that the theory of ordinals as developed in this section holds for ordinals by this definition without use of the Axiom of Regularity (i.e. the Axiom of Regularity is not used after II.8.2.4.).

II.8.9.7. The ZF Universe. Define V_α as in (). Then $V = \cup_\alpha V_\alpha$ is the class of all sets.

(a) Show that if x is a set and $x \subseteq V$, then $x \in V$. [For each $y \in x$, let σ_y be the smallest τ such that $y \in V_\tau$, and set $\alpha = \lim\{\sigma_y : y \in x\}$.]

(b) Show that if $x \in V_\alpha$, then $\cup x \in V_{\alpha+1}$.

(c) Give an explicit description of the sets in V_ω. (Note that all these sets are finite, and $card(V_\omega) = \aleph_0$.)

II.8.9.8. An ordinal τ is called *even* if it is a limit ordinal or of the form $\rho +_o 2n$ for a limit ordinal ρ; otherwise τ is *odd*. If σ is an ordinal, let σ_{ev} and σ_{odd} be the set of even and odd ordinals less than σ, respectively. Then σ_{ev} and σ_{odd} are disjoint, and $\sigma_{ev} \cup \sigma_{odd} = \sigma$. Show that if σ is a limit ordinal, then the maps $f(\rho +_o 2n) = \rho +_o n$ and $f(\rho +_o 2n +_o 1) = \rho +_o n$, as ρ runs over all limit ordinals less than σ, give bijections from σ_{ev} and σ_{odd} onto σ.

II.8.9.9. (a) Show that if $\rho \geq 1$ and σ, τ are ordinals, then $\rho \cdot \sigma < \rho \cdot \tau$ if and only if $\sigma < \tau$. (One direction is false in general for $\sigma \cdot \rho$ and $\tau \cdot \rho$: $1 \cdot \omega = 2 \cdot \omega$.)

(b) Show that if $\rho > 1$ and σ, τ are ordinals, then $\rho^\sigma < \rho^\tau$ if and only if $\sigma < \tau$.

II.8.9.10. "Decimal" Expansions of Ordinals. ([KM76], [Pot04]) Fix an ordinal $\beta > 1$ (the "base").
(a) Show that if σ is any ordinal, there are ordinals η, γ, ρ such that

$$\sigma = \beta^\eta \cdot \gamma + \rho$$

with $\gamma < \beta$ and $\rho < \beta^\eta$. If $\sigma \leq \beta^\tau$, then $\eta \leq \tau$, and if If $\sigma < \beta^\tau$, then $\eta < \tau$.
(b) Show that if $\sigma \geq 1$ is any nonzero ordinal, there is a natural number n and ordinals $\eta_1, \ldots, \eta_n$ and $\gamma_1, \ldots, \gamma_n$ with $\eta_1 > \eta_2 > \cdots > \eta_n \geq 0$ and $0 < \gamma_k < \beta$ for all k, such that

$$\sigma = \beta^{\eta_1} \cdot \gamma_1 + \cdots + \beta^{\eta_n} \cdot \gamma_n \ .$$

(c) Show that the expansion in (b) is unique. It is called the *normal form* of σ with base β.

This generalizes ordinary "decimal" expansions (the case where β and σ are finite). But ordinal expansions can be degenerate: for example, if $b > 1$ is a natural number, the base b expansion of ω is $\omega = b^\omega \cdot 1$; and if ϵ is an ϵ-number (II.8.9.2.), then $\epsilon = \omega^\epsilon \cdot 1$ is the base ω expansion of ϵ.

II.8.9.11. "Decimal" Expansions of Ordinals Continued. Use the language of II.8.9.10..
(a) For any $\beta > 1$, there is a smallest ordinal σ such that $\sigma = \beta^\sigma$. This σ is the *salient ordinal* of β. If β is finite, its salient ordinal is ω; the salient ordinal of ω is ϵ_0.
(b) If σ is less than the salient ordinal of β, then in the normal form

$$\sigma = \beta^{\eta_1} \cdot \gamma_1 + \cdots + \beta^{\eta_n} \cdot \gamma_n$$

of σ with base β, we have $\eta_1 < \sigma$. Using the normal forms of the η_j we may thus write

$$\sigma = \beta^{(\beta^{\eta_{11}} \cdot \gamma_{11} + \cdots + \beta^{\eta_{1m_1}} \cdot \gamma_{1m_1})} \cdot \gamma_1 + \cdots + \beta^{(\beta^{\eta_{n1}} \cdot \gamma_{n1} + \cdots + \beta^{\eta_{nm_n}} \cdot \gamma_{nm_n})} \cdot \gamma_n$$

where $\eta_{jk} < \eta_j$ for all j, k.
(c) If $\eta_{11} > \beta$, repeat the process to get third-order exponents. Show that the process terminates after a finite number of steps, giving an expression of finite height in which all ordinals are $\leq \beta$. This expression is unique (if factors and exponents of 1 are omitted), and is the *complete normal form* of σ with base β.

The complete normal form is of interest even if β is finite, so σ must also be finite. For example, if $\beta = 2$ and $\sigma = 8201$, we have

$$8201 = 2^{13} + 2^3 + 1 = 2^{2^{2+1}+2+1} + 2^{2+1} + 1 = 2^{2^{2^1 \cdot 1 + 2^0 \cdot 1} + 2^1 \cdot 1 + 2^0 \cdot 1} \cdot 1 + 2^{2^1 \cdot 1 + 2^0 \cdot 1} \cdot 1 + 2^0 \cdot 1 \ .$$

We usually take the middle expression for notational ease, leaving out factors and exponents of 1. If $\beta = 3$ and $\sigma = 2350$, we have

$$2350 = 3^7 + 3^4 \cdot 2 + 1 = 3^{3 \cdot 2 + 1} + 3^{3+1} \cdot 2 + 1 \ .$$

II.8.9.12. The Axiom of Infinity Revisited. Replace the inductive form of the Axiom of Infinity by apparently weaker axiom "There exists a T-infinite set." The definition of ordinal still makes sense (we can also eliminate the Axiom of Regularity by taking the definition of ordinal in II.8.9.6.). Say an ordinal σ is *finite* if it is T-finite, i.e. every collection of subsets has an $\subseteq$-maximal element.
(a) Show that $\omega = \{\sigma : \sigma \text{ is a finite ordinal}\}$ is a (definable) class.

(b) Show that if y is any T-finite set, then there is a unique finite ordinal n equipotent with y. [Consider the set $\mathcal{S}$ of subsets of y equipotent to a finite ordinal. For uniqueness, use () and that a finite ordinal is Dedekind-finite.]

(c) Suppose x is a T-infinite set. Define a function F from $\mathcal{P}(x)$ to ω by setting $F(y)$ to be the unique finite ordinal equipotent to y if y is T-finite, and $F(y) = \emptyset$ if y is T-infinite. Show that $Ran(F)$ is all of ω. By the Axiom of Replacement, ω is a set.

(d) Show that ω is inductive, and is the smallest inductive set N of II.7.1.35.. Thus the T-infinite form of the Axiom of Infinity is equivalent to the inductive form (in the presence of the other ZF axioms other than Regularity). We also have $\omega = \cup\omega$ and is an ordinal, the smallest limit ordinal.

II.9. Cardinals

We now want to make precise the notion of the cardinality of a set. In the course of this section, we will define $card(X)$ for any set X. To avoid confusion, we will use $|X|$ to denote the informal notion of cardinality developed in (); so the statements $|X| = |Y|$, $|X| \leq |Y|$, $|X| < |Y|$ make sense for sets X, Y, although we avoid attaching any meaning to $|X|$ itself.

The idea is that we want to have a standard "model set" $card(X)$ for any set X, with the properties that $|card(X)| = |X|$ for any X and, for any X, Y, $card(X) = card(Y)$ if and only if $|X| = |Y|$. This process only really works well in the presence of the Axiom of Choice. Indeed, what we want to do is a form of Choice, identifying one model set from each equipotence equivalence class. We will, however, not make a blanket assumption of the Axiom of Choice, since the delineation of what can and cannot be proved without it gives good insight into the nature and role of the Axiom of Choice. Results which are dependent on the Axiom of Choice will be denoted by (AC); some of these can be obtained using weakened Choice axioms such as Dependent Choice ().

II.9.1. Cardinals as Ordinals

Cardinals are ordinals of a special type:

II.9.1.1. Definition. An ordinal σ is a *cardinal*, or *cardinal number*, if $|\tau| < |\sigma|$ for every $\tau < \sigma$. A cardinal is *finite* [*infinite*] if it is finite [infinite] as an ordinal.

The letters κ, λ, χ are frequently used to denote general cardinals, and $\aleph$ is also used to denote a general infinite cardinal (if the Axiom of Choice is not assumed, the symbol $\aleph$ is reserved for only certain infinite cardinals: see ().) The class of cardinals has a natural ordering induced from Ω, under which it is well ordered (in particular, totally ordered.)

II.9.1.2. Examples.

(i) Any finite ordinal is a cardinal.

(ii) ω is a cardinal. No other countable infinite ordinal is a cardinal.

(iii) ω_1 is a cardinal.

When regarded as cardinals, ω and ω_1 are usually denoted $\aleph_0$ and $\aleph_1$ respectively. $\aleph_0$ is the smallest infinite cardinal, and $\aleph_1$ is the smallest uncountable cardinal and is the smallest cardinal greater than $\aleph_0$.

II.9.1.3. Proposition. If κ and λ are cardinals, then $|\kappa| < |\lambda|$ if and only if $\kappa < \lambda$.

This is obvious from the definition.

So far, we have not needed the Axiom of Choice. But the Axiom of Choice is crucial in the next result, which is the key fact about cardinals.

II.9.1.4. THEOREM. (AC) Let X be any set. Then there is a unique cardinal κ such that $|X| = |\kappa|$.

PROOF: By the Well-Ordering Principle, there is a $\sigma \in \Omega$ with $|X| = |\sigma|$. Let κ be the smallest such σ. Then κ is clearly a cardinal. Uniqueness follows from II.9.1.3.. 👍

[This result is actually equivalent to the Well-Ordering Principle, as is easily seen.]

II.9.1.5. DEFINITION. If X is a set, then the *cardinality* of X, denoted $card(X)$, is the unique cardinal κ with $|X| = |\kappa|$.

II.9.1.6. EXAMPLES.

(i) If X is a finite set with n elements, then $card(X) = n$.

(ii) If X is countably infinite, then $card(X) = \aleph_0$.

(iii) If κ is a cardinal, then $card(\kappa) = \kappa$.

II.9.1.7. If $Y \subseteq X$, then $card(Y) \leq card(X)$. We can have $card(Y) = card(X)$ even if Y is a proper subset of X, provided X and Y are infinite. If σ and τ are ordinals, and $\tau < \sigma$, then $card(\tau) \leq card(\sigma)$. We can have $card(\tau) = card(\sigma)$; for example, $card(\omega) = card(\omega + 1) = \aleph_0$. We have $card(\sigma) \leq \sigma$ for any ordinal σ.

II.9.1.8. If $\{\kappa_i : i \in I\}$ is a set of cardinals, there is a smallest cardinal κ such that $\kappa_i \leq \kappa$ for all $i \in I$. This κ is denoted $\lim\{\kappa_i : i \in I\}$ (it is the same as $\lim\{\kappa_i : i \in I\}$ in the ordinal sense.)

II.9.1.9. THEOREM. If σ is any ordinal, then there is a cardinal κ with $\sigma < \kappa$.

This is easy to prove using the Axiom of Choice: we can take $\kappa = card(\mathcal{P}(\sigma))$. However, it can be proved without the Axiom of Choice. We first prove a lemma:

II.9.1.10. LEMMA. Let σ be an ordinal. Then the class $Z(\sigma) = \{\tau \in \Omega : |\tau| = |\sigma|\}$ is a set.

PROOF: Let s be a set with $|s| = |\sigma|$ (we could take $s = \sigma$.) If $\tau \in Z(\sigma)$, then there is a well-ordering on s making s order-isomorphic to τ; thus the collection $\mathcal{R}_\tau$ of all well orderings on s order-isomorphic to τ is nonempty. $\mathcal{R}_\tau$ is contained in the set of all relations on s, which can be identified with $\mathcal{P}(s \times s)$. If $\tau, \rho \in Z(\sigma)$, $\tau \neq \rho$, then $\mathcal{R}_\tau$ and $\mathcal{R}_\rho$ are disjoint and thus distinct. Thus $\tau \mapsto \mathcal{R}_\tau$ is an injective map from $Z(\sigma)$ to $\mathcal{P}(\mathcal{P}(s \times s))$, which is a set. 👍

II.9.1.11. Corollary. Let σ be an infinite ordinal, and $\kappa = \cup Z(\sigma)$. Then κ is a cardinal; $|\sigma| < \kappa$, and κ is the smallest cardinal greater than $card(\sigma)$. In fact, $\kappa = \{\tau \in \Omega : |\tau| \leq |\sigma|\}$.

Proof: The last formula is easy to verify, and the other statements follow immediately (note that $\kappa \notin \kappa$.) 👍

II.9.1.12. Corollary. The class of cardinals is not a set.

II.9.1.13. Definition. For each ordinal σ, we (inductively) define $\aleph_\sigma$ to be the smallest infinite cardinal (aleph) greater than $\{\aleph_\tau : \tau < \sigma\}$.

By II.9.1.11., $\aleph_\sigma$ is defined for every ordinal σ. This gives a one-to-one order-preserving correspondence $\sigma \longleftrightarrow \aleph_\sigma$ between the ordinals and the infinite cardinals (alephs).

II.9.1.14. Definition. The cardinal $\aleph_\sigma$ is a *successor cardinal* if σ is a successor ordinal, and a *limit cardinal* if σ is a limit ordinal.

By this definition, $\aleph_0$ and $\aleph_\omega$ are examples of limit cardinals, and $\aleph_1$ is a successor cardinal.

II.9.1.15. Since every cardinal is an ordinal, the function $\sigma \mapsto \aleph_\sigma$ can be regarded as a map from Ω to Ω. It is easy to prove by induction that $\sigma \leq \aleph_\sigma$ for every σ, and intuitively it seems that $\aleph_\sigma$ is *much* bigger than σ. Put another way: if σ is an ordinal, then it seems intuitively that $\{\kappa : \kappa \text{ a cardinal}, \kappa < \sigma\}$ should be quite a small subset of σ, i.e. the cardinals should be sparsely distributed among the ordinals. However, it is not hard to show (Exercise II.9.9.10.) that there is a σ for which $\sigma = \aleph_\sigma$. (Such a σ is obviously a cardinal.) We discuss related matters in ().

II.9.2. General Cardinals and Alephs

II.9.2.1. We have so far only defined cardinals of sets which can be well ordered; these cardinals, if infinite, are called *alephs*. In the absence of the Axiom of Choice, there will be sets which cannot be well ordered and whose cardinal is as yet undefined. As described in the introduction, we want to pick a "model set" to serve as the cardinal of such a set. This can only be done rigorously in the presence of the Axiom of Regularity, and even then in a not entirely satisfactory way without the Axiom of Choice. However, even without the Axiom of Regularity it is possible to use the whole cardinal terminology as simply shorthand for statements about sets which would otherwise be cumbersome and inelegant. For example, II.9.9.11.(a) can be phrased without the language of cardinals as follows:
"Let X be a set containing a countably infinite subset. If Y is any set disjoint from X such that $X \cup Y$ is equipotent with $\mathcal{P}(X)$, then Y is equipotent with $\mathcal{P}(X)$."

We will thus be slightly imprecise and assume that $card(X)$ has been defined for every set X, so that we can write such statements as "Let κ be a cardinal." The usual "ordering" gives a partial ordering on the class of all cardinals (the antisymmetric property follows from the Schröder-Bernstein Theorem.)

II.9.2.2. Infinite cardinals which are not alephs (if they exist) are not comparable with alephs in general. Although infinite cardinals, alephs are "small" in the sense that any infinite cardinal dominated by an aleph is again an aleph. However, there is no cardinal larger than all the alephs, since the class of alephs do not form a set (II.9.1.12..) Thus the statements in the previous section must be modified in the absence of the Axiom of Choice by interpreting "cardinal" to mean "cardinal of a well ordered set", i.e. a natural number or aleph. For example, $\aleph_\sigma$ is the smallest *aleph* greater than $\{\aleph_\tau : \tau < \sigma\}$: there could be another cardinal κ greater than all these alephs but incomparable with $\aleph_\sigma$. There could even be an infinite cardinal which is incomparable with $\aleph_0$! ()

II.9.2.3. A potential definition (proposed by RUSSELL) is to say that $card(X)$ is the class of all sets equipotent with X. The problem is that this is a proper class, not a set (unless $X = \emptyset$). A clever device (due to D. SCOTT) to modify this definition to actually give a set is to use the theory of levels (which needs the Replacement Axiom) to first let α be the smallest ordinal for which there is a set in V_α equipotent with X, and define $card(X)$ to be the set of all $y \in V_\alpha$ equipotent with X. This gives a well-defined set $card(X)$, and if Y is another set equipotent with X, then $card(Y) = card(X)$. If X can be well ordered, this is a different definition of $card(X)$ than the ordinal definition, but a straightforward translation can be given. This definition is slightly clumsy to work with but gives a fairly satisfactory solution.

II.9.2.4. One different way "cardinality" of a set could be defined is to associate to a set X the partially ordered set $S(X)$ of equipotence classes of subsets of X. This partially ordered set always has a largest element, and, if X is infinite, contains a hereditary subset order-isomorphic to ω. The set $S(X)$ is totally ordered if and only if S can be well ordered; thus the AC is equivalent to the statement that $S(X)$ is always totally ordered. In this case, $S(X)$ is well ordered; if X has cardinality $\aleph_\alpha$, then $S(X) \cong \omega + \alpha + 1$. However, in the absence of AC it is unclear whether $S(X) \cong S(Y)$ implies that X and Y are equipotent. It would be interesting to characterize which partially ordered sets occur as $S(X)$ for some X (in the absence of AC).

II.9.3. Cardinal Arithmetic

Cardinal arithmetic is easy to define. However, it must be kept in mind that it is *not* the same as ordinal arithmetic for infinite cardinals (alephs), and one must always carefully distinguish whether cardinal or ordinal arithmetic is being used. The distinction is usually clear from the context; but if there is any possibility of confusion we will use $+_c$ and $+_o$ for cardinal and ordinal addition respectively, and similarly $\cdot_c$ and $\cdot_o$. We can also use the vertical arrow notation for exponentiation to distinguish between cardinal and ordinal exponentiation (where the distinction is more crucial), writing $\sigma \uparrow_o \tau$ for σ^τ in the ordinal sense, and $\kappa \uparrow_c \lambda$ for cardinal exponentiation.

II.9.3.1. DEFINITION. Let κ and λ be cardinals, and let X and Y be disjoint sets (see Exercise II.5.6.15.) with $card(X) = \kappa$, $card(Y) = \lambda$. Then

$\kappa + \lambda = \kappa +_c \lambda = card(X \cup Y)$.

$\kappa \cdot \lambda = \kappa \cdot_c \lambda = card(X \times Y)$.

$\kappa^\lambda = \kappa \uparrow_c \lambda = card(X^Y)$ (X^Y denotes the set of functions from Y to X.)

It is clear that this definition does not depend on the choice of the sets X and Y.

The definition of cardinal exponentiation gives the explanation for the notation $card(\mathcal{P}(\mathbb{N})) = card(\{0,1\}^{\mathbb{N}}) = 2^{\aleph_0}$, since for set theorists $2 = \{0,1\}$.

Cardinal addition and multiplication agree with ordinary addition and multiplication for finite cardinals. Cardinal addition and multiplication are not very interesting for infinite cardinals, at least for alephs:

II.9.3.2. THEOREM. Let κ and λ be nonzero cardinals of well-ordered sets, not both finite. Then $\kappa + \lambda = \kappa \cdot \lambda = \max(\kappa, \lambda)$.

Note that II.5.4.4. and II.5.4.9.(i) are special cases.
PROOF: See Exercises II.9.9.2. and II.9.9.3.. 👍

It can be seen easily that $\kappa +_c \lambda = card(\kappa +_o \lambda)$ and $\kappa \cdot_c \lambda = card(\kappa \cdot_o \lambda)$.

II.9.3.3. Exponentiation of cardinals is very interesting, however, and is quite different from ordinal exponentiation (for example, $\omega^\omega = \omega \uparrow_o \omega$ is a countable ordinal, hence $card(\omega^\omega) = \aleph_0$, but $\aleph_0^{\aleph_0} = \aleph_0 \uparrow_c \aleph_0$ is an uncountable cardinal.)

Cardinal addition, multiplication, and exponentiation satisfy the usual algebraic rules (Exercise II.9.9.15..) However, subtraction and division are another matter: they are usually not well defined.

II.9.3.4. The sum of an infinite set of cardinals can also be defined. If $\{\kappa_i : i \in I\}$ is an indexed set of cardinals, and $\{X_i : i \in I\}$ is a collection of disjoint sets (see Exercise ()) with $card(X_i) = \kappa_i$ (e.g. $X_i = \{i\} \times \kappa_i$), then $\sum_i \kappa_i = card(\cup_i X_i)$. If $\kappa = \lim\{\kappa_i : i \in I\}$ is an aleph and $card(I) \leq \kappa$ (in particular, if the κ_i are all distinct), then $\sum_{i \in I} \kappa_i = \kappa$. Infinite products can be defined similarly.

II.9.4. The Role of the Axiom of Choice

In this subsection, we explore what can and cannot be proved about cardinals without assuming the Axiom of Choice (AC).

II.9.4.1. If $\aleph$ is an aleph, then $\aleph + \aleph = \aleph$ and $\aleph^2 = \aleph\aleph = \aleph$. It cannot be proved without AC that $\kappa + \kappa = \kappa$ or $\kappa^2 = \kappa$ for every infinite cardinal κ. In fact, the statement that $\kappa^2 = \kappa$ for every infinite cardinal κ implies (hence is equivalent to) AC [**?**, XVI.2 Theorem 6]. The statement that $\kappa + \kappa = \kappa$ for all infinite κ is strictly weaker than AC ([]; cf. Exercise II.9.9.4.). It can be proved without AC that some infinite cardinals not provably alephs, such as $\kappa = 2^{\aleph_0}$, satisfy $\kappa + \kappa = \kappa^2 = \kappa$; see Exercise (). [**?**, XVI.2] contains several other similar statements equivalent to AC (see Exercise ().)

If AC is not assumed, the class of cardinals is not totally ordered by $\leq$ (in fact, $\leq$ being a total order is equivalent to AC, cf. II.9.9.1..) In the absence of AC, there is another interesting and potentially different "ordering" on cardinals:

II.9.4.2. DEFINITION. Let X and Y be sets, $\kappa = |X|$, $\lambda = |Y|$. Write $\lambda \leq^* \kappa$ if there is a surjective function from X to Y, or if $Y = \emptyset$.

This relation clearly depends only on κ and λ and not on the choice of X or Y, and is transitive.

II.9.4.3. PROPOSITION. Let κ and λ be cardinals. If $\lambda \leq \kappa$, then $\lambda \leq^* \kappa$.

PROOF: Let $\kappa = |X|$, $\lambda = |Y|$. The result is obvious if $Y = \emptyset$. Otherwise, let $f : Y \to X$ be injective, and fix $y_0 \in Y$. Define $g : X \to Y$ as follows. If $x \in X$, set

$$g(x) = \begin{cases} y & \text{if} \quad f(y) = x \\ y_0 & \text{if} \quad x \notin f(Y) \end{cases} .$$

Then g is surjective. 👍

II.9.4.4. The Axiom of Choice easily implies that if $\lambda \leq^* \kappa$, then $\lambda \leq \kappa$, so $\leq$ and $\leq^*$ are identical (more generally, the relations are identical if κ and λ are alephs.) It is an open question whether agreement of $\leq$ and $\leq^*$ implies the AC. In the absence of AC, it cannot be proved that $\leq^*$ is antisymmetric (i.e. there is no analog of the Schröder-Bernstein Theorem; see Exercise ()(b).) It is provable without AC that $\kappa \leq^* \lambda$ implies $2^\kappa \leq 2^\lambda$ (Exercise ().)

II.9.4.5. PROPOSITION. Let $\kappa = card(I)$ and $\lambda = card(Y)$ be cardinals. Then $\kappa \leq^* \lambda$ if and only if Y can be written as $\cup_{i \in I} Y_i$, with the Y_i nonempty and disjoint.

PROOF: If $f : Y \to I$ is surjective, let $Y_i = f^{-1}(\{i\})$. Conversely, if $Y = \cup Y_i$, define $f : Y \to I$ by $f(y) = i$ if $y \in Y_i$. 👍

II.9.4.6. THEOREM. [HARTOGS] Let κ be an infinite cardinal, and X a set with $card(X) = \kappa$. Let $\aleph$ be the collection of all ordinals equipotent with a subset of X. Then $\aleph$ is a set, hence a cardinal, and is the smallest aleph $\aleph_\sigma$ satisfying $\aleph_\sigma \not\leq \kappa$.

PROOF: Let W be the collection of all well-orderings on subsets of X. Each well-ordering is a relation on X, hence an element of $\mathcal{P}(X \times X)$; thus W is a subset of $\mathcal{P}(\mathcal{P}(X \times X))$, hence a set. Define a function f from W to Ω by sending an element of W to the unique ordinal it is order-isomorphic to. By the Axiom of Replacement, $\aleph = f(W)$ is a set. If $\alpha \in \aleph$, then $\beta \in \aleph$ for all β with $card(\beta) \leq card(\alpha)$, and in particular for all β with $\beta < \alpha$. Thus $\aleph$ is an ordinal and a cardinal (aleph). It is obvious that $\aleph \not\leq \kappa$ (since $\aleph \notin \aleph$) and that if $\aleph_\sigma < \aleph$, then $\aleph_\sigma \in \aleph$ so $\aleph_\sigma \leq \kappa$. 👍

II.9.4.7. This aleph, called the *Hartogs cardinal* of κ, is denoted $\aleph^+(\kappa)$. (Many authors use the notation $\aleph(\kappa)$, but this is confusing since $\aleph^+(\aleph_\sigma) = \aleph_{\sigma+1}$, not $\aleph_\sigma$.) It is easily checked that, if κ is infinite,

$$\aleph^+(\kappa) = \{\sigma \in \Omega : card(\sigma) \leq \kappa\} .$$

II.9.4.8. THEOREM. Let κ be an infinite cardinal. Then $\aleph^+(\kappa) \leq^* 2^{2^\kappa}$, and hence $\aleph^+(\kappa) < 2^{\aleph^+(\kappa)} \leq 2^{2^{2^\kappa}}$.

For the proof, note that if X is a set with $card(X) = \kappa$, then each subset of $\mathcal{P}(X)$ (element of $\mathcal{P}(\mathcal{P}(X))$) has a natural partial order induced by set inclusion.

It cannot be proved without AC that $\aleph^+(\kappa) \leq 2^{2^\kappa}$ (but see Exercise ().)

II.9.4.9. LEMMA. For each ordinal $\sigma < \aleph^+(\kappa)$, there is a subset of $\mathcal{P}(X)$ (element of $\mathcal{P}(\mathcal{P}(X))$) which is well ordered of order type σ.

PROOF: Let f be a bijection from σ to a subset Y of X. For $\rho \leq \sigma$, let $Y_\rho = \{f(\tau) : \tau < \rho\}$. Then $\rho \mapsto Y_\rho$ is an order-embedding of σ into $\mathcal{P}(X)$. 👍

PROOF: We now prove Theorem II.9.4.8.. Construct a function $f : \mathcal{P}(\mathcal{P}(X)) \to \aleph^+(\kappa)$ as follows. If $S \in \mathcal{P}(\mathcal{P}(X))$ is well ordered of order type σ for some $\sigma < \aleph^+(\kappa)$, set $f(S) = \sigma$; otherwise set $f(S) = 0$. By the lemma, f is surjective. The last statement follows from () and (). 👍

II.9.4.10. If κ is finite, then set $\aleph^+(\kappa) = \aleph_0$. Without AC, it is conceivable that there could be infinite κ for which $\aleph^+(\kappa) = \aleph_0$, i.e. $\aleph_0 \not\leq \kappa$. These are precisely the infinite Dedekind-finite sets (II.5.2.1.), cf. (). It is even conceivable that there could be infinite κ for which $\aleph_0 \not\leq^* \kappa$. This possibility is discussed in the next section.

II.9.5. Dedekind-Finite Sets

This section, which is off the mainstream, explores what can be said about infinite Dedekind-finite sets. It turns out there is quite a bit we can say. But it is of limited interest since such sets are incompatible with even the Countable Axiom of Choice. So we will not assume the Countable AC, which means we have to be very careful in doing set operations we usually take for granted; but note that we can still choose one element out of any nonempty set (or n elements successively out of any set of cardinality at least n), and we can still define sequences recursively.

I am indebted to A. Karagila for much illuminating discussion on this topic.

II.9.5.1. Definition. A set X is *Dedekind-finite* if there is no bijection from X to a proper subset of X, i.e. every injective function from X to X is surjective. Otherwise X is *Dedekind-infinite*. A cardinal κ is Dedekind-finite if it is the cardinal of a Dedekind-finite set.

Whether a set X is Dedekind-finite or Dedekind-infinite depends only on the cardinal of X: if X and Y are equipotent and one is Dedekind-finite, so is the other. Any subset of a Dedekind-finite set is Dedekind-finite.

The next result is fundamental:

II.9.5.2. Proposition. Let X be a set. Then X is Dedekind-infinite if and only if there is an injective function from $\mathbb{N}$ to X (i.e. there is a sequence of distinct elements of X). A cardinal κ is Dedekind-infinite if and only if $\aleph_0 \leq \kappa$.

Proof: If (x_n) is a sequence of distinct elements of X, define $f : X \to X$ by $f(x_n) = x_{n+1}$ for each n and $f(x) = x$ if $x \neq x_n$ for all n. Then f is an injective function from X onto $X \setminus \{x_1\}$, so X is Dedekind-infinite. Conversely, suppose X is Dedekind-infinite, and let $f : X \to X$ be injective but not surjective. Choose $x_0 \in X \setminus f(X)$. Define a sequence (x_n) recursively by setting $x_n = f(x_{n-1})$ for $n \in \mathbb{N}$. It is easily checked that the x_n are distinct (using that $f_n = f \circ \cdots \circ f$ (n iterates) is injective for all n). 👌

II.9.5.3. Any finite set is Dedekind-finite. An infinite set has n-element subsets (in fact many) for any n. So if we assume the Countable AC, if X is an infinite set, we can choose an n-element subset A_n for each n; then by II.5.4.8. $\cup_n A_n$ is countably infinite, so $\aleph_0 \leq |X|$. Thus, under Countable AC, every Dedekind-finite set is finite.

The next proposition gives some alternate characterizations of Dedekind-infiniteness:

II.9.5.4. Proposition. Let X be a set, and $\kappa = card(X)$. The following are equivalent:

(i) X is Dedekind-infinite.

(ii) $\kappa = \kappa + \aleph_0$.

(iii) $\kappa = \kappa + 1$.

(iv) If $\kappa = \lambda + 1$, then $\lambda = \kappa$.

(v) If Y is any set obtained from X by adding or removing finitely many points, then $card(Y) = \kappa$.

(vi) There is a proper subset Y of X which is equipotent with X.

PROOF: (i) $\Rightarrow$ (ii), (i) $\Rightarrow$ (iii), (i) $\Rightarrow$ (iv), (i) $\Rightarrow$ (v) are all proved the same way: if Z is a countably infinite subset of X, and W is Z with a finite number of elements removed, or a finite or countable number of points added, then $|Z| = |W|$ and $|X| = |(X \setminus Z) \cup Z| = |(X \setminus Z) \cup W|$.
It is obvious that any of (ii), (iii), (iv), (v) imply (vi).
(vi) $\Rightarrow$ (i) is (). 👍

II.9.5.5. If we have one infinite Dedekind-finite cardinal, we have many. If X is an infinite Dedekind-finite set and $n \in \mathbb{N}$, let F_n be an n-element subset of X, and $\kappa_n = |X \setminus F_n|$ (note that we can choose F_n for any particular n, or for finitely many n, but we may not choose F_n simultaneously for all n, and in particular we cannot choose the F_n so that $F_n \subseteq F_{n+1}$ for all n). It is easily seen that the cardinality of $X \setminus F_n$ depends only on n and not on the choice of F_n, so κ_n is well defined for each n. We clearly have $\kappa_{n+1} \leq \kappa_n$ for each n, and we have $\kappa_{n+1} \neq \kappa_n$ since $X \setminus F_n$ is Dedekind-finite. Thus there is a strictly decreasing sequence of infinite Dedekind-finite cardinals smaller than $|X|$, which dramatically contradicts the Well-Ordering Principle (cf. II.9.5.21.). More generally, if Y is any proper subset of X, then $|Y| < |X|$. (Note that there is a bijection between the cardinals $\leq |X|$ and a subset of $\mathcal{P}(\mathcal{P}(X))$, so this collection of cardinals is a "set," although the collection of all Dedekind-finite cardinals is not clearly a set.)

Recall the definition of $\leq^*$ (II.9.4.2.).

II.9.5.6. PROPOSITION. Let X be an infinite subset of $\mathbb{R}$. Then $\aleph_0 \leq^* |X|$, i.e. X is the union of a sequence of pairwise disjoint nonempty subsets.

PROOF: [Sie65] We first define a function $f : \mathcal{P}(\mathbb{R}) \to \mathcal{P}(\mathbb{R})$ such that if A is any subset of $\mathbb{R}$ with more than one element, then $f(A)$ is a nonempty proper subset of A. Note that if A is such a set, then there is a rational number r such that $\{x \in A : x < r\}$ is a nonempty proper subset of A (choose r to lie between two elements of A). So let (r_n) be an enumeration of $\mathbb{Q}$ (no form of Choice is needed to do this), and for each $A \subseteq \mathbb{R}$ with more than one element, let n be the smallest natural number for which $f(A) := \{x \in A : x < r_n\}$ is a nonempty proper subset of A. The function f can be defined arbitrarily on $\emptyset$ and singletons.
We now define a function g from the set of infinite subsets of $\mathbb{R}$ to itself, such that $g(A)$ is a proper subset of A, for each infinite set $A \subseteq \mathbb{R}$. For such an A, $f(A)$ and $A \setminus f(A)$ are both nonempty, and at least one is infinite. Set $g(A) = f(A)$ if $f(A)$ is infinite, and $g(A) = A \setminus f(A)$ otherwise.
Now let X be an infinite subset of $\mathbb{R}$. Let $X_1 = X \setminus g(X)$, and inductively set $X_{n+1} = g^n(X) \setminus g^{n+1}(X)$ (where $g^n = g \circ g \circ \cdots \circ g$ n times). Set $X_\infty = \cap_n g^n(X)$ if nonempty. Then the X_n are nonempty, pairwise disjoint, and their union is X. 👍

This result is a theorem of ZF (i.e. uses no form of Choice) and applies, of course, also to any infinite set equipotent with a subset of $\mathbb{R}$. In particular, any infinite cardinal comparable to $2^{\aleph_0}$ is weakly Dedekind-infinite (II.9.5.9.).

The next result is due to KURATOWSKI and TARSKI. The proof of (ii) $\Rightarrow$ (i) is from [Sie65]. This is a theorem of ZF. Under Countable AC, the conditions are just equivalent to κ being infinite.

II.9.5.7. THEOREM. Let κ be a cardinal, and X a set with $|X| = \kappa$. The following are equivalent:

(i) $\aleph_0 \leq^* \kappa$, i.e. there is a sequence of pairwise disjoint nonempty subsets of X.

(ii) $\aleph_0 \leq 2^\kappa$, i.e. X has a sequence of distinct subsets.

(iii) $2^{\aleph_0} \leq 2^\kappa$.

PROOF: (i) $\Rightarrow$ (iii): Let (A_n) be a sequence of pairwise disjoint nonempty subsets of X. For each $S \subseteq \mathbb{N}$, let $B_S = \cup_{n \in S} A_n$. Then the B_S are pairwise distinct, i.e. $S \mapsto B_S$ is an injective function from $\mathcal{P}(\mathbb{N})$ to $\mathcal{P}(X)$.
(ii) $\Rightarrow$ (i): Let (A_n) be a sequence of distinct subsets of X. Let S be the set of sequences of 0's and 1's. For $s = (s_n) \in S$, define

$$C_s = \{x \in X : \forall n \; x \in A_n \text{ if } s_n = 1 \text{ and } x \notin A_n \text{ if } s_n = 0\} \; .$$

Then the C_s are pairwise disjoint subsets of X. For each n, we have

$$A_n = \bigcup_{s_n = 1} C_s$$

so infinitely many distinct sets are unions of the C_s; thus infinitely many C_s are nonempty. Set

$$E = \{s \in S : C_s \neq \emptyset\} \; .$$

Then E is an infinite subset of S. Since S is equipotent with $\mathbb{R}$, E is equipotent with a subset of $\mathbb{R}$. So by II.9.5.6. there is a surjective function from E to $\mathbb{N}$. Composing this with the function from X to E sending $x \in X$ to the C_s with $x \in C_s$, we get a surjective function from X to $\mathbb{N}$.
(iii) $\Rightarrow$ (ii) is trivial (given II.5.3.3.). 👍

II.9.5.8. By a slight modification of the (i) $\Rightarrow$ (iii) argument, if κ and λ are cardinals and $\lambda \leq^* \kappa$, then $2^\lambda \leq 2^\kappa$.

II.9.5.9. DEFINITION. Let X be a set. Then

(i) X is *dually Dedekind-infinite* if there is a surjective function from X to X which is not injective.

(ii) X is *weakly Dedekind-infinite* if $\aleph_0 \leq^* |X|$, i.e. X satisfies the equivalent conditions of II.9.5.7..

X is *dually Dedekind-finite* [resp. *weakly Dedekind-finite* or *power Dedekind-finite*] if it is not dually Dedekind-infinite [resp. weakly Dedekind-infinite].

II.9.5.10. We have (X Dedekind-infinite) $\Rightarrow$ (X dually Dedekind-infinite) $\Rightarrow$ (X weakly Dedekind-infinite) $\Rightarrow$ (X infinite). Actually, the term *weakly Dedekind-finite* is misleading language, since the condition of being weakly Dedekind-finite is considerably *stronger* than being Dedekind-finite (II.9.5.17.); in fact, by II.9.5.7. X is weakly Dedekind-finite if and only if $\mathcal{P}(X)$ is Dedekind-finite, so the alternate term *power Dedekind-finite* seems more appropriate. These conditions can be regarded as degrees of Dedekind-finiteness. ✋

By II.9.5.6., every infinite subset of $\mathbb{R}$ is weakly Dedekind-infinite. Subsets of $\mathbb{R}$ are special among all sets (in the absence of AC) in part because they can be totally ordered.

Amorphous Sets

II.9.5.11. Definition. An infinite set X is *amorphous* if X cannot be written as a disjoint union of two infinite subsets.

II.9.5.12. Any infinite subset of an amorphous set is amorphous. If X is an amorphous set, Y an infinite set, and there is a surjective function $f : X \to Y$, then Y is also amorphous [if Y is the disjoint union of Y_1 and Y_2, then X is the disjoint union of $f^{-1}(Y_1)$ and $f^{-1}(Y_2)$]. In particular, if X is weakly Dedekind-infinite, since $\mathbb{N}$ is not amorphous X cannot be amorphous, so if X is amorphous then X is power Dedekind-finite. Thus amorphous sets are Dedekind-finite in a strong sense. In particular, amorphous sets are incompatible with the Countable AC.

As an example of how strange an amorphous set is, we have:

II.9.5.13. Proposition. An amorphous set cannot be totally ordered.

Proof: Let X be an infinite totally ordered set. If there is an element of X with infinitely many predecessors and infinitely many successors, then X can be split as the disjoint union of two infinite sets and is not amorphous. If every element of X has only finitely many predecessors, then X is well ordered and is thus equipotent with $\mathbb{N}$. Similarly, if every element has only finitely many successors, X is reverse well ordered and again equipotent with $\mathbb{N}$. Otherwise, the sets $A = \{x \in X : x \text{ has infinitely many predecessors}\}$ and $B = \{x \in X : x \text{ has infinitely many successors}\}$ are nonempty and disjoint, and their union is X. If A is finite, it has a smallest element, and its set of predecessors is order-isomorphic to $\mathbb{N}$. Similarly, if B is finite, X has a subset equipotent with $\mathbb{N}$. In any case, X is not amorphous. 👍

Examples and Constructions

The discussion so far could conceivably be vacuous since we do not yet know whether there can be infinite Dedekind-finite sets. We now discuss existence of various kinds of Dedekind-finite sets, although the actual constructions are mostly beyond the scope of this book.

II.9.5.14. The first example was due to A. Fraenkel [Fra22], who, assuming a certain extension ZFA of ZF is consistent (ZFA was later proved to be equiconsistent with ZF) constructed a model of ZF containing an amorphous set X. This X is not only Dedekind-finite, but even power Dedekind-finite (II.9.5.12.).

II.9.5.15. The most famous Dedekind-finite set was constructed by P. Cohen in his work on the independence of the Axiom of Choice and the Continuum Hypothesis [Coh66]: a canonical infinite Dedekind-finite subset X of $\mathbb{R}$ in a certain model of ZF. This X is weakly Dedekind-infinite (II.9.5.6.) and of course can be totally ordered, hence has quite different properties from Fraenkel's example (although Cohen's construction was modeled on Fraenkel's); the cardinalities of these sets are even incomparable, in fact totally incomparable: the only subsets of either which are equipotent with a subset of the other are the finite subsets.

II.9.5.16. If X is a set (infinite), there are two closely related constructions which are relevant. We denote by SX the set of finite sequences in X with no repeated terms (including the "empty sequence"), and $\mathcal{F}(X)$ (set-theorists' notation $[X]^{<\omega}$) the set of finite subsets of X. There is an obvious surjective function from SX to $\mathcal{F}(X)$ with finite preimages of singletons.

II.9.5.17. PROPOSITION. (i) If X is Dedekind-finite, then SX is also Dedekind-finite.
(ii) If X is infinite, then SX is dually Dedekind-infinite, hence weakly Dedekind-infinite.

PROOF: (i): If SX is Dedekind-infinite, let (s_n) be a sequence of distinct elements of SX. Concatenate the s_n into a sequence s. Since each s_n has no repeated elements, there must be a subsequence of s consisting of distinct elements of X, so X is Dedekind-infinite.
(ii): Define $f : SX \to SX$ by $f(s) = s$ with its last term deleted (with $f(\emptyset) = \emptyset$). Then f is surjective but not injective. 👍

II.9.5.18. So if X is any infinite Dedekind-finite set (e.g. FRAENKEL'S amorphous set), then SX is Dedekind-finite but dually Dedekind-infinite. Since SX has a subset equipotent with X (the sequences of length 1), if X cannot be totally ordered (e.g. if X is amorphous) neither can SX. Note that SX is never amorphous (for X infinite).

II.9.5.19. The situation with $\mathcal{F}(X)$ is quite different: if X is infinite and Dedekind-finite, then $\mathcal{F}(X)$ can be either Dedekind-finite or Dedekind-infinite, and it is somewhat subtle to characterize which holds. [KS20] has a thorough study of conditions equivalent to Dedekind-finiteness of $\mathcal{F}(X)$. For example, if X is a Dedekind-finite subset of $\mathbb{R}$ (e.g. COHEN'S generic subset), then $\mathcal{F}(X)$ is Dedekind-finite. But if X is a Dedekind-finite set containing a set of "Russell's socks" (a sequence of pairwise disjoint two-element sets), $\mathcal{F}(X)$ is Dedekind-infinite (there do exist such sets in some models of ZF). A set X is called *Cohen finite* [resp. *Cohen infinite*] if $\mathcal{F}(X)$ is Dedekind-finite [resp. Dedekind-infinite].

Other set-theoretic constructions preserve Dedekind-finiteness:

II.9.5.20. PROPOSITION. A finite Cartesian product of Dedekind-finite sets is Dedekind-finite. A finite union of Dedekind-finite sets is Dedekind-finite.

PROOF: If (x_j) is a sequence of distinct elements of $X_1 \times \cdots \times X_n$, then for some k, $1 \leq k \leq n$, the sequence of k'th coordinates of the x_j must contain infinitely many distinct elements of X_k, hence X_k is not Dedekind-finite. Similarly, if $X = \cup_{k=1}^{n} X_k$, and (x_j) is a sequence of distinct elements of X, then infinitely many of the x_j must be in some X_k, so X_k is not Dedekind-finite. 👍

In particular, if X is Dedekind-finite, then X^n and $X \times F$ are Dedekind-finite for any n and any finite set F. These sets all have different cardinalities by II.9.5.5.. By repeating these constructions and taking subsets, a very large collection of Dedekind-finite cardinals can be obtained. In this connection, note the following (due to TARSKI):

II.9.5.21. PROPOSITION. If there exists a Dedekind-finite cardinal, there exists a collection of Dedekind-finite cardinals order-isomorphic to $\mathbb{R}$.

PROOF: Let Y be a Dedekind-finite set, and $X = SY$. Then X is Dedekind-finite, but is the disjoint union of a countably infinite collection of nonempty subsets. Index these subsets by $\mathbb{Q}$, so X is the disjoint union of $\{A_q : q \in \mathbb{Q}\}$, with each A_q nonempty. For each $r \in \mathbb{R}$, set

$$X_r = \bigcup_{q<r} A_q \ .$$

Then $X_r \subseteq X$, and if $r < s$ X_r is a proper subset of X_s, so as in II.9.5.5. we have $|X_r| < |X_s|$ for $r < s$. 👍

Russell Sets and Russell Cardinals

II.9.5.22. DEFINITION. A set X is a *Russell set* if it is Dedekind-finite but can be written as a countable union of pairwise disjoint two-element sets (called *pairs of Russell socks*). A cardinal κ is a *Russell cardinal* if it is the cardinal of a Russell set.

II.9.5.23. There are models of ZF containing Russell sets. For a thorough and entertaining description of Russell sets, including motivation for the sock terminology, see [HT06]; some results described there graphically show that the world of Dedekind-finite cardinals (when they exist at all) is even more bizarre than we have indicated (for example, there is a family of $2^{\aleph_0}$ pairwise incomparable Dedekind-finite cardinals).

Rigid Sets

Recall that a *permutation* of a set X is a bijection from X to X.

II.9.5.24. DEFINITION. An infinite set X is *rigid* if every permutation of X moves only finitely many elements (i.e. is the identity on the complement of a finite subset).

II.9.5.25. Any set equipotent with a rigid set is rigid. A subset of a rigid set is rigid (since a permutation of a subset extends to a permutation of the whole set). Since $\mathbb{N}$ is not rigid (it has many permutations moving infinitely many elements, even all elements), a rigid set must be Dedekind-finite. But "most" Dedekind-finite sets are not rigid (e.g. if X is any infinite set, then $X \times Y$ is not rigid for any Y with more than one element).

II.9.5.26. Being rigid seems to be similar to being amorphous. But it turns out that the notions are distinct. Say an infinite set X is *strongly rigid* if in any partition of X into nonempty subsets, all but finitely many cells are singletons, and *strongly amorphous* if it is amorphous and strongly rigid. A strongly rigid set is rigid. It can be shown that FRAENKEL'S set is strongly amorphous, hence rigid, and that COHEN'S generic subset of $\mathbb{R}$ is rigid, but it is not amorphous. There are amorphous sets containing infinitely many pairwise disjoint two-element sets, and such a set is not rigid (just interchange each pair of socks). And there are rigid sets which can be written as a countable disjoint union of three-element sets in some models of ZF.

II.9.6. Exercises

II.9.6.1. Let X be a set.

(a) Show that $|X| \leq |\mathcal{F}(X)| \leq |\mathcal{P}(X)|$ and that both equalities cannot hold (for fixed X).

(b) If X is finite, then $\mathcal{F}(X) = \mathcal{P}(X)$, so in particular $|\mathcal{F}(X)| = |\mathcal{P}(X)|$.

(c) If the AC is assumed, show that if X is infinite, then $|X| = |\mathcal{F}(X)|$ ().

(d) Show without any form of choice that if X is countably infinite, so is $\mathcal{F}(X)$, so $\aleph_0 = |X| = |\mathcal{F}(X)|$. [Take $X = \mathbb{N}$, and define a "lexicographic" order on finite subsets of $\mathbb{N}$.]

II.9.6.2. Let X and Y be nonempty sets.

(a) $X \times Y$ is Dedekind-finite if and only if both X and Y are Dedekind-finite (II.9.5.20.).

(b) $X \times Y$ is Cohen-finite if and only if both X and Y are Cohen-finite. [If (F_n) is a sequence of finite subsets of $X \times Y$, consider their projections into X and Y.]

(c) $X \times Y$ is power Dedekind-finite if and only if both X and Y are power Dedekind-finite. [If F is a surjection from $X \times Y$ to $\mathbb{N}$, consider the images of vertical cross-sections, and use II.9.6.1.(d).] Thus if X and Y are power Dedekind-finite, there is no sequence of distinct relations from X to Y.

(d) Let Y^X denote the set of all functions from X to Y. Show that (if $|Y| \geq 2$) $|2^X| \leq |Y^X| \leq |2^{X \times Y}|$. In particular, X^X is Dedekind-finite (i.e. there is no sequence of distinct functions from X to X) if and only if X is power Dedekind-finite.

(e) Show that if X is power Dedekind-finite, then X is dually Dedekind-finite. [If $f : X \to X$ is surjective but not injective, then (f^n) is a sequence of distinct functions from X to X, contradicting (d).]

II.9.7. The Continuum Hypothesis

If X is any set, and $card(X) = \kappa$, then $card(\mathcal{P}(X)) = card(\{0,1\}^X) = 2^\kappa$. By Cantor's Theorem (), $\kappa < 2^\kappa$. In aleph notation, $\aleph_{\sigma+1} \leq 2^{\aleph_\sigma}$ for every σ. (Without AC, it can only be proved that $\aleph_{\sigma+1} \leq^* 2^{\aleph_\sigma}$ (Exercise ().) In particular, $\aleph_1 \leq 2^{\aleph_0}$. Do we have equality? This is equivalent (under AC) to the question whether every uncountable subset of $\mathbb{R}$ has cardinality $2^{\aleph_0}$.

II.9.7.1. Continuum Hypothesis (CH): $2^{\aleph_0} = \aleph_1$.

II.9.7.2. It was a big mathematical industry in the first part of the 20th century (actually beginning with Cantor in the 19th century) to try to prove (or, sometimes, to disprove) this hypothesis. Perhaps surprisingly, it turns out to be independent of ZFC, i.e. it cannot be either proved or disproved in ZFC![16] Mathematicians to this day have had various opinions about whether the CH "should be" true, and it is much more customary than with AC to carefully note any use of the CH in proofs and to avoid using it whenever possible.

The Axiom of Choice is also independent of ZF+CH. However, the CH implies a weak version of AC: $\mathbb{R}$ can be well ordered since it is equipotent with ω_1. This in turn implies that there is a choice function for $\mathcal{P}(\mathbb{R})$ (). But the CH does not even imply the Countable Axiom of Choice.

If the AC is not assumed, then there is a weaker version of the Continuum Hypothesis which is not equivalent to the one above:

II.9.7.3. Weak Continuum Hypothesis: There is no cardinal strictly between $\aleph_0$ and $2^{\aleph_0}$, i.e. if $X \subseteq \mathcal{P}(\mathbb{N})$ (or $X \subseteq \mathbb{R}$), then either X is countable or $|X| = |\mathcal{P}(\mathbb{N})| = |\mathbb{R}|$.

In fact there are many inequivalent ways to phrase the CH in the absence of the AC. See http://math.stackexchange.com/questions/404807/how-to-formulate-continuum-hypothesis-without-the-axiom-of-choice?rq=1 for a discussion.

There is a generalized version of the Continuum Hypothesis:

II.9.7.4. Generalized Continuum Hypothesis (Version 1): $2^{\aleph_\sigma} = \aleph_{\sigma+1}$ for every σ.

This is also independent of ZFC (?). However, without the Axiom of Choice there is another version analogous to the Weak Continuum Hypothesis:

II.9.7.5. Generalized Continuum Hypothesis (Version 2): If κ is any infinite cardinal, then there is no cardinal strictly between κ and 2^κ.

II.9.7.6. Note that Version 2 is not really weaker than Version 1, since it is an assertion about all sets, not just those which can be well ordered. In fact, it turns out that Version 2 implies the Axiom of Choice, and thus also Version 1 (Exercise ().)

The Continuum Hypothesis does not imply the Generalized Continuum Hypothesis Version 1. There are intermediate hypotheses, saying $2^{\aleph_\sigma} = \aleph_{\sigma+1}$ for every σ up to some point but not beyond.

[16] This statement is again under the assumption that ZF is consistent.

II.9.8. Large Cardinals

We will give only the briefest introduction to the bizarre world of large cardinals, emphasizing the measurable cardinals (as is appropriate in a book on real analysis.) In this section, "cardinal" means "cardinal of a well ordered set"; we may as well assume the Axiom of Choice.

II.9.8.1. DEFINITION. Let κ be a cardinal. Then κ is *regular* if κ is not equal to $\sum_{i\in I}\kappa_i$ for any set $\{\kappa_i : i \in I\}$ with $card(I) < \kappa$ and $\kappa_i < \kappa$. Otherwise κ is *singular*.

Equivalently, κ is regular if a set of cardinality κ cannot be written as a union of fewer than κ subsets, each of cardinality less than κ. Among the finite cardinals, only 0,1, and 2 are regular. Among infinite cardinals, every successor cardinal is regular; $\aleph_0$ is regular, but "most" limit cardinals are singular, for example $\aleph_\omega = \sum_{n\in\mathbb{N}}\aleph_n$.

II.9.8.2. DEFINITION. An infinite cardinal κ is *weakly inaccessible* if it is an uncountable regular limit cardinal.

II.9.8.3. A weakly inaccessible cardinal cannot be reached from below by taking successor cardinals or sums of smaller sets of smaller cardinals. The set of cardinals less than a weakly inaccessible cardinal κ has cardinality κ, the same as the set of ordinals less than κ (i.e. $\kappa = \aleph_\kappa$), so the cardinals less than κ are in this sense not so sparsely distributed among the ordinals. (But $\kappa = \aleph_\kappa$ does not imply that κ is weakly inaccessible (Exercise ()).)

II.9.8.4. There is one potential way to reach a weakly inaccessible cardinal: in the absence of the Continuum Hypothesis, even $2^{\aleph_0}$ might be so huge as to be weakly inaccessible, or exceed a weakly inaccessible cardinal. In fact, it appears to be consistent with ZFC to assume this. So we strengthen the definition of inaccessibility:

II.9.8.5. DEFINITION. A cardinal κ is *inaccessible*, or *strongly inaccessible*, if it is weakly inaccessible and $\lambda < \kappa$ implies $2^\lambda < \kappa$.

If the Generalized Continuum Hypothesis is assumed, there is no difference between weak and strong inacessibility.

II.9.8.6. The question then becomes: do (weakly) inaccessible cardinals exist?

Inaccessible cardinals cannot be reached by any of the operations of set theory, and thus their existence cannot be proved in ZFC. For if there does exist an inaccessible cardinal in some model of ZFC, let κ be the smallest inaccessible cardinal. Then the set V_κ of Exercise () can be taken to be the universe of a model for ZFC in which there are no inaccessible cardinals. On the other hand, it appears to be consistent with ZFC to assume that there exists an inaccessible cardinal.

II.9.8.7. Note that the cardinals immediately larger than an inaccessible cardinal are not inaccessible. In fact, if there is a second inaccessible cardinal, it must be vastly larger than the first one. The same proof as in the previous paragraph shows that even if an inaccessible cardinal is assumed to exist, it cannot be proved in ZFC that a second one also exists. Thus the inaccessible cardinals are (seemingly) very sparsely distributed even among all cardinals, let alone among all ordinals.

II.9.8.8. Inaccessible cardinals are often called *large cardinals*. (Set theorists make a technical distinction between inaccessible cardinals and "large cardinals", which have no precise widely-accepted definition, but inaccessible cardinals are large cardinals by any reasonable definition, and the classes of large cardinals which have been studied are inaccessible cardinals.) There is an enormous hierarchy of types of large cardinals which have been considered. For an analyst, the most interesting class is the class of measurable cardinals, discussed in the next section. See [Kan03] for a full discussion of all types of large cardinals, or [Ruc82] for a readable semitechnical description of some of them.

Large cardinals may seem far removed from the "real world", but it turns out that existence axioms for certain kinds of large cardinals have close connections with natural questions about subsets of $\mathbb{R}$ and even $\mathbb{N}$. See () for a discussion of a few of these; [Kan03] has a full treatment.

II.9.9. Exercises

II.9.9.1. (a) Show that the AC (Well-Ordering Principle) implies that $\leq$ is a total order (in fact, well order) on the class of all cardinals.

(b) Conversely, suppose $\leq$ is a total order on the class of all cardinals. If κ is a cardinal, show that $\kappa \leq \aleph^+(\kappa)$ (II.9.4.7.), and hence κ is an aleph. Conclude that AC holds.

Thus the AC is equivalent to the statement that the class of cardinals is totally ordered.

II.9.9.2. (a) If κ is an infinite cardinal (aleph), then $card(\kappa_{ev}) = card(\kappa_{odd}) = \kappa$ (II.8.9.8.). Thus $\kappa +_c \kappa = \kappa$.

(b) Conclude that if κ and λ are infinite cardinals (alephs), then $\kappa +_c \lambda = \max(\kappa, \lambda)$.

(c) Show that if κ is an infinite cardinal (aleph), then $\kappa +_c n = \kappa$ for each finite cardinal n ().

II.9.9.3. Let κ be an infinite cardinal (aleph). Define an ordering on $\kappa \times \kappa$ by setting $(\pi, \rho) < (\sigma, \tau)$ if $\max(\pi, \rho) < \max(\sigma, \tau)$, or if $\max(\pi, \rho) = \max(\sigma, \tau)$ and $\pi < \sigma$, or if $\max(\pi, \rho) < \max(\sigma, \tau)$, $\pi = \sigma$, and $\rho < \tau$.

(a) Show that this is a well ordering on $\kappa \times \kappa$. Draw a picture illustrating the ordering.

(b) Assume that $\lambda \cdot \lambda < \kappa$ for all cardinals $\lambda < \kappa$. Show that every element of $\kappa \times \kappa$ has fewer than κ predecessors under this ordering, so if σ is the order type of the well ordering, then $\sigma \leq \kappa$. Conclude that $card(\kappa \cdot \kappa) \leq \kappa$. (The opposite inequality also obviously holds.)

(c) Prove by transfinite induction on κ that $\kappa \cdot \kappa = \kappa$ if κ is an aleph.

(d) Conclude that if $\lambda \neq 0$ is finite or an aleph, then $\kappa \cdot \lambda = \max(\kappa, \lambda)$.

(e) If λ is a cardinal with $2 \leq \lambda \leq 2^\kappa$, then $\lambda^\kappa = 2^\kappa$ and, if $\lambda \leq \kappa$, $2^\lambda \cdot 2^\kappa = 2^\kappa$. [$2^\kappa \leq \lambda^\kappa \leq (2^\lambda)^\kappa \leq (2^\kappa)^\kappa = 2^{\kappa \cdot \kappa}$ and $2^\lambda \cdot 2^\kappa \leq 2^\kappa \cdot 2^\kappa = 4^\kappa$ (cf. II.9.9.15.).]

II.9.9.4. (a) Let κ be a cardinal. If $\kappa + \kappa = \kappa$, show that $\aleph_0 \cdot \kappa = \kappa$. [Let X be a set with $|X| = \kappa$ and f and g one-to-one functions from X to X with disjoint ranges. Consider the functions

$$\underbrace{g \circ g \circ \cdots \circ g}_{n} \circ f \ .]$$

(b) Show that AC implies that $\kappa + \kappa = \kappa$ for every infinite cardinal κ. The converse implication does not hold [].

(c) Assume $\kappa + \kappa = \kappa$ for all $\kappa \geq \aleph_0$. Show (without AC) that every infinite cardinal is Dedekind-infinite, i.e. $\geq \aleph_0$ [**?**].

II.9.9.5. Suppose X is a set of cardinality χ, and there is a purely nonatomic (countably additive) $\{0,1\}$-valued measure μ on $(X, \mathcal{P}(X))$ with $\mu(X) = 1$.

(a) Suppose, for some $\kappa < \chi$, there are sets $\{A_i : i \in I\}$ with each $A_i \subseteq X$ and $card(I) = \kappa$, such that $\mu(A_i) = 0$ for all i but $\mu(\cup_{i\in I} A_i) = 1$, and suppose κ is the smallest cardinal for which this happens. If $E \subseteq I$, define $\nu(E) = \mu(\cup_{i\in E} A_i)$. Then ν is a $\{0,1\}$-valued measure on I with $\nu(\{i\}) = 0$ for all $i \in I$ and $\nu(I) = 1$. Also, ν is $\underline{\kappa}$-additive by minimality of κ, so κ is measurable.

(b) If no $\kappa < \chi$ is measurable, conclude that μ is $\underline{\chi}$-additive, so χ is measurable.

(c) If χ is the first measurable cardinal, and κ is a cardinal, use the argument of (a) to show that every $\{0,1\}$-valued measure (on $(Y, \mathcal{P}(Y))$ for some set Y) is κ-additive if and only if $\kappa < \chi$.

II.9.9.6. [**?**] Suppose X is a set, and ν is a purely nonatomic probability measure on $(X, \mathcal{P}(X))$. Show that there is a purely nonatomic $\{0,1\}$-valued measure μ on $(X, \mathcal{P}(X))$ with $\mu(X) = 1$. Conclude that $card(X)$ dominates a measurable cardinal. (This requires that $2^{\aleph_0}$ does not dominate an inaccessible cardinal.)

II.9.9.7. Let κ and λ be cardinals.

(a) Show (without AC) that if $\kappa \leq \lambda$, then $\kappa \leq^* \lambda$. [Mimic the proof of II.5.4.1..]

(b) Show (without AC) that if $\kappa \leq^* \lambda$, then $2^\kappa \leq 2^\lambda$. [If $card(X) = \kappa$ and $card(Y) = \lambda$, and $f : Y \to X$, define $f^* : \mathcal{P}(X) \to \mathcal{P}(Y)$ by $f^*(A) = f^{-1}(A)$. If f is surjective, then f^* is injective.]

(c) Use (b) to show that if κ is any cardinal, then $2^\kappa \not\leq^* \kappa$. This also follows directly from the proof of II.5.3.3..

II.9.9.8. (a) Show (without AC) that if κ is any infinite cardinal, then $\aleph_0 \leq^* 2^\kappa$. [If $card(X) = \kappa$ and $n \in \mathbb{N}$, the set of subsets of X of cardinality n is nonempty. Define $f : \mathcal{P}(X) \to \{0\} \cup \mathbb{N} \cup \{\infty\}$ by $f(A) = card(A)$ if $A \subseteq X$ is finite, and $f(A) = \infty$ otherwise.]

(b) Conclude from II.9.9.7.(b) that $2^{\aleph_0} \leq 2^{2^\kappa}$. Explicitly describe a subset of $\mathcal{P}(\mathcal{P}(X))$ of cardinality $2^{\aleph_0}$. [For $S \subseteq \mathbb{N}$, consider $\{A \subseteq X : card(A) \in S\}$.]

II.9.9.9. (a) [**?**] Show (without AC or CH) that $\aleph_1 \leq^* 2^{\aleph_0}$. [Define $f : \mathcal{P}(\mathbb{Q}) \to \aleph_1 = \omega_1$ as follows. If $A \subseteq \mathbb{Q}$, set $f(A) = \sigma$ if A is well ordered (in the induced order from $\mathbb{Q}$) and order-isomorphic to σ; set $f(A) = 0$ if A is not well ordered.] Thus, from II.9.9.7.(b), $\aleph_1 < 2^{\aleph_1} \leq 2^{2^{\aleph_0}}$.

(b) [**?**] If $\aleph_1$ and $2^{\aleph_0}$ are not comparable (i.e. if $\aleph_1 \not\leq 2^{\aleph_0}$), then there is a cardinal κ satisfying $\kappa > 2^{\aleph_0}$ but $\kappa \leq^* 2^{\aleph_0}$. [Set $\kappa = \aleph_1 + 2^{\aleph_0}$. Let $f : [0,1] \to \aleph_1$ be surjective, and $g : \mathbb{R} \setminus [0,1] \to \mathcal{P}(\mathbb{N})$ be a bijection. Then f, g give a surjective map from $\mathbb{R}$ to $\aleph_1 \cup \mathcal{P}(\mathbb{N})$.] Then $2^\kappa = 2^{2^{\aleph_0}}$.

It cannot be proved in ZF that $\aleph_1 \leq 2^{\aleph_0}$ [].

II.9.9.10. (a) Set $\kappa_0 = \aleph_0$, $\kappa_{n+1} = \aleph_{\kappa_n}$ for each n, and $\kappa = \lim\{\kappa_n\}$. Show that $\kappa = \aleph_\kappa$. [We have that $\kappa = \aleph_\sigma$ for some σ. Since $\tau \leq \aleph_\tau$ for all τ by (), $\sigma \leq \kappa$. If $\sigma < \kappa$, then $\sigma < \kappa_n$ for some n, so $\aleph_\sigma < \aleph_{\kappa_n} = \kappa_{n+1} \leq \kappa$.] But κ is not inaccessible.

(b) Let κ be an inaccessible cardinal, and let $C = \{\lambda < \kappa : \lambda = \aleph_\lambda\}$. Show that $card(C) = \kappa$.

II.9.9.11. (a) Show (without AC) that if κ is a cardinal with $\aleph_0 \leq \kappa$, and λ is a cardinal with $\kappa + \lambda = 2^\kappa$, then $\lambda = 2^\kappa$. [This is a hard problem; see [Sie65, IX.5] for a solution.]

(b) Show (without AC) that if κ is a cardinal with $\aleph_0 \leq \kappa$, then $2^\kappa + 2^\kappa = 2^\kappa$. [It follows from $\aleph_0 \leq \kappa$ that $\kappa = 1 +_c \kappa$, so $2^\kappa = 2^{1+\kappa} = 2 \cdot 2^\kappa$.]

(c) Using (a), show (without AC) that if the GCH2 () is assumed, then every cardinal is dominated by an aleph, hence is an aleph, yielding the Axiom of Choice. [If κ_0 is an infinite cardinal, set $\kappa = \kappa_0 + \aleph_0$, and $\lambda = \kappa + \aleph^+(\kappa)$ (). Then $2^{2^\kappa} + \lambda \leq 2^{2^{2^\kappa}}$ by () and (b). Conclude from (a) and the GCH2 that either $\lambda = 2^{2^{2^\kappa}}$, in which case by (a), $\aleph^+(\kappa) = 2^{2^{2^\kappa}}$; or else $\lambda \leq 2^{2^\kappa}$. If $\lambda \leq 2^{2^\kappa}$, then $2^\kappa + \lambda \leq 2^{2^\kappa}$, so by the same argument either $\aleph^+(\kappa) = 2^{2^\kappa}$ or $\lambda \leq 2^\kappa$; if the latter, apply (a) and the GCH2 again to conclude that $2^\kappa = \aleph^+(\kappa)$.]

II.9.9.12. Show (without AC) that for any infinite cardinal κ, we have $\aleph^+(\kappa) \leq 2^{2^{\kappa^2}}$.[Mimic the proof of II.9.1.10..]

II.9.9.13. [Sie65, VIII.2 Ex. 2] Show (without AC) that if κ is infinite and $\kappa \leq 2^{\aleph_0}$, then $\aleph_0 \leq^* \kappa$ (and hence $2^{\aleph_0} \leq 2^\kappa$ by ()), as follows. (This appears to be a curious result: it apparently says that if κ is infinite and "not too large", then κ is "not too small".)

(a) Fix an enumeration $(r_1, r_2, \dots)$ of $\mathbb{Q}$. If X is an infinite subset of $\mathbb{R}$, let r_n be the first rational number r such that there are $a, b \in X$ with $a \leq r < b$. Then, if $Y = X \cap \{x \in \mathbb{R} : x \leq r_n\}$ and $Z = X \setminus Y$, Y and Z are disjoint and nonempty, and $X = Y \cup Z$. At least one of Y, Z is infinite.

(b) Use the procedure in (a) inductively, beginning with an infinite subset X of $\mathbb{R}$: write $X = X_1 \cup Y_1$ with X_1 nonempty, Y_1 infinite, and X_1, Y_1 disjoint. Similarly decompose $Y_1 = X_2 \cup Y_2$, $\dots$, $Y_n = X_{n+1} \cup Y_{n+1}$.

(c) Set $X_\infty = \cap_n Y_n = X \setminus \cup_n X_n$. If $X_\infty \neq \emptyset$, then X is the disjoint union

$$X = X_1 \cup X_2 \cup \dots \cup X_\infty$$

of $\aleph_0$ nonempty sets, and otherwise $X = \cup_n X_n$ is a disjoint union of $\aleph_0$ nonempty sets.

(d) Conclude that if κ is a cardinal and $2^\kappa < 2^{\aleph_0}$, then κ is finite.

It cannot be proved without some form of Choice (Countable AC suffices) that an infinite subset of $\mathbb{R}$ contains a countably infinite subset [].

II.9.9.14. (?) Show (without AC) that if κ and λ are cardinals ≥ 2, at least one of which is infinite, and $\kappa \leq \lambda$, then $\kappa^\lambda (= \kappa \uparrow_c \lambda) \geq \lambda^\kappa$. In particular, if κ is infinite, then $\kappa^2 \leq 2^\kappa$. (We can have $\kappa^\lambda = \lambda^\kappa$ even if $\kappa < \lambda$ and both are infinite. For example, if $\kappa = \lim\{\aleph_0, 2^{\aleph_0}, 2^{2^{\aleph_0}}, \dots\}$, then $2^\kappa = \aleph_0^\kappa = \kappa^{\aleph_0} = \kappa^\kappa$ [**?**, VIII.3].)

II.9.9.15. (a) Show (without AC) that cardinal addition and multiplication satisfy the associative, commutative, and distributive laws.

(b) Show (without AC) that cardinal exponentiation satisfies the usual rules of exponents: if κ, λ, μ are cardinals, then $\kappa^\lambda \cdot \kappa^\mu = \kappa^{\lambda+\mu}$, $\kappa^\mu \cdot \lambda^\mu = (\kappa \cdot \lambda)^\mu$, $(\kappa^\lambda)^\mu = \kappa^{\lambda \cdot \mu}$.

(c) Show (without AC) that if $\kappa_1 \leq \kappa_2$ and $\lambda_1 \leq \lambda_2$, then $\kappa_1 + \lambda_1 \leq \kappa_2 + \lambda_2$, $\kappa_1 \cdot \lambda_1 \leq \kappa_2 \cdot \lambda_2$, and $\kappa_1^{\lambda_1} \leq \kappa_2^{\lambda_2}$.

(d) If $\kappa_1 < \kappa_2$ and $\lambda_1 \leq \lambda_2$, it is not necessarily true that $\kappa_1 + \lambda_1 < \kappa_2 + \lambda_2$ and $\kappa_1 \cdot \lambda_1 < \kappa_2 \cdot \lambda_2$, even with AC: we have $\aleph_0 + \aleph_1 = \aleph_1 + \aleph_1 = \aleph_1$ and $\aleph_0 \cdot \aleph_1 = \aleph_1^2 = \aleph_1$. If $\kappa_1 < \kappa_2$ and $\lambda_1 < \lambda_2$, and the Axiom of Choice is assumed, then $\kappa_1 + \lambda_1 < \kappa_2 + \lambda_2$ and $\kappa_1 \cdot \lambda_1 < \kappa_2 \cdot \lambda_2$. (This cannot be proved without AC; in fact, each of these is equivalent to AC [Sie65, XVI.2]. [Let κ be an infinite cardinal and $\lambda = \aleph_0 \cdot \kappa$; then $2 \cdot \lambda = \lambda$. If $\lambda \not\leq \aleph^+(\lambda)$, then $\lambda < \lambda + \aleph^+(\lambda)$ and $\aleph^+(\lambda) < \lambda + \aleph^+(\lambda)$, so

$$\lambda + \aleph^+(\lambda) < 2\lambda + 2\aleph^+(\lambda) = \lambda + \aleph^+(\lambda)$$

which is a contradiction. Thus $\kappa \leq \lambda \leq \aleph^+(\lambda)$ and κ is an aleph. The argument for the second inequality is similar, using $\lambda = \kappa^{\aleph_0}$.])

(e) Even if $\kappa_1 < \kappa_2$ and $\lambda_1 < \lambda_2$, and the Axiom of Choice is assumed, it is not necessarily true that $\kappa_1^{\lambda_1} < \kappa_2^{\lambda_2}$: if κ is as in Exercise (), then $\kappa^2 = \kappa$, and we have $(2^\kappa)^\kappa = 2^{\kappa^2} = 2^\kappa = \kappa^{\aleph_0}$ (A. Tarski; cf. [Sie65, VIII.3].)

II.9.9.16. Suppose $\{X_i : i \in I\}$ is an indexed collection of sets. Let $X = \cup_{i \in I} X_i$, and for each i let $Y_i = \{i\} \times X_i \subseteq I \times X$. Then $\{Y_i : i \in I\}$ is an indexed collection of *disjoint* sets with $|Y_i| = |X_i|$ for each i.

II.9.9.17. Fix an ordinal σ. Let ω_σ be $\aleph_\sigma$ regarded as an ordinal. Let Q_σ be the set of all sequences from ω_σ which are eventually 0. Give Q_σ the lexicographic ordering.

(a) Show that Q_0 is order-isomorphic to $\mathbb{Q} \cap [0, \infty)$.

(b) Show that $card(Q_\sigma) = \aleph_\sigma$.

(c) Show that Q_σ has a cofinal subset order-isomorphic to ω_σ, and an explicit order-isomorphism from Q_σ to each half-open segment between consecutive elements of this cofinal subset. [Consider the set of sequences which are 0 after the first term.]

(d) If α is an ordinal with $card(\alpha) \leq \aleph_\sigma$, show that there is a cofinal subset of α of order type $\leq \omega_\sigma$, with the half-open segments between consecutive elements of the subset of order type $< \alpha$. [If $card(\alpha) = \aleph_\sigma$, let $\{\beta_\tau : \tau < \omega_\sigma\}$ be an enumeration of the elements of α. For each $\rho < \omega_\sigma$, let $\gamma_\rho = \beta_\tau$ for the first τ for which β_τ is larger than γ_π for all $\pi < \rho$, if there is one.]

(e) Using (c) and (d), define by transfinite induction a specific order-embedding of α into Q_σ, for each ordinal α with $card(\alpha) \leq \aleph_\sigma$.

(f) Show (without AC or CH), by the argument of Exercise () using Q_σ in place of $\mathbb{Q}$, that $\aleph_{\sigma+1} \leq^* 2^{\aleph_\sigma}$ (A. Tarski [].)

*****Does this actually work without AC? In (d), must choose a bijection between α and ω_σ.

II.9.9.18. Let χ be the first measurable cardinal. This problem will show that χ is inaccessible.

(a) Use problem II.9.9.5.(c) to show that, if $\{\kappa_i : i \in I\}$ is a set of cardinals with $\kappa_i < \chi$ for all i and $card(I) < \chi$, then $\kappa = \lim\{\kappa_i\} < \chi$. [If X has cardinality κ, write $X = \cup_{i \in I} X_i$, where $card(X_i) = \kappa_i$.]

(b) If $\kappa < \chi$, show that $2^\kappa < \chi$ as follows. Let X be a set of cardinality κ, and μ a $\{0,1\}$-valued measure on $(\mathcal{P}(X), \mathcal{P}(\mathcal{P}(x)))$. For each $x \in X$, let

$$\mathcal{A}_x = \{Y \subseteq X : x \in Y\} \ , \quad \mathcal{B}_x = \{Y \subseteq X : x \notin Y\} \ .$$

Then $\mathcal{A}_x$ and $\mathcal{B}_x$ are complementary subsets of $\mathcal{P}(X)$, so either $\mu(\mathcal{A}_x) = 1$ or $\mu(\mathcal{B}_x) = 1$. Set

$$S = \{x \in X : \mu(\mathcal{A}_x) = 1\}$$

$$\mathcal{S} = \bigcap_{x \in S} \mathcal{A}_x \cap \bigcap_{x \notin S} \mathcal{B}_x \ .$$

(i) Show that $\mu(\mathcal{S}) = 1$. [Consider the complement, and use the fact from Problem 8(c) that μ is κ-additive.]

(ii) Show that $\mathcal{S} = \{S\}$.

(iii) Conclude that $2^\kappa < \chi$.

(c) If $\kappa < \chi$, use (b) to conclude that the successor of κ is also less than χ. Conclude that χ is inaccessible.

II.10. Measurable Cardinals

II.10.1. Introduction

In this section, we discuss two fascinating set-theoretic problems which originally arose in the development of Lebesgue measure. Throughout this section, we will need to assume the Axiom of Choice, which will be used without comment.

II.10.1.1. In Lebesgue's original work, he asked whether every subset of $\mathbb{R}$ is Lebesgue measurable, or if Lebesgue measure can at least be extended to a translation-invariant measure defined on all subsets of $\mathbb{R}$. Vitali's example of a nonmeasurable set (XVII.4.10.5.) gave a negative answer (if AC is assumed).

Banach weakened the question to just ask: is there a measure defined on all subsets of $\mathbb{R}$, not translation-invariant, which extends Lebesgue measure? More generally, is there a finite nonzero measure on $[0,1]$, or any set X, defined on all subsets of X, in which every singleton has measure 0? It is obvious that the answer to this question for a set X depends only on the cardinality of X.

II.10.1.2. Definition. Let X be a set. A measure μ on $(X, \mathcal{P}(X))$ is a *full real-valued measure* on X if it is a nonzero finite measure in which $\mu(\{x\}) = 0$ for every $x \in X$. A *full* $\{0,1\}$*-valued measure* is a full real-valued measure taking only the values 0 and 1.

In set theory, a full real-valued measure on a set X is usually just called a *real-valued measure*, but this term is ambiguous in a measure-theory context (the main point, of course, is that a full real-valued measure is defined for all subsets). A full real-valued measure μ can be, and usually is, renormalized so that $\mu(X) = 1$ (a *full real-valued probability measure*).

We then can phrase Banach's question as:

II.10.1.3. Question. Does $[0,1]$ (or $\mathbb{R}$) have a full real-valued measure? Does there exist a set X with a full real-valued measure on X?

It is obvious from countable additivity that a set with a full real-valued measure must be uncountable.

II.10.1.4. Definition. Let κ be a cardinal. A measure μ on a set $(X, \mathcal{P}(X))$ is κ^+*-additive* if, whenever $\{A_i : i \in I\}$ is an indexed collection of pairwise disjoint subsets of X, with $|I| \leq \kappa$, then

$$\mu(\cup_{i\in I} A_i) = \sum_{i\in I} \mu(A_i) \ .$$

The measure μ is κ*-additive* if it is λ^+-additive for all $\lambda < \kappa$.

This terminology is a little confusing and somewhat counterintuitive, but is standard in set theory and the theory of large cardinals. Note that "$\aleph_0^+$-additive" and "$\aleph_1$-additive" both mean "countably additive"; "$\aleph_0$-additive" is the same as "finitely additive."

II.10.1.5. Proposition. Let κ be a cardinal, X a set of cardinality κ, and μ a finite (or σ-finite) measure on $(X, \mathcal{P}(X))$. Then the following are equivalent:

(i) Whenever $\{A_i : i \in I\}$ is a collection of subsets of X with $|I| \leq \kappa$ and $\mu(A_i) = 0$ for all i, we have $\mu(\cup_{i\in I} A_i) = 0$.

(ii) Whenever $\{A_i : i \in I\}$ is a collection of pairwise disjoint subsets of X with $|I| \leq \kappa$ and $\mu(A_i) = 0$ for all i, we have $\mu(\cup_{i\in I} A_i) = 0$.

(iii) μ is κ^+-additive.

PROOF: (i) $\Rightarrow$ (ii) and (iii) $\Rightarrow$ (ii) are trivial.
(ii) $\Rightarrow$ (i): Suppose μ satisfies (ii), and $\mu(A_i) = 0$ for $i \in I$. Fix a well ordering on I, and for each i let $B_i = A_i \setminus [\cup_{j<i} A_j]$ (set $B_{i_0} = A_{i_0}$ if i_0 is the first element of I). Then the B_i are pairwise disjoint, $\mu(B_i) = 0$ since $B_i \subseteq A_i$, and

$$\mu(\cup_{i\in I} A_i) = \mu(\cup_{i\in I} B_i) = \sum_{i\in I} \mu(B_i) = 0 \ .$$

(ii) $\Rightarrow$ (iii): Suppose μ satisfies (ii), and that $\{A_i : i \in I\}$ is a collection of pairwise disjoint subsets of X with $|I| \leq \kappa$. Then $J = \{i \in I : \mu(A_i) > 0\}$ is countable by (), and we have $\mu(\cup_{i\in I\setminus J} A_i) = 0$ by (ii), since $|I \setminus J| \leq \kappa$. Thus

$$\mu(\cup_{i\in I} A_i) = \mu(\cup_{i\in J} A_i) + \mu(\cup_{i\in I\setminus J} A_i) = \sum_{i\in J} \mu(A_i) + 0 = \sum_{i\in I} \mu(A_i) \ .$$



So, if μ is a full real-valued measure on X, then μ is κ-additive if and only if the union of fewer than κ sets of measure zero has measure zero.

A key observation made by BANACH was

II.10.1.6. PROPOSITION. Let κ be the smallest cardinal for which there exists a set of cardinality κ with a full real-valued measure, and let μ be a full real-valued measure on a set X (not necessarily of cardinality κ). Then μ is κ-additive.

PROOF: Suppose that $\{A_i : i \in I\}$ is an indexed collection of pairwise disjoint subsets of X with $\mu(A_i) = 0$ for all i and $|I| < \kappa$. Let $A = \cup_{i\in I} A_i$. If $\mu(A) > 0$, for $S \subseteq I$ let $\nu(S) = \mu(\cup_{i\in S} A_i)$. Then it is easy to check that ν is a full real-valued measure on I, contradicting minimality of κ. Thus $\mu(A) = 0$ and μ is κ-additive.

II.10.2. Real-Valued Measurable Cardinals

In light of II.10.1.6., we make the following definition:

II.10.2.1. DEFINITION. A cardinal κ is *real-valued measurable* if there is a κ-additive full real-valued measure on a set of cardinality κ.

Note that if κ is any cardinal larger than a real-valued measurable cardinal, then there is a full real-valued measure on a set of cardinality κ (supported on a subset of real-valued measurable cardinality), but the measure will not be κ-additive in general. But II.10.1.6. says that if κ is the *smallest* cardinal of a set supporting a full real-valued measure, then κ is a real-valued measurable cardinal. Thus there exists a set with a full real-valued measure if and only if there exists a real-valued measurable cardinal.

II.10.2.2. PROPOSITION. Let κ be a real-valued measurable cardinal. Then κ is an uncountable regular cardinal ().

PROOF: We have already observed that κ is uncountable.

Let X be a set of cardinality κ and μ a κ-additive full real-valued measure on X. Note that if A is a subset of X with $|A| < \kappa$, then $\mu(A) = 0$ by κ-additivity. So if $\{A_i : i \in I\}$ is an indexed collection of pairwise disjoint subsets of X with $|A_i| < \kappa$ for all i and $|I| < \kappa$, then $\mu(A_i) = 0$ for all i and hence $\mu(\cup_{i\in I} A_i) = 0$, so we cannot have that $X = \cup_{i\in I} A_i$. 👍

BANACH and KURATOWSKI proved that if the Continuum Hypothesis is assumed, there is no full real-valued measure on $\mathbb{R}$ or $[0, 1]$. Then BANACH extended the arguments to show that if the Generalized Continuum Hypothesis () is assumed, then any real-valued measurable cardinal is inaccessible ().

The next major contributions were made by S. ULAM. In his Ph.D. dissertation, he showed that every real-valued measurable cardinal is weakly inaccessible without the use of the GCH. The device he used is now called an *Ulam matrix*. The construction uses cardinal and ordinal notation and terminology; see () for an explanation.

II.10.2.3. LEMMA. [ULAM MATRIX] Let λ be a cardinal, and κ the successor cardinal of λ. Regard λ and κ as sets of ordinals. Then there are subsets $\{A^\gamma_\alpha : \alpha < \kappa, \gamma < \lambda\}$ of $\kappa = \{\sigma : \sigma < \kappa\}$ (α, β, γ, σ, τ denote ordinals) such that

(i) If $\alpha < \beta < \kappa$ and $\gamma < \lambda$, then $A^\gamma_\alpha \cap A^\gamma_\beta = \emptyset$.

(ii) For any $\alpha < \kappa$, $[\kappa \setminus \cup_{\gamma<\lambda} A^\gamma_\alpha] \subseteq \alpha$, and in particular $|\kappa \setminus \cup_{\gamma<\lambda} A^\gamma_\alpha| < \kappa$.

The A^γ_α can be thought of as an array of subsets with κ rows and λ columns, with α denoting the row and γ the column. The conditions say that the sets in each column are pairwise disjoint, and the union of each row is all but a "small" subset of the whole set κ.

PROOF: If $\sigma < \kappa$, then $|\sigma + 1| \leq \lambda$, so there is a surjective function $f_\sigma : \lambda \to \sigma + 1$. Fix such a function for each σ. If $\alpha < \kappa$ and $\gamma < \lambda$, let

$$A^\gamma_\alpha = \{\sigma < \kappa : f_\sigma(\gamma) = \alpha\} .$$

Property (i) is obvious. For (ii), fix $\alpha < \kappa$. Suppose $\sigma < \kappa$ and $\sigma \notin \cup_{\gamma<\lambda} A^\gamma_\alpha$. Then there is no γ such that $f_\sigma(\gamma) = \alpha$, i.e. α is not in the range of f_σ, which is $\sigma + 1 = \{\tau : \tau \leq \sigma\}$. Thus $\sigma < \alpha$. 👍

II.10.2.4. THEOREM. Let κ be a real-valued measurable cardinal. Then κ is weakly inaccessible.

PROOF: It has already been shown (II.10.2.2.) that κ is uncountable and regular; it remains to show that it is a limit cardinal. If κ were the successor of a cardinal λ, let $\{A^\gamma_\alpha\}$ be an Ulam matrix as in II.10.2.3.. If μ is a κ-additive full real-valued measure on κ, set $S = \{(\alpha, \gamma) : \mu(A^\gamma_\alpha) > 0\} \subseteq \kappa \times \lambda$. For each α, we have $\mu(\kappa \setminus \cup_{\gamma<\lambda} A^\gamma_\alpha) = 0$ by II.10.2.3.(ii) and κ-additivity, and hence $\mu(\cup_{\gamma<\lambda} A^\gamma_\alpha) > 0$; then by κ-additivity $\mu(A^\gamma_\alpha) > 0$ for some γ, i.e. for each $\alpha < \kappa$ there is at least one γ with $(\alpha, \gamma) \in S$. Thus $|S| \geq \kappa$. On the other hand, for each $\gamma < \lambda$ there are at most countably many α such that $(\alpha, \gamma) \in S$ by II.10.2.3.(i) and (), since μ is finite. Hence $|S| \leq \aleph_0 \cdot \lambda = \lambda$, a contradiction. Thus κ is not a successor cardinal. 👍

More recently, R. Solovay [Sol71] proved that if κ is a real-valued measurable cardinal, then the cardinality of the set of weakly inaccessible cardinals less than κ is κ (Solovay's result is actually considerably stronger: a real-valued measurable cardinal is weakly Mahlo).

Ulam's Dichotomy

Ulam then observed that Question II.10.1.3. naturally divides into two rather distinct questions, both of considerable interest. Recall that a set $A \in \mathcal{A}$ in a measure space $(X, \mathcal{A}, \mu)$ is an *atom* if $\mu(A) > 0$ and for every $B \in \mathcal{A}$, $B \subseteq A$, either $\mu(B) = 0$ or $\mu(A \setminus B) = 0$, and that $(X, \mathcal{A}, \mu)$ is *atomless* if there are no atoms.

II.10.2.5. Proposition. Let μ be a full real-valued measure on a set X. If $(X, \mathcal{P}(X), \mu)$ has an atom, then there is a full $\{0,1\}$-valued measure on X.

Proof: Let A be an atom in $(X, \mathcal{P}(X), \mu)$. For $B \subseteq X$, define

$$\nu(B) = \frac{\mu(A \cap B)}{\mu(A)} \ .$$

It is easily checked that ν is a full $\{0,1\}$-valued measure on X. 👍

So Question II.10.1.3. splits into the next two questions:

II.10.2.6. Question. Does there exist an atomless full real-valued measure on a set X?

II.10.2.7. Question. Does there exist a full $\{0,1\}$-valued measure on a set X?

II.10.3. Atomless Full Real-Valued Measures

We will discuss Questions II.10.2.6. and II.10.2.7. separately. We first analyze the properties of an atomless full real-valued measure (if one exists). Ulam proved the following theorem:

II.10.3.1. Theorem. Let μ be an atomless full real-valued measure on a set X. Then μ is not $(2^{\aleph_0})^+$-additive.

Before we prove this theorem, let us draw the most important consequence:

II.10.3.2. Corollary. Suppose there is a set with an atomless full real-valued measure. Let κ be the smallest real-valued measurable cardinal. Then $\kappa \leq 2^{\aleph_0}$.

Proof: Let X be a set with an atomless full real-valued measure μ on X. Then μ is not $(2^{\aleph_0})^+$-additive by II.10.3.1.. But μ is κ-additive by II.10.1.6.. Thus $\kappa \leq 2^{\aleph_0}$. 👍

This result, combined with II.10.2.4., puts severe set-theoretic limitations on the existence of atomless full real-valued measures: they can only exist if $2^{\aleph_0}$ is at least as large as the first weakly inaccessible cardinal (in fact, by Solovay's result, the first weakly Mahlo cardinal). It is believed that this is consistent with ZFC, but it would be in dramatic contradiction to the Continuum Hypothesis and would imply that $2^{\aleph_0}$ is very large.

The proof of II.10.3.1. mostly consists of the next three lemmas. These results hold for arbitrary atomless finite measures, not just full ones.

II.10.3.3. Lemma. Let $(X, \mathcal{A}, \mu)$ be an atomless measure space, and $Y \in \mathcal{A}$ with $0 < \mu(Y) < \infty$. For any $\epsilon > 0$, there is an $A \in \mathcal{A}$, $A \subseteq Y$, with $(1-\epsilon)\mu(Y) < \mu(A) < \mu(Y)$.

Proof: Since Y is not an atom, it decomposes as a disjoint union of sets A_1, B_1 of strictly positive measure. We may assume that $\mu(A_1) \geq \mu(B_1)$; then $\mu(A_1) \geq \mu(Y)/2$. Then B_1 decomposes as a disjoint union of A_2 and B_2, of strictly positive measure, with $\mu(A_2) \geq \mu(B_1)/2$. Continue inductively; for any n we have

$$(1 - 2^{-n})\mu(Y) \leq \mu(\cup_{k=1}^{n} A_k) < \mu(Y)$$

and we can take $A = \cup_{k=1}^{n} A_k$ for n large enough that $2^{-n} < \epsilon$. 👍

II.10.3.4. Lemma. Let $(X, \mathcal{A}, \mu)$ be an atomless measure space, and $Y \in \mathcal{A}$ with $0 < \mu(Y) < \infty$. If $0 < t < \mu(Y)$, there is a subset A of Y, $A \in \mathcal{A}$, with $\mu(A) = t$.

Proof: For countable ordinals σ, define subsets $A_\sigma \in \mathcal{A}$ by transfinite induction as far as possible by setting $A_0 = Y$; if $\mu(A_\sigma) > t$, let $A_{\sigma+1}$ be a subset of A_σ with $t < \mu(A_{\sigma+1}) < \mu(A_\sigma)$ (using II.10.3.3.); and if σ is a (countable) limit ordinal, set $A_\sigma = \cap_{\tau<\sigma} A_\tau$. If $\mu(A_\sigma) = t$ for some σ, terminate the construction and we are done. In fact, this must happen at some countable (limit) ordinal, since the $\mu(A_\sigma)$ strictly decrease, and if $\mu(A_\sigma) > t$ for all $\sigma < \omega_1$ we would have that $\sigma \mapsto -\mu(A_\sigma)$ would be an order-embedding of ω_1 into $\mathbb{R}$, which is impossible. (Alternatively, if for each σ we set $B_\sigma = A_\sigma \setminus A_{\sigma+1}$, then the B_σ would be an uncountable family of pairwise disjoint subsets of Y of strictly positive measure.) 👍

II.10.3.5. Lemma. Let $(X, \mathcal{A}, \mu)$ be an atomless measure space, and $Y \in \mathcal{A}$ with $0 < \mu(Y) < \infty$. Then there are subsets $\{A_t : 0 \leq t \leq 1\} \subseteq \mathcal{A}$ of Y such that

(i) $A_s \subseteq A_t$ for $s < t$.

(ii) $A_t = \cap_{s>t} A_s$ for $0 \leq t < 1$ and $A_1 = Y$.

(iii) $\mu(A_t) = t \cdot \mu(Y)$ for $0 \leq t \leq 1$.

Proof: Let $\{r_0, r_1, r_2, \dots\}$ be an enumeration of $\mathbb{Q} \cap [0,1]$ with $r_0 = 0$ and $r_1 = 1$. Set $B_0 = \emptyset$ and $B_1 = Y$. Inductively define B_{r_n} for $n \geq 3$ so that $B_{r_n} \subseteq B_{r_m}$ if $r_n < r_m$ and $\mu(B_{r_n}) = r_n \cdot \mu(Y)$, using II.10.3.4. [at stage n, if r_p and r_q are the closest numbers with $r_p < r_n < r_q$ and $p, q < n$, apply II.10.3.4. to $B_{r_q} \setminus B_{r_p}$

and $t = (r_n - r_p)\mu(Y)$ and add B_{r_p}]. Set $A_1 = B_1 = Y$ and for $0 \leq t < 1$ set $A_t = \cap_{r_n > t} B_{r_n}$. Then (i) is obvious, and if $r_n < t < r_m$, we have $B_{r_n} \subseteq A_t \subseteq B_{r_m}$, so (ii) and (iii) follow easily. 👍

We now prove Theorem II.10.3.1..

PROOF: Let μ be an atomless full real-valued measure on a set X. Let $\{A_t : 0 \leq t \leq 1\}$ be a collection of subsets as in II.10.3.5. with $Y = X$. For each t, set $C_t = A_t \setminus [\cup_{s<t} A_s]$. We have that $\mu(C_t) = 0$ for each t. If $x \in X$, let $t = \inf(s : x \in A_s)$ (note that this set of s is nonempty since $x \in A_1 = X$). Then $x \in A_t$ by II.10.3.5.(i)–(ii), but $x \notin A_s$ if $s < t$, so $x \in C_t$. Thus $\cup_{0 \leq t \leq 1} C_t = X$, so $\mu(\cup_{0 \leq t \leq 1} C_t) > 0$ and μ is not $(2^{\aleph_0})^+$-additive. 👍

Almost the same proof shows one direction of the following:

II.10.3.6. THEOREM. There is an extension of Lebesgue measure to a measure on $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ if and only if there is a set with an atomless full real-valued measure.

PROOF: Suppose there is an atomless full real-valued probability measure μ on a set X. Define the sets A_t and C_t as in the proof of II.10.3.1., and define $f : X \to [0,1]$ by $f(x) = t$ if $x \in C_t$ (note that the C_t are pairwise disjoint). Set $\nu = f_*(\mu)$; then ν is a full real-valued measure on $[0,1]$. To see that ν extends Lebesgue measure λ on $[0,1]$, let $0 \leq t_1 < t_2 \leq 1$; then $f^{-1}((t_1, t_2]) = A_{t_2} \setminus A_{t_1}$, so

$$\nu((t_1, t_2]) = \mu(A_{t_2} \setminus A_{t_1}) = t_2 - t_1 = \lambda((t_1, t_2])$$

and thus by () ν agrees with λ on Borel subsets of $[0,1]$. Since for every Lebesgue measurable subset E of $[0,1]$ there exist Borel sets A and B with $A \subseteq E \subseteq B$ and $\lambda(A) = \lambda(E) = \lambda(B)$ (), ν must agree with λ on all Lebesgue measurable subsets of $[0,1]$. Then ν can be defined by translation on $[n, n+1]$ for all n and hence on all of $\mathbb{R}$, extending Lebesgue measure.

Conversely, suppose μ is an extension of λ to all subsets of $\mathbb{R}$. Then the restriction of μ to $[0,1]$ is a full real-valued measure on $[0,1]$. It is not obvious that this measure is atomless; however, if it were not there would be a full $\{0,1\}$-valued measure on $[0,1]$ by II.10.2.5., which is impossible by II.10.4.3.. 👍

II.10.4. Measurable Cardinals

Now we turn to Question II.10.2.7.. The proof of the next result is essentially identical to the proof of II.10.1.6., and is left to the reader.

II.10.4.1. PROPOSITION. Let κ be the smallest cardinal for which there exists a set of cardinality κ with a full $\{0,1\}$-valued measure, and let μ be a full $\{0,1\}$-valued measure on a set X (not necessarily of cardinality κ). Then μ is κ-additive.

II.10.4.2. DEFINITION. A cardinal κ is *measurable* if, on a set X of cardinality κ, there is a κ-additive full $\{0,1\}$-valued measure on X.

By II.10.4.1., the condition of κ-additivity is not crucial for the question of existence of a measurable cardinal; but it does insure that if there is a second measurable cardinal, it is tremendously larger than the first (if this condition were eliminated from the definition, any cardinal larger than a measurable cardinal would be measurable.) The κ-additivity is very important in the set-theoretic interpretations of measurability.

A measurable cardinal is, of course, a real-valued measurable cardinal. By Ulam's dichotomy and II.10.3.1., a real-valued measurable cardinal is either measurable or $\leq 2^{\aleph_0}$. Measurable cardinals are larger than $2^{\aleph_0}$ (in fact "much larger"), because of the following fact:

II.10.4.3. THEOREM. A measurable cardinal is inaccessible.

PROOF: Let κ be a measurable cardinal. Then κ is weakly inaccessible by II.10.2.4., so it remains to show that if $\lambda < \kappa$, then $2^\lambda < \kappa$. Fix $\lambda < \kappa$, and suppose $2^\lambda \geq \kappa$. Then there is a κ-additive full $\{0,1\}$-valued measure μ on $\mathcal{P}(X)$ for a set X of cardinality λ (μ is supported on a subset of $\mathcal{P}(X)$ of cardinality κ). For each $x \in X$, let

$$\mathcal{A}_x = \{Y \subseteq X : x \in Y\} \ , \qquad \mathcal{B}_x = \{Y \subseteq X : x \notin Y\} \ .$$

Then $\mathcal{A}_x$ and $\mathcal{B}_x$ are complementary subsets of $\mathcal{P}(X)$, so either $\mu(\mathcal{A}_x) = 1$ or $\mu(\mathcal{B}_x) = 1$. Set

$$S = \{x \in X : \mu(\mathcal{A}_x) = 1\}$$

$$\mathcal{S} = \bigcap_{x \in S} \mathcal{A}_x \cap \bigcap_{x \notin S} \mathcal{B}_x \ .$$

The complement of $\mathcal{S}$ is

$$\mathcal{S}^c = \bigcup_{x \notin S} \mathcal{A}_x \cup \bigcup_{x \in S} \mathcal{B}_x$$

which is a union of λ sets of measure 0, hence $\mu(\mathcal{S}^c) = 0$, $\mu(\mathcal{S}) = 1$. But it is easily seen that $\mathcal{S} = \{S\}$, and $\mu(\mathcal{T}) = 0$ for every singleton $\mathcal{T}$, a contradiction. 👍

It turns out that measurable cardinals, if they exist, are enormous:

II.10.4.4. THEOREM. [HANF-TARSKI] If κ is a measurable cardinal, then the cardinality of the set of inaccessible cardinals less than κ is κ.

In fact, "inaccessible" can be replaced by other types of large cardinal conditions weaker than measurability. See II.10.4.11. and II.10.4.12. for stronger statements about the size of measurable cardinals.

In a set X of measurable cardinal κ, a κ-additive full $\{0,1\}$-valued measure gives a way of dividing all subsets of X into "large" ones (of measure 1) and "small" ones (of measure 0). If A is any subset of X, exactly one of A, $X \setminus A$ is "large". Any subset of cardinality less than κ is "small".

Ultrafilters and Measurable Cardinals

A good alternative description of the principle of the previous paragraph uses the language of ultrafilters. Recall that a *filter* on a set X is a nonempty subset $\mathcal{F}$ of $\mathcal{P}(X)$ such that

(i) $\mathcal{F}$ is closed under supersets, i.e. if $A \in \mathcal{F}$ and $A \subseteq B \subseteq X$, then $\mathcal{B} \in \mathcal{F}$.

(ii) $\mathcal{F}$ is closed under finite intersections.

(iii) $\emptyset \notin \mathcal{F}$.

An *ultrafilter* on X is a maximal filter on X. It is easily seen that a filter $\mathcal{F}$ is an ultrafilter if and only if, for every $A \subseteq X$, either $A \in \mathcal{F}$ or $(X \setminus A) \in \mathcal{F}$. An ultrafilter $\mathcal{F}$ is *fixed* if $\cap_{A\in\mathcal{F}} A$ is nonempty (in this case, we must have $\cap_{A\in\mathcal{F}} A = \{x\}$ for some $x \in X$); $\mathcal{F}$ is *free* if it is not fixed, i.e. if $\cap_{A\in\mathcal{F}} A = \emptyset$. $\mathcal{F}$ is free if and only if the complement of any finite subset of X is in $\mathcal{F}$. (Note that AC is needed to show that free ultrafilters exist on any infinite set; in fact, by Zorn's Lemma, every filter is contained in an ultrafilter, and in particular the filter consisting of complements of finite sets is contained in a free ultrafilter.)

If κ is a cardinal, a filter $\mathcal{F}$ on a set X is called *κ-complete* if $\mathcal{F}$ is closed under taking the intersection of fewer than κ sets. Every fixed ultrafilter is κ-complete for any κ.

II.10.4.5. Example. Let κ be a measurable cardinal, and μ a κ-additive full $\{0,1\}$-valued measure on a set X of cardinality κ (e.g. $X = \kappa = \{\sigma : \sigma < \kappa\}$). Let $\mathcal{F} = \{A \subseteq X : \mu(A) = 1\}$. Then $\mathcal{F}$ is a κ-complete free ultrafilter on X.

Conversely, suppose $\mathcal{F}$ is a κ-complete free ultrafilter on a set X of cardinality κ. Define $\mu : \mathcal{P}(X) \to \{0,1\}$ by $\mu(A) = 1$ if $A \in \mathcal{F}$ and $\mu(A) = 0$ if $A \notin \mathcal{F}$. It is easily checked that μ is a κ-additive full $\{0,1\}$-valued measure on X, so κ is a measurable cardinal.

So we obtain a rephrasing of the definition of measurable cardinals in the language of ultrafilters:

II.10.4.6. Proposition. Let κ be a cardinal. Then κ is measurable if and only if $\kappa > \aleph_0$ and there is a κ-complete free ultrafilter on κ (or any set of cardinality κ).

In fact, there is a natural one-one correspondence between full $\{0,1\}$-valued measures on a set X and free $\aleph_1$-complete ultrafilters on X, as in II.10.4.5..

Normal Ultrafilters and Nodal Sets

If $\mathcal{F}$ is an ultrafilter on a set X and $f : X \to X$ is a bijection, then $f(\mathcal{F}) = \{f(A) : A \in \mathcal{F}\}$ is also an ultrafilter on X, which is κ-complete if $\mathcal{F}$ is. Thus if X is a set of measurable cardinality κ, $\mathcal{F}$ is a κ-complete free ultrafilter on X, $A \in \mathcal{F}$ with $|A| = |X \setminus A| = \kappa$, and $f : X \to X$ is a bijection interchanging A and $X \setminus A$, then $f(\mathcal{F})$ is a κ-complete free ultrafilter on X with $(X \setminus A) \in f(\mathcal{F})$. Thus the intersection of all κ-complete free ultrafilters on X is the (κ-complete) filter of all subsets whose complement has cardinality less than κ. So there does not appear to be any way to systematically distinguish "size" for sets like A using κ-complete free ultrafilters.

However, it turns out that if $X = \kappa$, the ordering on κ does allow a distinction to be made on the "size" of certain sets $A \subseteq \kappa$ even if $|A| = |\kappa \setminus A| = \kappa$. This is done through the notion of a normal ultrafilter.

A subset of κ is *cobounded* if it contains $\{\sigma : \alpha \leq \sigma < \kappa\}$ for some $\alpha < \kappa$. A filter is *cobounded* if it contains all cobounded sets; the collection $\mathcal{C}_\kappa$ of all cobounded sets is a filter which is the smallest cobounded filter. A κ-complete filter on κ must be cobounded. Any cobounded filter is free.

II.10.4.7. DEFINITION. Let κ be a cardinal, and $\mathcal{F}$ a cobounded filter on κ. Then $\mathcal{F}$ is *normal* if, whenever $\{S_\alpha : \alpha < \kappa\}$ is an indexed collection of sets in $\mathcal{F}$, we have that

$$\Delta_{\alpha<\kappa} S_\alpha = \{\sigma < \kappa : \sigma \in \cap_{\alpha<\sigma} S_\alpha\} \in \mathcal{F} \ .$$

$\Delta_{\alpha<\kappa} S_\alpha$ is called the *diagonal intersection* of the collection $\{S_\alpha\}$.

By definition, a normal filter is cobounded. It is a little hard to motivate this definition intuitively. It means that for any collection $\{S_\alpha\}$ of sets in $\mathcal{F}$, "most" σ's in κ belong to X_α for all $\alpha < \sigma$. Normal filters are rather rare. For example, for each $n \in \aleph_0$ let $S_n = \{m \in \aleph_0 : m > n+1\}$. If $\mathcal{F}$ is any cobounded filter on $\aleph_0$, then this is an indexed set in $\mathcal{F}$; but $\Delta_{n<\aleph_0} S_n = \emptyset$. Thus there are no normal filters on $\aleph_0$.

II.10.4.8. PROPOSITION. If there is a normal filter $\mathcal{F}$ on κ, then $\mathcal{F}$ is κ-complete.

PROOF: Let $\{X_i : i \in I\}$ be an indexed collection of sets in $\mathcal{F}$, with $|I| = \lambda < \kappa$. Let $f : \lambda \to I$ be a bijection; set $S_\alpha = X_{f(\alpha)}$ for $\alpha < \lambda$, and $S_\alpha = \kappa$ for $\alpha \in B := \{\sigma : \lambda \le \sigma < \kappa\}$. Then $\Delta_{\alpha<\kappa} S_\alpha \in \mathcal{F}$, so

$$A := B \cap [\Delta_{\alpha<\kappa} S_\alpha] \in \mathcal{F}$$

since $\mathcal{F}$ is cobounded. But it is easy to see that $A = B \cap [\cap_{i\in I} X_i] \subseteq \cap_{i\in I} X_i$, so $\cap_{i\in I} X_i \in \mathcal{F}$. 👍

II.10.4.9. COROLLARY. If there is a normal ultrafilter on κ, then κ is measurable.

The converse is true also (cf. [**?**], [Kan03, 5.12]):

II.10.4.10. THEOREM. A measurable cardinal has a normal ultrafilter.

Actually, a measurable cardinal may have many normal ultrafilters: it is equiconsistent with the existence of measurable cardinals that there are 2^{2^κ} normal ultrafilters over each measurable cardinal κ ([KP71]; cf. [Kan03, 17.8]) (it is also equiconsistent that the smallest measurable cardinal has only one normal ultrafilter [Kun70]). There are 2^{2^κ} free ultrafilters on κ for any (infinite) cardinal κ.

Sets which are in the intersection of all normal ultrafilters on a measurable cardinal κ are called *nodal*, and can be considered "large" in a set-theoretic sense. The set of cardinals less than κ is intuitively a "small" subset of κ (there are intuitively many more ordinals which are not cardinals); however, it turns out that the set of cardinals less than κ is a nodal subset of κ. Much more is true:

II.10.4.11. THEOREM. If κ is a measurable cardinal, then the set of inaccessible cardinals less than κ is a nodal subset of κ.

This is at first glance surprising, since the inaccessible cardinals are seemingly very sparsely distributed even among the cardinals. The Hanf-Tarski Theorem (II.10.4.4.) is an immediate corollary. There are several much stronger results in [Kan03], replacing "inaccessible" with stronger large cardinal properties. The best known result along this line is the following result of F. ROWBOTTOM (cf. [Kan03, 7.19]):

II.10.4.12. THEOREM. Let κ be a measurable cardinal. Then the set of Ramsey cardinals less than κ is a nodal subset of κ. Thus the cardinality of the set of Ramsey cardinals less than κ is κ.

This result gives some idea of the enormity of measurable cardinals.

II.10.5. Consistency and Equiconsistency Results

Perhaps remarkably, it appears to be consistent with ZFC to assume the existence of a measurable cardinal. Even more remarkably, even though measurable cardinals seem to be far removed from any aspect of the "real world" (even the "real mathematical world"), the existence of a measurable cardinal has significant implications in mathematical logic about subsets of $\mathbb{N}$! (Specifically, there are subsets of $\mathbb{N}$ which are not "constructible.")

We list without proof several relative consistency results. All are discussed in detail in [Kan03]. If A and B are theories, then B is of *greater consistency strength* than A if A can be proved consistent assuming that B is consistent, and A and B are *equiconsistent* if each is of greater consistency strength than the other. We also use the term *strictly greater consistency strength*, with the obvious meaning. For example, ZF, ZFC, and ZFC+CH are equiconsistent by the results of GÖDEL. (Recall the fundamental theorem of model theory: a theory is consistent if and only if it has a model. The proofs that B is of greater consistency strength than A generally proceed by constructing a model for A from a model for B.)

II.10.5.1. ZFC is not known to be consistent, and this may not be provable using existing theories. ZFC is presumed to be consistent, though, and all of set theory needs to be drastically altered if it is not. In the rest of this section (and throughout this book), we assume ZFC is consistent.

"God exists since mathematics is consistent, and the Devil exists since we cannot prove it."

André Weil[17]

II.10.5.2. It cannot be proved in ZFC that an inaccessible cardinal exists, i.e.

ZFC+(There does not exist an inaccessible cardinal)

is equiconsistent with ZFC. This is discussed in (). It cannot even be proved in ZFC that a weakly inaccessible cardinal exists. Moreover, it cannot be proved that existence of an inaccessible cardinal is relatively consistent with ZFC, since existence of such a proof would contradict Gödel's Incompleteness Theorem (cf. [**?**]). Thus, unlike the situation with CH,

ZFC+(There exists an inaccessible cardinal)

is of strictly greater consistency strength than ZFC.

There is no current reason to believe, however, that even the existence of a measurable cardinal is inconsistent with ZFC, or with ZFC+GCH.

[17] In H. Eves, *Mathematical Circles Adieu*, Boston: Prindle, Weber and Schmidt, 1977.

II.10.5.3. R. Solovay [Sol70] proved that

$$\text{ZFC+(There exists an inaccessible cardinal)}$$

is of greater consistency strength than

$$\text{ZF+DC+(Every subset of } \mathbb{R} \text{ is Lebesgue measurable)}$$

(plus several other properties of $\mathbb{R}$). S. Shelah [She84] showed that

$$\text{ZF+(Countable AC)+(Every subset of } \mathbb{R} \text{ is Lebesgue measurable)}$$

is of greater consistency strength than

$$\text{ZFC+(There exists an inaccessible cardinal)}$$

and thus all of these are equiconsistent. So, roughly speaking, the necessity of using AC to construct a nonmeasurable subset of $\mathbb{R}$ is equivalent to the existence of an inaccessible cardinal.

II.10.5.4. (Solovay [Sol71]) The following are equiconsistent:

(i) ZFC+(There exists a measurable cardinal).

(ii) ZFC+($2^{\aleph_0}$ is real-valued measurable).

II.10.5.5. (J. Silver [Sil71]) If ZFC+(There exists a measurable cardinal) is consistent, so is

$$\text{ZFC+GCH+(There exists a measurable cardinal)}$$

and thus these two are equiconsistent.

II.10.5.6. The existence of a real-valued measurable cardinal $\leq 2^{\aleph_0}$ is obviously (from II.10.2.4.) inconsistent with the Continuum Hypothesis. It is also inconsistent with Martin's Axiom. One version of Martin's Axiom is a variant of the Baire property:

Martin's Axiom. If X is a compact Hausdorff space with the *Countable Chain Condition* (every family of pairwise disjoint nonempty open subsets in X is countable), then the intersection of fewer than $2^{\aleph_0}$ dense open sets in X is dense.

Martin's Axiom (MA) can be thought of as a weak version of CH which is more palatable to some set theorists, although it is far from universally accepted. Since CH implies MA, ZFC+MA is equiconsistent with ZFC. See [Fre84] for details about Martin's Axiom and its consequences.

Incidentally, both real-valued measurability of $2^{\aleph_0}$ and Martin's Axiom imply the following curious result: for every cardinal κ with $\aleph_0 \leq \kappa < 2^{\aleph_0}$, we have $2^{\kappa} = 2^{\aleph_0}$.

II.10.5.7. (D. Scott [Sco61]) The existence of a measurable cardinal is inconsistent with Gödel's constructibility hypothesis $V = L$ (II.8.9.7.). This result can be interpreted two ways: it can be taken as evidence that measurable cardinals should not exist, or that Gödel's hypothesis is not reasonable. Most set theorists seem to take the latter view. Thus the existence of a measurable cardinal implies that there is a nonconstructible subset of $\mathbb{N}$. See [Dev84], [Dev77], and [Dev73] for more on constructibility.

II.10.5.8. **Grothendieck Universe.** A *Grothendieck universe* (in ZF) is a set $\mathfrak{U}$ with the following properties:

(i) $\omega \in \mathfrak{U}$ (in particular, $\mathfrak{U}$ is not empty).

(ii) If $x \in \mathfrak{U}$, then $x \subseteq \mathfrak{U}$ (i.e. $\mathfrak{U}$ is transitive).

(iii) If $x, y \in \mathfrak{U}$, then $\{x, y\} \in \mathfrak{U}$.

(iv) If $x \in \mathfrak{U}$, then $\mathcal{P}(x) \in \mathfrak{U}$ (so $\mathcal{P}(x) \subseteq \mathfrak{U}$).

(v) If $\{x_i : i \in I\}$ is an indexed collection of elements of $\mathfrak{U}$, and $I \in \mathfrak{U}$, then $\cup_{i \in I} x_i \in \mathfrak{U}$.

(a) If $\mathfrak{U}$ is a Grothendieck universe, show that the elements of $\mathfrak{U}$ form a model for ZF set theory. If "set" is interpreted as "element of $\mathfrak{U}$" and "class" is interpreted as "subset of $\mathfrak{U}$," then these sets and classes form a model of NBG set theory.

(b) If inaccessible cardinals exist, let κ be the smallest inaccessible cardinal. Show that V_κ (II.8.9.7.) is a Grothendieck universe, and is contained in every Grothendieck universe. Thus ZF+(Inaccessible cardinals exist) implies the existence of a Grothendieck universe.

(c) If ZF is consistent, show that the existence of a Grothendieck universe cannot be proved in ZF. [V_κ would give a model of ZF with no Grothendieck universe.]

(d) If a Grothendieck universe $\mathfrak{U}$ exists, show that an inaccessible cardinal exists. [Consider the set of all cardinals contained in $\mathfrak{U}$.] Thus existence of a Grothendieck universe is equivalent to existence of an inaccessible cardinal.

The theories ZF+(a Grothendieck universe exists) and NBG+(a Grothendieck universe exists) are called ZFU and NBGU respectively. These theories are needed to interpret category theory properly in ZF set theory.

II.10.5.9. (a) If κ is a regular cardinal and X is a set with $\kappa = card(X)$, the set

$$\mathcal{F} = \{Y \subseteq X : card(X \setminus Y) < \kappa\}$$

is a κ-complete (II.10.4.5.) free filter on X.

(b) A cardinal κ is *strongly compact* (the name comes from model theory, not topology) if it is uncountable and regular, and every κ-complete filter on a set of cardinality κ is contained in a κ-complete ultrafilter on X (we need only consider free filters). A strongly compact cardinal is measurable (II.10.4.6.).

In fact, the existence of a strongly compact cardinal is considerably stronger logically than the existence of a measurable cardinal (although the first strongly compact cardinal is not necessarily larger than the first measurable cardinal, if both exist).

There are notions of cardinals larger than strongly compact cardinals: supercompact and extendible cardinals. See [Jec03] or [Kan03].

II.11. Categories and Functors

> "[T]he discovery of ideas as general as these [categories] is chiefly the willingness to make a brash or speculative abstraction, in this case supported by the pleasure of purloining words from the philosophers: 'Category' from Aristotle and Kant, 'Functor' from Carnap, and 'natural transformation' from then current informal parlance."
>
> *Saunders MacLane*[18]

Category theory is a "theory of mathematical theories." It was originally intended to provide a uniform abstract language for mathematical theories and to formulate and prove properties common to many theories. As often happens in mathematics, abstract category theory not only fulfills this original purpose but also turns out to be applicable to many other settings in mathematics as diverse as group theory, the theory of partially ordered sets, and mathematical physics; there is even a way to base an entire formulation of the foundations of mathematics on category theory rather than set theory ().

Category theory is characterized by a high degree of abstraction, and it is fashionable in some circles to view it pejoratively and call it "abstract nonsense." However, a knowledge of the bare rudiments (the barest of which we describe here) is very useful in understanding the general nature of mathematical theories and how different parts of mathematics can be viewed as parts of a common whole.

A value-neutral nutshell description of category theory is "the theory of arrow-theoretic results." Standard references for category theory are [Mit65], [ML98], [Sch72], and [LS97].

II.11.1. Mathematical Theories

II.11.1.1. A mathematical theory generally consists of two parts:

(1) A collection of objects which are sets with additional structure (algebraic, topological, etc.)

(2) Functions between the objects which preserve the structure.

II.11.1.2. There are many familiar examples, some of which are described in more detail in this book. Here are a few:

Vector spaces (over a fixed field, e.g. $\mathbb{R}$) and linear transformations

Groups and group homomorphisms

Metric spaces and continuous functions

Topological spaces and continuous functions

Partially ordered sets and order-preserving functions

There is an all-inclusive theory: sets (with no additional structure) and arbitrary functions between them.

In order to really understand a mathematical theory such as group theory, one must understand not only the objects (groups) but also the structure-preserving functions between them (group-homomorphisms.)

[18] [ML98, p. 29-30].

II.11.2. Categories

II.11.2.1. A *category* $\mathcal{C}$ consists of:

A collection $\mathrm{Ob}(\mathcal{C})$ of *objects*.

For each pair (A, B) of objects, a collection $\mathrm{Hom}(A, B)$ (or $\mathrm{Hom}_{\mathcal{C}}(A, B)$) of *morphisms*.

For all objects A, B, C, a rule for composing morphisms, i.e. a function from $\mathrm{Hom}(B, C) \times \mathrm{Hom}(A, B)$ to $\mathrm{Hom}(A, C)$, written $(g, f) \mapsto g \circ f$. Composition must be associative, i.e. for any objects A, B, C, D, and any $f \in \mathrm{Hom}(A, B)$, $g \in \mathrm{Hom}(B, C)$, $h \in \mathrm{Hom}(C, D)$, we must have $h \circ (g \circ f) = (h \circ g) \circ f$.

For any object A, an *identity morphism* i_A satisfying $f \circ i_A = f$ for all B and all $f \in \mathrm{Hom}(A, B)$, and $i_A \circ g = g$ for all B and all $g \in \mathrm{Hom}(B, A)$.

II.11.2.2. These requirements are reasonable in light of the examples to be modeled, and are the minimum amount of structure needed for a satisfactory theory. In practically all the applications we will consider in this book, the objects will be sets and the morphisms certain (*allowable*) functions between them, with "composition" the usual composition of functions. The requirements are then that a composition of allowable functions in the theory is again an allowable function, and that the identity function from any set in the theory to itself is always an allowable function. Associativity of composition is automatic in this case. Such a category is called a *category of sets and functions*.

II.11.2.3. The examples described earlier give standard categories:

Set: Sets and functions (i.e. the objects of **Set** are all sets, and $\mathrm{Hom}(X, Y)$ is the set of all functions from X to Y)

$\mathbf{Vect}_{\mathbb{R}}$: Real vector spaces and linear transformations

Gp: Groups and group homomorphisms

Ab: Abelian groups and group homomorphisms

MSp: Metric spaces and continuous functions

Top: Topological spaces and continuous functions

POSet: Partially ordered sets and order-preserving functions

II.11.2.4. There is, however, no requirement that the morphisms of a category be functions. Some interesting categories in which the morphisms are not functions are given in the exercises.

II.11.2.5. When categories are defined within set theory, the collection of objects need not be a set (although it is a class, i.e. the objects themselves are sets); but $\mathrm{Hom}(A, B)$ must be a set for every A, B. A category $\mathcal{C}$ in which $\mathrm{Ob}(\mathcal{C})$ is a set is called a *small category*. None of the examples above are small categories; but one could construct small categories from them by limiting the size of the objects by restricting the underlying set to be a subset of some large fixed "universal" set. (Actually the interpretation of category theory within set theory depends on what set theory is used. The above interpretation, the one commonly used by mathematicians, is for NBG set theory, which includes proper classes; the interpretation for ZF set theory is more complicated and subtle, cf. II.10.5.8..)

II.11.2.6. If $\mathcal{C}$ and $\mathcal{D}$ are categories, then $\mathcal{C}$ is a *subcategory* of $\mathcal{D}$ if

(i) $\mathrm{Ob}(\mathcal{C}) \subseteq \mathrm{Ob}(\mathcal{D})$.

(ii) For all $A, B \in \mathrm{Ob}(\mathcal{C})$, $\mathrm{Hom}_{\mathcal{C}}(A, B) \subseteq \mathrm{Hom}_{\mathcal{D}}(A, B)$.

(iii) Composition of morphisms in $\mathcal{C}$ is the same as composition in $\mathcal{D}$.

(iv) if $A \in \mathrm{Ob}(\mathcal{C})$, the identity morphism ι_A in $\mathrm{Hom}_{\mathcal{D}}(A, A)$ is in $\mathrm{Hom}_{\mathcal{C}}(A, A)$ (hence is the identity morphism in $\mathcal{C}$).

A subcategory is a "subset" which is itself a category under the same operations. $\mathcal{C}$ is a *full subcategory* of $\mathcal{D}$ if $\mathrm{Hom}_{\mathcal{C}}(A, B) = \mathrm{Hom}_{\mathcal{D}}(A, B)$ for all $A, B \in \mathrm{Ob}(\mathcal{C})$. The category **Ab** is a full subcategory of **Gp**. The category of metric spaces (or separable metric spaces) and uniformly continuous functions is a subcategory of **MSp** which is not full. Note that a category like **Gp** is technically not a subcategory of **Set** since different groups can have the same underlying set.

A full subcategory $\mathcal{C}$ of a category $\mathcal{D}$ is completely characterized by the subclass $\mathrm{Ob}(\mathcal{C})$ of $\mathrm{Ob}(\mathcal{D})$, and any subclass S of $\mathrm{Ob}(\mathcal{D})$ gives a (unique) full subcategory $\mathcal{C}$ with $\mathrm{Ob}(\mathcal{C}) = S$.

II.11.2.7. Properties of injective and surjective functions can be abstracted to general categories. If $\mathcal{C}$ is a category and $A, B \in \mathrm{Ob}(C)$, and $f \in \mathrm{Hom}(A, B)$, then

> The morphism f is a *monomorphism* if, for every $C \in \mathrm{Ob}(\mathcal{C})$ and $g, h \in \mathrm{Hom}(C, A)$, $f \circ g = f \circ h$ implies $g = h$ (i.e. f can be "canceled" on the left).
>
> The morphism f is a *epimorphism* if, for every $C \in \mathrm{Ob}(\mathcal{C})$ and $g, h \in \mathrm{Hom}(B, C)$, $g \circ f = h \circ f$ implies $g = h$ (i.e. f can be "canceled" on the right).

The adjectives *monic* and *epi* [or *epic* (!)] are used for monomorphisms and epimorphisms.

II.11.2.8. It is easy to check that in the category **Set**, a morphism (function) is a monomorphism [resp. an epimorphism] if and only if it is injective [resp. surjective]. However, this is not true in general categories, even categories of sets and functions: for example, in the category of Hausdorff topological spaces and continuous functions, a morphism (continuous function) is an epimorphism if and only if its range is dense.

II.11.2.9. If $\mathcal{C}$ is a category, $A, B \in \mathrm{Ob}(\mathcal{C})$, a morphism $f \in \mathrm{Hom}(A, B)$ is an *isomorphism* if there is a $g \in \mathrm{Hom}(B, A)$ such that $g \circ f = \iota_A$ and $f \circ g = \iota_B$. Such g is necessarily unique; an isomorphism is both a monomorphism and an epimorphism. In many categories such as **Set**, **Gp**, and $\mathbf{Vect}_{\mathbb{R}}$, a morphism is an isomorphism if and only if it is both a monomorphism and an epimorphism, but this is not true in general categories: for example, in the category of Hausdorff topological spaces and continuous functions, an injective continuous function with dense range is both a monomorphism and an epimorphism, but not an isomorphism in general, even if it is actually surjective (since the inverse function is not necessarily continuous).

II.11.2.10. Other set-theoretic constructions such as products, coproducts, limits and colimits (inductive and projective limits), pullbacks, etc., can be nicely phrased in categorical (arrow-theoretic) terms with interpretations as universal objects in general categories.

One very important case is where the Hom-sets have a structure as abelian groups (plus some other technical properties). Such a category is called an *abelian category*. The categories **Ab** and $\mathbf{Vect}_{\mathbb{R}}$ are abelian categories. See [Fre64] for more on abelian categories.

II.11.3. Functors

II.11.3.1. If $\mathcal{C}$ and $\mathcal{D}$ are categories, a *functor* is a "function" F from $\mathcal{C}$ to $\mathcal{D}$ consisting of

A function $F : \mathrm{Ob}(\mathcal{C}) \to \mathrm{Ob}(\mathcal{D})$.

For each pair A, B of objects in $\mathcal{C}$, a function, also denoted F (or sometimes $F_{A,B}$), from $\mathrm{Hom}_{\mathcal{C}}(A, B)$ to $\mathrm{Hom}_{\mathcal{D}}(F(A), F(B))$ such that

(i) If $A, B, C \in \mathrm{Ob}(\mathcal{C})$ and $f \in \mathrm{Hom}_{\mathcal{C}}(A, B)$, $g \in \mathrm{Hom}_{\mathcal{C}}(B, C)$, then $F(g \circ f) = F(g) \circ F(f)$.

(ii) If $A \in \mathcal{C}$, then $F(\iota_A) = \iota_{F(A)}$.

II.11.3.2. Examples. (i) A simple example is the "underlying topology" functor from **MSp** to **Top**: if (X, ρ) is a metric space, set $F(X, \rho) = (X, \mathcal{T}_\rho)$ and $F(f) = f$ if $f : X \to Y$ is a continuous function from (X, ρ) to (Y, σ).

(ii) If $\mathcal{C}$ is a category of sets and functions, there is a "forgetful functor" F from $\mathcal{C}$ to **Set**, with $F(A) = A$ (just regarded as a set) for $A \in \mathrm{Ob}(\mathcal{C})$ and $F(f) = f$ (just regarded as a function) for any morphism f. There can also be forgetful functors from a category with more structure to one with less: for example, there is a forgetful functor from $\mathbf{Vect}_{\mathbb{R}}$ to **Ab** which assigns to each real vector space its underlying additive group, "forgetting" the scalar multiplication. (Note that the functor in (i) is essentially a forgetful functor.)

(iii) An inclusion of a subcategory, such as the inclusion of **Ab** into **Gp**, is a functor.

(iv) The "power set" construction is a functor from **Set** to **Set**. If X is a set, set $F(X) = \mathcal{P}(X)$. If $f : X \to Y$ is a function, define $F(f) : \mathcal{P}(X) \to \mathcal{P}(Y)$ by $[F(f)](A) = f(A)$ for $A \subseteq X$.

(v) Here are some examples from group theory. The commutator subgroup construction $G \mapsto [G, G]$ defines a functor F from **Gp** to **Gp** since a group-homomorphism $f : G \to H$ maps $[G, G]$ into $[H, H]$, so we can define $F(G) = [G, G]$ and $F(f) = f|_{[G,G]}$. Similarly, the abelianization construction $G \mapsto G/[G, G]$ gives a functor from **Gp** to **Ab** since a group-homomorphism $f : G \to H$ induces a homomorphism from $G/[G, G]$ to $H/[H, H]$ in a natural way. But the center construction $G \mapsto Z(G)$ (where $Z(G)$ is the center of G) does not define a functor from **Gp** to **Ab** (i.e. is not "functorial") since a group-homomorphism $f : G \to H$ does not necessarily map $Z(G)$ into $Z(H)$, so does not induce a homomorphism from $Z(G)$ to $Z(H)$.

Some more interesting examples are given in the exercises.

II.11.3.3. A functor in the sense of II.11.3.1. is also called a *covariant functor* since it preserves the "direction" of arrows. One can also define contravariant functors which reverse the direction of arrows: a *contravariant functor* F from a category $\mathcal{C}$ to a category $\mathcal{D}$ consists of

A function $F : \mathrm{Ob}(\mathcal{C}) \to \mathrm{Ob}(\mathcal{D})$.

For each pair A, B of objects in $\mathcal{C}$, a function, also denoted F (or sometimes $F_{A,B}$), from $\mathrm{Hom}_{\mathcal{C}}(A, B)$ to $\mathrm{Hom}_{\mathcal{D}}(F(B), F(A))$ such that

(i) If $A, B, C \in \mathrm{Ob}(\mathcal{C})$ and $f \in \mathrm{Hom}_{\mathcal{C}}(A, B)$, $g \in \mathrm{Hom}_{\mathcal{C}}(B, C)$, then $F(g \circ f) = F(f) \circ F(g)$.

(ii) If $A \in \mathcal{C}$, then $F(\iota_A) = \iota_{F(A)}$.

II.11.3.4. EXAMPLES. (i) Here is a simple but very important example of a contravariant functor (there are many slight variations). Let $\mathcal{C}$ be any category, and C a fixed object of $\mathcal{C}$. Define a functor F from $\mathcal{C}$ to **Set** by setting $F(A) = \mathrm{Hom}(A, C)$ for $A \in \mathrm{Ob}(\mathcal{C})$, and if $A, B \in \mathrm{Ob}(\mathcal{C})$ and $f \in \mathrm{Hom}(A, B)$, let $F(f)$ be the function $[F(f)](g) = g \circ f$ if $g \in \mathrm{Hom}(B, C) = F(B)$. Then $F(f)$ is a function from $\mathrm{Hom}(B, C) = F(B)$ to $\mathrm{Hom}(A, C) = F(A)$, i.e. an element of $\mathrm{Hom}(F(B), F(A))$ in **Set**. The dual space construction on $\mathbf{Vect}_{\mathbb{R}}$ () and the Pontryagin dual construction () are special cases.

Similarly, putting C in the first argument gives a covariant functor: set $F(A) = \mathrm{Hom}(C, A)$ for $A \in \mathrm{Ob}(\mathcal{C})$, and if $A, B \in \mathrm{Ob}(\mathcal{C})$ and $f \in \mathrm{Hom}(A, B)$, let $F(f)$ be the function $[F(f)](g) = f \circ g$ if $g \in \mathrm{Hom}(C, A)$.

(ii) The power set construction also gives a contravariant functor from **Set** to **Set**. If X is a set, set $F(X) = \mathcal{P}(X)$, and if $f : X \to Y$ is a function, define $F(f) : \mathcal{P}(Y) \to \mathcal{P}(X)$ by $[F(f)](B) = f^{-1}(B)$ for $B \subseteq Y$.

II.11.3.5. Historically, the first important example of a contravariant functor (although the term has been used here only recently and it is not usually thought of quite this way) is the Galois group construction in field theory. Let $\mathcal{E}$ be the collection of all pairs (K, F), where F is a field and K is a Galois extension (finite normal separable extension) of F. A morphism from (K, F) to (K', F') is an embedding $\alpha : K \to K'$ with $\alpha(F) \subseteq F'$. Then $\mathcal{E}$ is a category. The assignment $(K, F) \mapsto Gal(K/F)$ is a contravariant functor from $\mathcal{E}$ to the category of finite groups: if $\alpha : (K, F) \to (K', F')$ and $\sigma \in Gal(K'/F')$ is an automorphism of K' over F', then σ must be the identity on $\alpha(F)$ and map $\alpha(K)$ to itself, hence gives an element of $Gal(K/F)$, and this assignment clearly respects composition and identity maps (in two senses).

II.11.3.6. Roughly speaking, contravariant functors send subobjects to quotients, and vice versa, i.e. interchange monomorphisms and epimorphisms (although this is not exactly true in general).

Contravariant functors can be regarded as ordinary (covariant) functors to or from an opposite category, although this is often unnatural in practice.

II.11.3.7. (Covariant) functors can be composed: if F is a functor from $\mathcal{C}$ to $\mathcal{D}$, and G is a functor from $\mathcal{D}$ to $\mathcal{E}$, then $G \circ F$ (often written GF), defined in the obvious way, is a functor from $\mathcal{C}$ to $\mathcal{E}$. Since functor composition is essentially composition of functions, it is automatically associative.

A covariant functor can also be composed with a contravariant functor in either order (provided the categories match up!) The composition is a contravariant functor. Two contravariant functors can be composed; the composition is a covariant functor.

Each category $\mathcal{C}$ has an "identity functor" denoted $\mathbf{1}_{\mathcal{C}}$, which is just the identity map on objects and morphisms. The identity functor is an identity for functor composition.

II.11.3.8. A (covariant or contravariant) functor F from a category $\mathcal{C}$ to a category $\mathcal{D}$ is *faithful* if each $F_{A,B}$ is injective as a function from $\mathrm{Hom}_{\mathcal{C}}(A, B)$ to $\mathrm{Hom}_{\mathcal{D}}(F(A), F(B))$ ($\mathrm{Hom}_{\mathcal{D}}(F(B), F(A))$ in the contravariant case). The functor F is *full* if each $F_{A,B}$ is surjective. The inclusion of a subcategory is faithful, and the inclusion of a full subcategory is full. The functor H from **Top** to **Toph** (II.11.9.11.) is full but not faithful. A functor which is full and faithful, and bijective on objects as well as Hom-sets, is a *category isomorphism*. Isomorphic categories are the "same" abstractly. See II.11.5.6. for a more flexible notion of equivalence of categories.

II.11.4. Natural Transformations

Some constructions in mathematics are "natural" and some are not. Roughly speaking, a construction is "natural" if it is always done by the same systematic procedure regardless of the specific situation, and "unnatural" if the construction relies on ad-hoc procedures or choices dependent on the situation.

The notion of "natural transformation" in category theory is designed to try to codify and make careful sense of this vague but important notion. The philosophy of EILENBERG and MACLANE, who invented category theory, was that "Categories were invented to define functors, and functors were invented to define natural transformations."

II.11.4.1. DEFINITION. Let $\mathcal{C}$ and $\mathcal{D}$ be categories, and F, G functors from $\mathcal{C}$ to $\mathcal{D}$. A *natural transformation* from F to G is a function τ from $\mathrm{Ob}(\mathcal{C})$ to the morphisms of $\mathcal{D}$ with $\tau_A = \tau(A) \in \mathrm{Hom}_{\mathcal{D}}(F(A), G(A))$ for each $A \in \mathrm{Ob}(\mathcal{C})$, such that, for all $A, B \in \mathrm{Ob}(\mathcal{C})$ and all $f \in \mathrm{Hom}_{\mathcal{C}}(A, B)$, er have

$$\tau_B \circ F(f) = G(f) \circ \tau_A \ .$$

The morphisms $\{\tau_A : A \in \mathrm{Ob}(\mathcal{C})\}$ are called the *components* of τ, and we write $\tau : F \to G$.

If morphisms are pictured as arrows, the condition is that for all A, B, f, the diagram

$$\begin{array}{ccc} F(A) & \xrightarrow{\tau_A} & G(A) \\ {\scriptstyle F(f)}\downarrow & & \downarrow{\scriptstyle G(f)} \\ F(B) & \xrightarrow{\tau_B} & G(B) \end{array}$$

commutes.

As explained in [ML98], a functor from $\mathcal{C}$ to $\mathcal{D}$ gives a "picture" of $\mathcal{C}$ in $\mathcal{D}$, and a natural transformation τ from F to G gives a "translation" of the F-picture of $\mathcal{C}$ to the G-picture of $\mathcal{C}$ in $\mathcal{D}$. In particular, any commutative diagram in $\mathcal{C}$ is "pictured" as a commutative diagram in $\mathcal{D}$ by both F and G, and τ gives a larger commutative diagram whose "ends" are the F- and G-pictures of the original diagram.

II.11.4.2. EXAMPLES. Here are two good examples from [ML98]:

(i) Let **CRng** be the category of unital commutative rings and unital ring-homomorphisms. Fix $n \in \mathbb{N}$. If $R \in \mathrm{Ob}(\mathbf{CRng})$, let $GL_n(R)$ be the set of invertible $n \times n$ matrices with entries in R. Then $GL_n(R)$ is a group under matrix multiplication. If $R, S \in \mathrm{Ob}(\mathbf{CRng})$ and $f \in \mathrm{Hom}(R, S)$, i.e. f is a unital ring-homomorphism, then f applied entrywise gives a group-homomorphism from $GL_n(R)$ to $GL_n(S)$ (this is because matrix multiplication is defined by the "same formula" over every ring). Thus GL_n may be viewed as a functor F from **CRng** to **Gp**, with $F(R) = GL_n(R)$ for each R and $F(f)$ the entrywise group-homomorphism. We have $GL_1(R) = R^\times$, the group of invertible elements of R.

For any $R \in \mathrm{Ob}(\mathbf{CRng})$, the determinant Det_R gives a group-homomorphism from $GL_n(R)$ to $R^\times$, i.e. an element of $\mathrm{Hom}_{\mathbf{Gp}}(GL_n(R), GL_1(R))$. If we define $\tau_R = Det_R$ for each R, then τ is a natural transformation from GL_n to GL_1 since if $f : R \to S$ is a unital ring-homomorphism and $\tilde{f}$ the corresponding group-homomorphism from $GL_n(R)$ to $GL_n(S)$, we have, for any $a \in GL_n(R)$, $Det_S(\tilde{f}(a)) = f(Det_R(a))$ ("the determinant is given by the same formula over any ring"). Thus the determinant is "natural."

(ii) Let F be the abelianization functor from II.11.3.2.(v), i.e. if G is a group, then $F(G) = G/[G, G]$, and if $f : G \to H$ is a group-homomorphism, then $F(f) = \bar{f}$, where $\bar{f}$ is the induced group-homomorphism from

$G/[G,G]$ to $H/[H,H]$. We may regard F as a functor from **Gp** to **Gp** since **Ab** is a subcategory of **Gp**. Let I be the identity functor from **Gp** to **Gp**. Define τ by letting τ_G be the quotient homomorphism from G to $G/[G,G]$, i.e. $\tau_G \in \mathrm{Hom}(I(G), F(G))$. Then τ is a natural transformation from I to F since for any G, H, and $f : G \to H$, we have a commutative diagram

$$\begin{array}{ccc} G & \xrightarrow{\tau_G} & G/[G,G] \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle \bar{f}} \\ H & \xrightarrow{\tau_H} & H/[H,H] \end{array}$$

so abelianization is "natural" (this example has many variations).

II.11.4.3. Natural transformations can be "composed": if F, G, H are functors from $\mathcal{C}$ to $\mathcal{D}$, τ is a natural transformation from F to G, and σ a natural transformation from G to H, then $\sigma \circ \tau$, defined by $(\sigma \circ \tau)_A = \sigma_A \circ \tau_A$ for each $A \in \mathrm{Ob}(\mathcal{C})$, is a natural transformation from F to H. Composition is associative. Each functor F has an "identity natural transformation" from F to F, define in the obvious way, which is an identity for composition.

II.11.4.4. Definition. Let F and G be functors from a category $\mathcal{C}$ to a category $\mathcal{D}$. A *natural isomorphism* (or *natural equivalence*) from F to G is a natural transformation $\tau : F \to G$ such that $\tau_A \in \mathrm{Hom}_{\mathcal{D}}(F(A), G(A))$ is an isomorphism for every $A \in \mathrm{Ob}(\mathcal{C})$. Write $F \cong G$ if there is a natural isomorphism from F to G.

If τ is a natural isomorphism from F to G, then τ^{-1}, defined by $(\tau^{-1})_A = (\tau_A)^{-1}$ for each $A \in \mathrm{Ob}(\mathcal{C})$, is a natural isomorphism from G to F. A composition of natural isomorphisms is a natural isomorphism, and the identity natural transformation of a functor is a natural isomorphism. In particular, natural isomorphism is an equivalence relation on the functors from $\mathcal{C}$ to $\mathcal{D}$.

II.11.4.5. Natural transformations and natural isomorphisms can be defined also for contravariant functors, or between a covariant and a contravariant functor; details are left to the reader.

II.11.5. Equivalence of Categories and Skeletons

II.11.5.1. A category such as **Gp** is in some respects "too large": we usually want to think of two groups a being "the same" abstract group if they are isomorphic, i.e. we want to consider a category whose objects are "abstract groups," i.e. isomorphism classes of groups.

It is a little tricky to decide how to do this, and in particular what the morphisms in such a category should be, though, for many reasons, such as:

(i) A group can have two distinct subgroups which are isomorphic. Should we regard these subgroups as the "same" or "different"?

(ii) A group can have automorphisms other than the identity, and the study of automorphisms is an important part of group theory.

At least these two considerations can be handled by taking $\mathrm{Hom}([G],[H])$ (where $[G]$ is the isomorphism class of G) to be a full copy of $\mathrm{Hom}_{\mathbf{Gp}}(G,H)$.

II.11.5.2. DEFINITION. A category $\mathcal{S}$ is *skeletal* if, whenever $A, B \in \mathrm{Ob}(\mathcal{S})$ and $f \in \mathrm{Hom}_{\mathcal{S}}(A, B)$ is an isomorphism, it follows that $A = B$ (i.e. the isomorphism classes in $\mathcal{S}$ are singletons).

A *skeleton* of a category $\mathcal{C}$ is a full skeletal subcategory $\mathcal{S}$ such that every object of $\mathcal{C}$ is isomorphic (in $\mathcal{C}$) to a (necessarily unique) object in $\mathcal{C}$.

A skeleton for $\mathcal{C}$ thus consists of a "model object" for each object of $\mathcal{C}$, such that morphisms between objects are fully modeled by the morphisms between the corresponding model objects.

II.11.5.3. EXAMPLE. Here is a simple example. Let **Fin** be the category of finite sets and functions. Set $\mathbf{0} = \emptyset$ and inductively for each $n \in \mathbb{N}$ set $\boldsymbol{n} = \{\mathbf{0}, \mathbf{1}, \ldots, \boldsymbol{n}\text{−}\mathbf{1}\}$. The sets $\{\boldsymbol{n} : n \in \mathbb{N} \cup \{0\}\}$ are called *finite ordinals* (it is customary in set theory to think of the standard n-element set as $\{0, \ldots, n-1\}$ instead of $\{1, \ldots, n\}$). Let **Ford** be the full subcategory of **Fin** whose objects are the finite ordinals. Then **Ford** is a skeleton of **Fin**. (Note that isomorphisms in **Set** are just bijections.) An isomorphism from a finite set A to a finite ordinal is just a "numbering" of the elements of A. A function f from A to B can be described as a function between the corresponding finite ordinals once numberings of the elements of A and B are fixed: f sends the first (actually 0'th) element of A to the n_0'th element of B, etc.

II.11.5.4. There is one part of the structure of finite sets not preserved in the skeleton: containment. If A, B are finite sets, we have fixed numberings of the elements of A and B. But if $A \subseteq B$, the numberings we have chosen do not necessarily agree on the elements of A. If we are just dealing with these two sets, we can of course make the numberings compatible by first numbering the elements of A and number the elements of $B \setminus A$ after that; but we cannot choose compatible numberings in any greater generality. However, containment is not really part of the category **Fin** (thus the category **Fin**, or **Set**, does not really capture the full structure of set theory); subobjects correspond (imperfectly) to monomorphisms, which are preserved in passing to the skeleton.

II.11.5.5. DEFINITION. Let $\mathcal{C}$ and $\mathcal{D}$ be categories. A *weak equivalence* from $\mathcal{C}$ to $\mathcal{D}$ is a functor F from $\mathcal{C}$ to $\mathcal{D}$ which is both faithful and full, with the property that every $D \in \mathrm{Ob}(\mathcal{D})$ is isomorphic to $F(A)$ for some $A \in \mathrm{Ob}(\mathcal{C})$.

The inclusion of a skeleton is an example, perhaps the most important example.

It is not obvious that weak equivalence is an equivalence relation on categories, despite the name (it is obviously reflexive and transitive, but not obviously symmetric.) But a weak equivalence between skeletal categories is a category isomorphism. There is a stronger notion of category equivalence which is more obviously a true equivalence:

II.11.5.6. DEFINITION. Categories $\mathcal{C}$ and $\mathcal{D}$ are *strongly equivalent* if there are functors F from $\mathcal{C}$ to $\mathcal{D}$ and G from $\mathcal{D}$ to $\mathcal{C}$ such that $G \circ F \cong \mathbf{1}_{\mathcal{C}}$ and $F \circ G \cong \mathbf{1}_{\mathcal{D}}$.

II.11.5.7. PROPOSITION. If $\mathcal{C}$ and $\mathcal{D}$ are categories and F and G are functors implementing a strong equivalence as in II.11.5.6., then F and G are weak equivalences.

II.11.5.8. What about the converse? In particular, if $\mathcal{S}$ is a skeleton for $\mathcal{C}$ and F is the inclusion, is there a functor G from $\mathcal{C}$ to $\mathcal{S}$ implementing a strong equivalence with F? The way to do this is to choose, for each $A \in \mathrm{Ob}(\mathcal{C})$, the object $G(A)$ of $\mathcal{S}$ it is isomorphic to. But this is not sufficient to define a functor: we must define $G(f)$ for each morphism in $\mathcal{C}$ in such a way that composition and identities are preserved. Preserving identities can obviously be done, but preserving composition is tricky. If for each $A \in \mathrm{Ob}(\mathcal{C})$ we fix an isomorphism $\theta_A \in \mathrm{Hom}(A, G(A))$, we can define $G(f) = \theta_B \circ f \circ \theta_A^{-1}$ for $f \in \mathrm{Hom}(A, B)$ and composition (and identities) will be preserved. The subtlety is that the Axiom of Global Choice (II.6.3.18.) is needed to choose a θ_A for each A. Thus, if we assume the AGC, a category is strongly equivalent to any skeleton.

Another important consequence of the AGC:

II.11.5.9. Proposition. (AGC) Every category has a skeleton. Any two skeletons for a category are isomorphic.

Proof: Let $\mathcal{C}$ be a category. The class $\mathrm{Ob}(\mathcal{C})$ partitions into isomorphism classes. Using AGC, there is a subclass of $\mathrm{Ob}(\mathcal{C})$ consisting of exactly one element from each equivalence class. The full subcategory corresponding to this class is a skeleton of $\mathcal{C}$. If $\mathcal{S}$ and $\mathcal{S}'$ are skeletons of $\mathcal{C}$, then for each $A \in \mathrm{Ob}(\mathcal{S})$ there is a unique $A' \in \mathrm{Ob}(\mathcal{S}')$ which is isomorphic to A. Fix an isomorphism $\phi_A \in \mathrm{Hom}_{\mathcal{C}}(A, A')$ for each $A \in \mathrm{Ob}(\mathcal{S})$ (the AGC is needed to choose these), and set $F(A) = A'$ and $F(f) = \phi_B \circ f \circ \phi_A^{-1}$ for $A, B \in \mathrm{Ob}(\mathcal{S})$ and $f \in \mathrm{Hom}(A, B)$. It is easily checked that F is an isomorphism from $\mathcal{S}$ to $\mathcal{S}'$. 👍

II.11.5.10. Proposition. (AGC) Let $\mathcal{C}$ and $\mathcal{D}$ be categories. Then $\mathcal{C}$ and $\mathcal{D}$ are weakly equivalent if and only if they are strongly equivalent.

Proof: One direction is II.11.5.7.. Conversely, suppose $\mathcal{C}$ and $\mathcal{D}$ are weakly equivalent. Let $\mathcal{S}$ and $\mathcal{S}'$ be skeletons of $\mathcal{C}$ and $\mathcal{D}$ respectively. Then there are strong, hence weak, equivalences from $\mathcal{S}$ to $\mathcal{C}$ and from $\mathcal{D}$ to $\mathcal{S}'$. Composing with a weak equivalence from $\mathcal{C}$ to $\mathcal{D}$, we get a weak equivalence from $\mathcal{S}$ to $\mathcal{S}'$. But $\mathcal{S}$ and $\mathcal{S}'$ are skeletal, so this weak equivalence is a category isomorphism, hence a strong equivalence. Thus we get strong equivalences from $\mathcal{D}$ to $\mathcal{S}'$ to $\mathcal{S}$ to $\mathcal{C}$. 👍

II.11.6. Adjoint Functors

II.11.7. The Category of Small Categories

II.11.7.1. Functors may be regarded as "morphisms" between categories. We may thus try to go to one further level of abstraction by regarding the collection of all categories as one supercategory, whose objects are categories and for which $\mathrm{Hom}(\mathcal{C}, \mathcal{D})$ is the collection of (covariant) functors from $\mathcal{C}$ to $\mathcal{D}$.

II.11.7.2. When working within set theory, this construction almost works, but not quite. We run into a version of Russell's Paradox if we try to take all categories as objects, and we also violate the condition that $\mathrm{Hom}(\mathcal{C}, \mathcal{D})$ be a set for any $\mathcal{C}$ and $\mathcal{D}$. But these logical difficulties disappear if we take as our objects only *small* categories. Thus we may define a category **SCat** whose objects are all small categories, and for which $\mathrm{Hom}(\mathcal{C}, \mathcal{D})$ is the set(!) of all functors from $\mathcal{C}$ to $\mathcal{D}$.

Higher Categories

II.11.7.3. One can instead think of the collection of all categories as a "second-order category." There is more structure than just a category: the objects are collections of objects with morphisms between them, and there are morphisms (functors) between the second-level objects preserving the first-order structure. One can formalize the notion of a second-order category.

II.11.7.4. Why stop there? The second-order category of all categories has morphisms between functors: natural transformations. One can define third-order categories by abstracting this. The process can be repeated indefinitely, apparently even transfinitely.

II.11.8. Categories as a Foundation of Mathematics

II.11.8.1. Categories were first defined by Eilenberg and MacLane in the early 1940s ([EM42a], [EM42b], [EM45]) to provide a setting for studying natural transformations and formulating parts of algebraic topology and, later, homological algebra. They began to be more than just a language with Grothendieck's 1957 paper [Gro57] recasting algebraic geometry in category terms. Lawvere in his 1963 Ph.D. thesis began an attempt to reformulate the foundations of mathematics in terms of categories rather than sets. See [Fre64] for an incomplete but illuminating discussion of the history of category theory up to 1964.

II.11.8.2. Lawvere's idea was that instead of focusing on *membership*, which is the basic concept of set theory, one should focus on *structure* as encoded in the transformations between objects. This represents a "top-down" approach to the foundations of mathematics where it is not relevant what the objects of study (e.g. numbers) "are", but rather what properties they have, rather than the "bottom-up" approach of set theory, which attempts to build up all the objects of mathematics from fundamental building blocks. This is arguably more in line with the way most mathematicians who are not logicians or set-theorists approach mathematics, and it can even be interpreted as being in line with logic and set theory, e.g. we don't normally worry about what sets *are*, but only about how to work with them structurally. A great deal of mainstream mathematics, if not all, can (apparently) be done from this point of view without ever involving set theory.

II.11.8.3. The notion of a *topos* (plural often written *topoi* as if it were a Greek word even though it isn't), invented by Grothendieck, can be used in a foundational way. A topos is a category with extra structure (limits, colimits, products, quotients, etc.) which is similar enough to the category of sets and functions that it can be used in a similar way to define and describe all the standard structures of mathematics. They can even encode alternative logics such as intuitionist logic. (Grothendieck viewed topoi as generalizations of topological spaces in his original work, which is another valid point of view; the fact that topoi so nicely generalize set theory and topology makes them even more appealing as foundational structures.)

II.11.8.4. There has been much work on recasting the foundations of mathematics on category theory rather than set theory. However, this work is incomplete: there are still substantial gaps in what might be called the "foundations of category theory," including avoiding potential paradoxes similar to the paradoxes of naive set theory which caused so much consternation in the early twentieth century. The potential, though, is to give a better foundation to the full spectrum of mathematics as it is actually practiced.

II.11.9. Exercises

II.11.9.5. Let R be a transitive, reflexive relation on a set X. Define a (small) category $\mathcal{C}$ with $\mathrm{Ob}(\mathcal{C}) = X$, and such that $\mathrm{Hom}(x, y)$ has exactly one element if xRy, and $\mathrm{Hom}(x, y) = \emptyset$ if x is not related to y. Composition is defined in the only possible way: if xRy and yRz, and f and g are the unique elements of $\mathrm{Hom}(x, y)$ and $\mathrm{Hom}(y, z)$, then $g \circ f$ is the unique element of $\mathrm{Hom}(x, z)$.
(a) Show that $\mathcal{C}$ is a (small) category.
(b) Let S be another transitive, reflexive relation on a set Y, and $\mathcal{D}$ the corresponding category. Describe all functors from $\mathcal{C}$ to $\mathcal{D}$.

II.11.9.6. Let $\mathcal{S}$ be the category associated with the partially ordered set $\{0, 1\}^2$ with the ordinary ordering (). If $\mathcal{C}$ is any category, a functor from $\mathcal{S}$ to $\mathcal{C}$ is called a *commutative square* in $\mathcal{C}$. Describe commutative squares in words, and draw a picture explaining the terminology (begin with a picture of the category $\mathcal{S}$ by placing the objects at vertices of a square and drawing arrows for the morphisms.)

II.11.9.7. (a) Let G be a group. Define a (small) category $\mathcal{G}$ with one object $\bullet$, with $\mathrm{Hom}(\bullet, \bullet) = G$. Define composition of morphisms via the group operation.

(b) Let H be another group, and $\mathcal{H}$ the corresponding category. Describe all functors from $\mathcal{G}$ to $\mathcal{H}$. Describe all functors from $\mathcal{G}$ to **Set**, $\mathbf{Vect}_{\mathbb{R}}$, and **Top** (these functors give representations of G).

(c) A (necessarily small) category with one object is called a *monoid*. Describe which monoids arise from groups.

(d) A small category in which every morphism is an isomorphism is called a *groupoid*. A group is thus a groupoid with exactly one object. An equivalence relation, regarded as a category as in II.11.9.5., is also a groupoid. A general groupoid is a sort of mixture of the extreme examples of groups and equivalence relations. There is an extensive theory of groupoids, with many applications; see ().

II.11.9.8. Define a category $\mathbf{Mat}_{\mathbb{R}}$ whose objects are $\mathbb{N}$, and for which the morphisms $\mathrm{Hom}(n, m)$ are the $m \times n$ matrices with real entries. Composition is matrix product. Show that $\mathbf{Mat}_{\mathbb{R}}$ is a (small) category.

II.11.9.9. Let **Rel** have as objects all sets, and $\mathrm{Hom}(X, Y)$ the set of relations from X to Y, subsets of $X \times Y$ (i.e. $\mathrm{Hom}(X, Y) = \mathcal{P}(X \times Y)$. Composition of relations is as in II.3.1.5.. Show that **Rel** is a category, and that **Set** is a subcategory (since a function as defined in II.3.2.1. is a relation).

II.11.9.10. Here is a type of algebraic object which is important in algebraic topology and homological algebra. A *chain complex* is a diagram of the form

$$\cdots \xrightarrow{d_{n+2}} C_{n+1} \xrightarrow{d_{n+1}} C_n \xrightarrow{d_n} C_{n-1} \xrightarrow{d_{n-1}} \cdots$$

where the C_n are abelian groups and $d_n : C_n \to C_{n-1}$ is a group homomorphism, which satisfies $d_n \circ d_{n+1} = 0$ for all n, i.e. the image B_n of d_{n+1} is contained in the kernel Z_n of d_n for each n. Often, but not always, $C_n = \{0\}$ for $n < 0$. The chain complex can formally be defined as a pair $C = \{(C_n), (d_n)\}$ consisting of the sequence (C_n) together with the sequence (d_n) of maps.

The elements of C_n are called *n-chains*; the elements of Z_n are called *n-cycles*, and the elements of B_n *n-boundaries* (these terms are motivated by standard constructions of chain complexes in topology). The quotient group Z_n/B_n is called the *n'th homology group of C*, denoted $H_n(C)$.

(a) Let $C = \{(C_n), (d_n)\}$ and $D = \{(D_n), (d'_n)\}$ be chain complexes. A *morphism* from C to D is a sequence $f = (f_n)$, where $f_n : C_n \to D_n$ is a group homomorphism making the squares in the following diagram commutative:

$$\begin{array}{ccccccccc}
\cdots & \xrightarrow{d_{n+2}} & C_{n+1} & \xrightarrow{d_{n+1}} & C_n & \xrightarrow{d_n} & C_{n-1} & \xrightarrow{d_{n-1}} & \cdots \\
 & & \downarrow{\scriptstyle f_{n+1}} & & \downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle f_{n-1}} & & \\
\cdots & \xrightarrow{d'_{n+2}} & D_{n+1} & \xrightarrow{d'_{n+1}} & D_n & \xrightarrow{d'_n} & D_{n-1} & \xrightarrow{d'_{n-1}} & \cdots
\end{array}$$

i.e. $d'_n \circ f_n = f_{n-1} \circ d_n$ for each n. Morphisms are composed in the obvious way by setting $(g \circ f)_n = g_n \circ f_n$ (check that this makes sense). Show that the chain complexes with these morphisms form a category **Chc**.

(b) Show that, for fixed n, H_n is a functor from **Chc** to **Ab**. [Here H_n sends the chain complex C to the group $H_n(C)$. If $f \in \mathrm{Hom}(C, D)$, show that there is a natural group homomorphism $H_n(f) : H_n(C) \to H_n(D)$ with the right properties.]

There is also a theory of *cochain complexes*, where the horizontal arrows are reversed and the subscripts written as superscripts. In an abstract algebraic sense, there is only a notational difference between chain complexes and cochain complexes, but they are used in a somewhat different and complementary way in topology. Chain and cochain complexes can be defined also over an arbitrary ring R (as diagrams of left R-modules). The algebra involved in the study of chain complexes and related matters is quite extensive. See books on homological algebra such as [], or on algebraic topology such as [], for a full treatment.

II.11.9.11. Here is an important category from topology which is not obviously a category of sets and functions. If X and Y are topological spaces and f_0, f_1 are continuous functions from X to Y, then f_0 and f_1 are *homotopic*, written $f_0 \simeq f_1$, if there is a continuous function $h : [0,1] \times X \to Y$ with $f_0(x) = h(0, x)$ and $f_1(x) = h(1, x)$ for all $x \in X$ ().

(a) Show that for any topological spaces X and Y, $\simeq$ is an equivalence relation on the set of continuous functions from X to Y. [Transitivity takes a little trick.] Write $[f]$ for the homotopy equivalence class of f.

(b) Show that if X, Y, Z are topological spaces and $f_0, f_1 : X \to Y$ and $g_0, g_1 : Y \to Z$ are continuous functions with $f_0 \simeq f_1$ and $g_0 \simeq g_1$, then $g_0 \circ f_0 \simeq g_1 \circ f_1$ (). Thus composition of homotopy equivalence classes of continuous functions is well defined. Show it is associative and respects identity maps.

(c) Let **Toph** have as objects all topological spaces, and for $X, Y \in \mathrm{Ob}(\mathbf{Toph})$, let $\mathrm{Hom}(X, Y)$ be the set of homotopy equivalence classes of continuous functions from X to Y. Compose equivalence classes as in (b). Show that **Toph** is a category with these objects and morphisms.

(d) There is an obvious functor H from **Top** to **Toph** which is the identity on objects and sends a morphism (continuous function) to its homotopy equivalence class.

(e) A functor F from **Top** to a category $\mathcal{D}$ is *homotopy-invariant* if, whenever $X, Y \in \mathrm{Ob}(\mathbf{Top})$ and $f_0, f_1 \in \mathrm{Hom}(X, Y)$ with $f_0 \simeq f_1$, we have $F(f_0) = F(f_1)$. Show that the functor F is homotopy-invariant if and only if there is a functor G from **Toph** to $\mathcal{D}$ such that $F = G \circ H$ (i.e. F "factors through **Toph**"). This is an example of a general "quotient category" construction.

II.11.9.12. [Fre64] Here is an example (a "minimal" example?) of a category which has no embedding into the category of sets and functions.

Take the objects of $\mathcal{C}$ to be a "zero object" O, a "universal object" S, and one object A_α for each ordinal α. There is one "zero morphism" between any two objects, an identity morphism for each object, and, for each pair $\beta < \alpha$ of ordinals three morphisms $f_{\alpha,\beta} : A_\alpha \to S$, $g_{\alpha,\beta} : S \to A_\alpha$, and $h_{\alpha,\beta} : A_\alpha \to A_\alpha$.

Define composition of morphisms as follows. Composition with an identity morphism is as usual. $g_{\alpha,\beta} \circ f_{\alpha,\beta'} = h_{\alpha,\beta''}$, where $\beta'' = \max(\beta, \beta')$. All other compositions are zero morphisms.

(a) Verify that $\mathcal{C}$ is a category (in particular, that composition of morphisms is associative).

(b) Suppose Φ is an embedding of $\mathcal{C}$ into the category $\mathcal{S}$ of sets and functions. Choose an ordinal α such that $|\alpha| > |\mathcal{P}(\Phi(S))|$. Then there must be $\beta < \beta' < \alpha$ such that the functions $\Phi(f_{\alpha,\beta})$ and $\Phi(f_{\alpha,\beta'})$ from $\Phi(A_\alpha)$ to $\Phi(S)$ have the same range R.

(c) Since $g_{\alpha,\beta} \circ f_{\alpha,\beta'} = g_{\alpha,\beta'} \circ f_{\alpha,\beta'}$, the functions $\Phi(g_{\alpha,\beta})$ and $\Phi(g_{\alpha,\beta'})$ must agree on R, and hence

$$\Phi(g_{\alpha,\beta} \circ f_{\alpha,\beta}) = \Phi(g_{\alpha,\beta}) \circ \Phi(f_{\alpha,\beta}) = \Phi(g_{\alpha,\beta'}) \circ \Phi(f_{\alpha,\beta}) = \Phi(g_{\alpha,\beta'} \circ f_{\alpha,\beta})$$

but $g_{\alpha,\beta'} \circ f_{\alpha,\beta} \neq g_{\alpha,\beta} \circ f_{\alpha,\beta}$, contradicting that Φ is an embedding.

(d) Do we really need the zero object O?

The heart of the problem is that if S is interpreted as a set, it must have cardinality greater than the cardinality of all ordinals, which is impossible by the Burali-Forti Paradox ().

II.11.9.13. Categories in Logic. Here is another way a mathematical theory can be thought of as a category. A *(formal) mathematical theory* consists of a language which determines a set of statements, a set of axioms which are statements taken as a starting point, and a logical scheme which allows statements to be deduced from axioms and previous statements. See () for details. A *formal proof* is a finite sequence of statements, each deduced from the axioms and the previous statements of the proof according to the logical scheme. A formal deduction of a statement Q from a statement P is a formal proof with first line P and last line Q. If Q can be formally deduced from P, we write $P \vdash Q$.

A formal deduction is *reduced* if it does not have two identical consecutive lines. A deduction with two or more consecutive identical lines can be *contracted* by replacing the consecutive identical lines with a single line. Any deduction can be contracted to a reduced deduction in a finite number of steps.

(a) If $\mathcal{T}$ is a mathematical theory, form a category $\mathcal{C}_\mathcal{T}$ whose objects are the statements of $\mathcal{T}$, and for which $\mathrm{Hom}(P, Q)$ is the set of reduced formal deductions of Q from P. The "identity deduction" of P from P is the one-line proof P. Morphisms are composed by concatenation and reduction beginning at the point of concatenation. Show that $\mathcal{C}_\mathcal{T}$ is a (small) category (this will include showing that any formal deduction contracts to a unique reduced deduction, and that composition is associative).

(b) Find the isomorphisms in this category. Is it true that P and Q are isomorphic if and only if $P \vdash Q$ and $Q \vdash P$?

II.11.9.14. (a) Show that the Axiom of Choice implies that every small category has a skeleton.

(b) Show that the AC is equivalent to the statement: if X is a set with an equivalence relation, then there is a choice function for the set of equivalence classes.

(c) Let X be a set with an equivalence relation $\sim$. Regard X as a small category as in II.11.9.5.. Show that a skeleton for this category defines a choice function for the set of equivalence classes.

(d) Conclude that the AC is equivalent to the statement that every small category has a skeleton. Show that the Axiom of Global Choice (II.6.3.18.) is equivalent to the statement that every category has a skeleton.

II.11.9.15. (a) Let $\mathcal{C}$ be a category. Show that an isomorphism is both a monomorphism and an epimorphism.

(b) Let $A, B \in \mathrm{Ob}(\mathcal{C})$, and $f \in \mathrm{Hom}(A, B)$. A *right inverse* for f is a $g \in \mathrm{Hom}(B, A)$ with $f \circ g = \iota_B$ (a right inverse for f is often called a *section* for f). Show that a morphism with a right inverse must be an epimorphism. Are right inverses unique in general? Does every epimorphism have a right inverse?

(c) Let $A, B \in \mathrm{Ob}(\mathcal{C})$, and $f \in \mathrm{Hom}(A, B)$. A *left inverse* for f is a $g \in \mathrm{Hom}(B, A)$ with $g \circ f = \iota_A$ (a left inverse for f is often called a *retraction* for f). Show that a morphism with a left inverse must be a monomorphism. Are left inverses unique in general? Does every monomorphism have a left inverse?

(d) Show that if a morphism f has both a left inverse and a right inverse, then the left and right inverses are the same and hence f is an isomorphism.

II.11.9.16. Let $\mathcal{C}$ be a category. An object I of $\mathcal{C}$ is an *initial object* of $\mathcal{C}$ if there is exactly one morphism in $\mathrm{Hom}(I, A)$ for every $A \in \mathrm{Ob}(\mathcal{C})$. An object Φ of $\mathcal{C}$ is a *terminal object* of $\mathcal{C}$ if there is exactly one morphism in $\mathrm{Hom}(A, \Phi)$ for every $A \in \mathrm{Ob}(\mathcal{C})$. (A category need not have any initial or terminal objects.)

(a) Find all initial and terminal objects, if any, in **Set**, **Gp**, $\mathbf{Vect}_{\mathbb{R}}$, and **Top**.

(b) Show that any two initial objects of $\mathcal{C}$ are isomorphic, as are any two terminal objects.

(c) An object O of $\mathcal{C}$ is a *zero object* if it is both an initial object and terminal object. Which of the above categories have zero objects?

(d) Let O be a zero object of $\mathcal{C}$. Show that for any $A, B \in \mathrm{Ob}(\mathcal{C})$ there is a unique morphism in $\mathrm{Hom}(A, B)$ which factors through O. This morphism is called the *zero morphism* from A to B. If $\mathcal{C}$ has more than one zero object, zero morphisms do not depend on what zero object is used in the definition.

II.11.9.17. Here is a "paradox" concerning Gödel's Incompleteness Theorem along the lines of the Berry Paradox (). We work with the following somewhat informal version of Gödel's Theorem:

THEOREM. If $\mathcal{T}$ is a consistent theory which is finitely describable and contains Peano arithmetic, then there is a statement in $\mathcal{T}$ which is neither provable nor disprovable in $\mathcal{T}$.

Fix a model for Peano arithmetic, e.g. the standard model (III.2.2.6.), and let $\mathcal{T}$ be the theory whose language is the language of Peano arithmetic and whose axioms are all the statements in Peano arithmetic which are true in the given model. Then $\mathcal{T}$ is complete since every statement in $\mathcal{T}$ or its negation is an axiom of $\mathcal{T}$; $\mathcal{T}$ is consistent since it has a model; $\mathcal{T}$ obviously contains Peano arithmetic; and $\mathcal{T}$ is finitely describable since we just gave a finite description!

Give a resolution of this "paradox." (Note that the standard model for Peano arithmetic is finitely describable in ZF set theory, which is itself a finitely describable theory in the technical sense used by Gödel, so the demolition of the paradox must lie elsewhere.)

Chapter III

The Numbers of Analysis

The most basic object of study in analysis is the set of real numbers (along with functions on this set). The various number sets culminating in the real numbers (and, one step farther, the complex numbers) are generally familiar, at least informally. In this chapter, we will be fairly careful in describing these number systems through sets of axioms which characterize them and which can be used to develop all their properties and prove the theorems of analysis, while at the same time trying to avoid getting bogged down in set-theoretic technicalities. In the first section we describe the real numbers axiomatically, and then take a constructive approach in later sections beginning with the natural numbers and successively constructing the integers, rational numbers, real numbers, and complex numbers. The axiomatic approach of the first section may be sufficient for many readers, at least temporarily.

III.1. The Real Numbers – Axiomatic Approach

> "Mathematics is fundamentally an egalitarian subject. It should in principle be possible for anyone to check the correctness of any argument; nothing needs to be taken on trust. This is the philosophy. The reality is that some parts of mathematics are much harder to grasp than others."
>
> *J. K. Truss*[1]

There are two common ways the set $\mathbb{R}$ of real numbers is specified as a basis for analysis. In the more complete approach, the real numbers are constructed from "scratch" beginning with the natural numbers. This approach is taken in succeeding sections. While this approach is the "right" one from a logical standpoint, it is complicated and time-consuming and is probably not the best way for a beginning analysis student to proceed.

The other approach is to simply begin with a list of reasonable assumptions about $\mathbb{R}$, called *axioms*, which reflect the properties of the real numbers one expects from the informal approach seen in calculus and precalculus, and which are comprehensive enough that all the results of analysis can be developed from them. It is pretty easy for a beginning student to accept on faith that there is a number system satisfying the axioms, and it can be shown without too much difficulty that there is only one such number system

[1] [Tru97, p. v]

(in a sense which will be precisely specified). A student who has any doubts about such an acceptance can look at later sections to whatever extent necessary to convince him/her/self that the number system can be carefully constructed. While in principle no part of mathematics should be simply accepted on faith, in practice it may be prudent to postpone working through the full justification of some complicated parts of mathematics which appear reasonable. (The history of mathematics is full of such temporary leaps of faith, which have sometimes lasted for centuries! There is, of course, a danger in doing this, especially at the frontiers of knowledge: one can never be entirely sure the details will work out as expected until they are actually done.)

In this section, we take an axiomatic approach to the real numbers. We will not attempt to justify the reasonableness of most of the axioms, which we hope will be self-evident in light of previous experience, nor will we try to motivate the specific choice of axioms from among all possible statements we would expect to be true about $\mathbb{R}$ – after the fact it will become clear that we can get anything we might want as a logical consequence of the axioms we have chosen. The one axiom we will discuss and motivate is the crucial Completeness Axiom, which is the most complicated and least obvious of the axioms (and, for analysis, arguably the most important).

> "Our evidence to the effect that the order of addition is immaterial is limited to a small number of cases, whereas in practice we use this fact with all numbers. Hence, we are really making an assumption, namely, that for *any* two numbers a and b, integral, fractional, irrational, and negative, the order of addition will not affect the result. ... [W]e should not lose sight of the fact that the mathematics built upon the number system is a deductive system. This point needs emphasis because we begin to learn arithmetic at an early age by rote and thereafter we tend to operate with numbers mechanically without perceiving that we are constantly using axioms of numbers."
>
> *M. Kline*[2]

(This quote only applies directly to the axiomatic approach: in the constructive approach, the axioms are theorems. But the constructive approach is also really deductive, from the axioms of set theory.)

Throughout this section, the real numbers $\mathbb{R}$ will be assumed to be a set with various algebraic operations and an ordering, satisfying properties specified in the axioms.

III.1.1. The Algebraic Axioms of $\mathbb{R}$

The axioms for the real numbers come in three groups: the algebraic axioms, the order axioms, and one additional axiom called the *Completeness Axiom*. In this subsection, we list the algebraic axioms and derive the usual rules of algebra from them.

III.1.1.1. There are four standard algebraic operations on $\mathbb{R}$: addition, subtraction, multiplication, and division (other algebraic operations such as exponentiation can be defined in terms of these). In fact, subtraction and division can be defined in terms of addition and multiplication respectively, so addition and multiplication are the operations which need to be specified in the axioms.

[2] [?, p. 75-76, 81].

III.1.1.2. The Algebraic Axioms

(Op) There are two binary operations + *(addition)* and · *(multiplication)* on $\mathbb{R}$ associating to each pair (x, y) of elements of $\mathbb{R}$ elements $x + y$ and $x \cdot y$ of $\mathbb{R}$ ($x \cdot y$ is usually written xy).

(A1) The binary operation + is *associative*: $x + (y + z) = (x + y) + z$ for all $x, y, z \in \mathbb{R}$.

(A2) The binary operation + is *commutative*: $x + y = y + x$ for all $x, y \in \mathbb{R}$.

(A3) There is an element $0 \in \mathbb{R}$ such that $0 + x = x$ for all $x \in \mathbb{R}$.

(A4) For every $x \in \mathbb{R}$ there is an element $(-x) \in \mathbb{R}$ with $x + (-x) = 0$.

(M1) The binary operation · is *associative*: $x(yz) = (xy)z$ for all $x, y, z \in \mathbb{R}$.

(M2) The binary operation · is *commutative*: $xy = yx$ for all $x, y \in \mathbb{R}$.

(M3) There is an element $1 \in \mathbb{R}$, $1 \neq 0$, such that $1 \cdot x = x$ for all $x \in \mathbb{R}$.

(M4) For every $x \in \mathbb{R}$, $x \neq 0$, there is an element $x^{-1} \in \mathbb{R}$ with $xx^{-1} = 1$.

(D) The operation · *distributes* over +: $x(y + z) = xy + xz[= (xy) + (xz)]$ for all $x, y, z \in \mathbb{R}$.

Axiom (D) is called the *distributive law* or *distributive property.*

All of the usual rules of arithmetic and algebra can be proved from the algebraic axioms. The next proposition gives some of the most important and often-used ones. We write out parts of the proof in excruciating detail to illustrate the techniques.

III.1.1.3. Proposition.

(i) If $x, y, z \in \mathbb{R}$ and $x + z = y + z$, then $x = y$. *(Additive Cancellation)*

(ii) If $x, y, z \in \mathbb{R}$, $z \neq 0$, and $xz = yz$, then $x = y$. *(Multiplicative Cancellation)*

(iii) If $x \in \mathbb{R}$, then $0 \cdot x = 0$.

(iv) If $x, y \in \mathbb{R}$, then $(-x)y = -(xy)$.

(v) If $x \in \mathbb{R}$, then $(-1)x = -x$.

(vi) If $x \in \mathbb{R}$, then $-(-x) = x$.

(vii) If $x, y \in \mathbb{R}$, then $(-x)(-y) = xy$. In particular, $(-1)(-1) = 1$.

(viii) If $x, y \in \mathbb{R}$, $x \neq 0$, and $y \neq 0$, then $xy \neq 0$ and $(xy)^{-1} = (x^{-1})(y^{-1})$.

Proof: (i): If $x + z = y + z$, then $(x + z) + (-z) = (y + z) + (-z)$. But

$$\begin{aligned}(x + z) + (-z) &= x + (z + (-z)) \text{ by (A1)}\\ &= x + 0 \text{ by (A4) (definition of } -z)\\ &= 0 + x \text{ by (A2)}\\ &= x \text{ by (A3)}\ .\end{aligned}$$

By a similar argument, $(y+z)+(-z)=y$. Thus $x=y$.

(ii): The proof is almost identical to the proof of (i), and is left to the reader (Exercise ()).

(iii): We have

$$\begin{aligned} 0\cdot x + x &= 0\cdot x + 1\cdot x \text{ by (M3)} \\ &= x\cdot 0 + x\cdot 1 \text{ by (M2)} \\ &= x(0+1) \text{ by (D)} \\ &= x\cdot 1 \text{ by (A3)} \\ &= 1\cdot x \text{ by (M2)} \\ &= x \text{ by (M3)} \\ &= 0 + x \text{ by (A3)} \ . \end{aligned}$$

Thus we conclude that $0\cdot x = 0$ by (i) (applied to $(0\cdot x, 0, x)$ in place of (x,y,z)).

(iv): We have

$$\begin{aligned} ((-x)y) + (xy) &= y(-x) + (yx) \text{ by (M2)} \\ &= y((-x)+x) \text{ by (D)} \\ &= y(x+(-x)) \text{ by (A2)} \\ &= y\cdot 0 \text{ by (A4) (definition of } -x) \\ &= 0\cdot y \text{ by (M2)} \\ &= 0 \text{ by (iii)} \\ &= (xy) + (-(xy)) \text{ by (A4) (definition of } -(xy)) \\ &= (-(xy)) + (xy) \text{ by (A2)} \end{aligned}$$

Therefore, we have $(-x)y = -(xy)$ by (i).

(v): By (iv), $(-1)x = -(1\cdot x)$. But $1\cdot x = x$ by (M3).

(vi): We abbreviate the argument (as we will from now on). We have

$$[-(-x)] + (-x) = 0 = x + (-x)$$

and so $-(-x) = x$ by (i).

(vii): We have

$$(-x)(-y) = -[x(-y)] = -[-(xy)] = xy \text{ by (vi)} \ .$$

(viii): We have

$$(xy)[(x^{-1})(y^{-1})] = [(xy)x^{-1}](y^{-1}) = [x(yx^{-1})](y^{-1}) = [x(x^{-1}y)](y^{-1}) = [(xx^{-1})y](y^{-1}) = yy^{-1} = 1 \ .$$

If $xy = 0$, then $(xy)[(x^{-1})(y^{-1})] = 0\cdot(x^{-1}y^{-1}) = 0$ by (iii); but $1 \neq 0$ by (M3), a contradiction. Thus $xy \neq 0$. We then have $(xy)[(xy)^{-1}] = 1 = (xy)[(x^{-1})(y^{-1})]$, so $(xy)^{-1} = (x^{-1})(y^{-1})$ by (ii). 👍

III.1.1.4. By repeated application of the associative laws, a sum or product of three or more factors is unambiguously defined, so we may unambiguously write $x + y + z$, xyz, etc., for sums and products of real numbers. In particular, if $x \in \mathbb{R}$, the product $x \cdot x \cdots x$ (n factors) is unambiguously defined; write x^n for this product. By repeated application of III.1.1.3.(viii), $(x^n)^{-1} = (x^{-1})^n$ for any nonzero $x \in \mathbb{R}$ and any n; write x^{-n} for this real number. (The last two arguments are a little too informal; they can be done carefully by induction (Exercise III.1.14.3.).) The usual rules of exponents hold (III.1.14.3.).

III.1.1.5. All the usual rules of algebra (with the exception of existence of roots; cf. III.1.12.) can be similarly derived from the algebraic axioms of $\mathbb{R}$. In the future, we will generally use standard operations of algebra without comment. However, the reader should be sure he/she is *able* to write out the complete tedious proof of any algebraic formula if required, along the lines of the proofs in III.1.1.3..

From now on, we will also use standard algebraic notation: if $x, y \in \mathbb{R}$, we will write $x - y$ for $x + (-y)$ and $\frac{x}{y}$ for xy^{-1} (if $y \neq 0$). We will also group multiplication before addition in the usual way: $xy + z$ means $(xy) + z$, etc.

III.1.1.6. A set with two binary operations satisfying (A1)–(A4), (M1)–(M4), and (D) is called a *field*, and these axioms are often called the *field axioms*. There are many fields with quite varied properties; see III.4.3.1. and III.4.3.5.(c), or books on abstract algebra for more details. Since the algebraic properties of $\mathbb{R}$ are proved just using the field axioms, they are valid in any field.

III.1.2. The Order Axioms for $\mathbb{R}$

In addition to the algebraic structure, there is an ordering on the real numbers which (among other things) allows $\mathbb{R}$ to be pictured geometrically as a line.

III.1.2.1. There are two equivalent ways the ordering on $\mathbb{R}$ can be axiomatized: either directly, or by describing the set of positive numbers. Each is useful in different contexts. We will begin with the direct approach:

(Ord) There is a relation (strict total order) $<$ on $\mathbb{R}$, such that

- (O1) If $x, y \in \mathbb{R}$, then exactly one of $x < y$, $y < x$, $x = y$ holds. *(Trichotomy)*
- (O2) If $x, y, z \in \mathbb{R}$, $x < y$, and $y < z$, then $x < z$. *(Transitivity)*
- (O3) If $x, y, z \in \mathbb{R}$ and $x < y$, then $x + z < y + z$. *(Translation Invariance)*
- (O4) If $x, y, z \in \mathbb{R}$, $x < y$, and $0 < z$, then $xz < yz$. *(Homogeneity or Scaling Invariance)*

III.1.2.2. We will freely use the usual extensions of the order notation. We write $x \leq y$ to mean $x < y$ or $x = y$. Write $x > y$ to mean $y < x$, and $x \geq y$ to mean $y \leq x$, i.e. $x > y$ or $x = y$.

As with the algebraic axioms, there are many standard order results which can be proved from the axioms. Here are some of the basic ones:

III.1.2.3. PROPOSITION.

(i) If $x, y \in \mathbb{R}$ and $x < y$, then $(-y) < (-x)$.

(ii) If $x, y \in \mathbb{R}$, $x < y$, and $z < 0$, then $yz < xz$.

(iii) $0 < 1$.

(iv) If $x \in \mathbb{R}$ and $0 < x$, then $0 < x^{-1}$.

(v) If $x, y \in \mathbb{R}$, $0 < x$, and $x < y$, then $y^{-1} < x^{-1}$.

(vi) If $x, y, z, w \in \mathbb{R}$, $x < y$, and $z < w$, then $x + z < y + w$.

(vii) If $x, y, z, w \in \mathbb{R}$, $0 < x < y$, and $0 < z < w$, then $xz < yw$.

PROOF: (i): If $x < y$, by (O3) we have $-y = x - x - y < y - x - y = -x$.
(ii): If $x < y$ and $z < 0$, by (i) we have $0 < (-z)$, so $-(xz) = x(-z) < y(-z) = -(yz)$, and hence $yz < xz$ by (i).
(iii): Since $1 \neq 0$ by (M3), either $0 < 1$ or $1 < 0$. If $1 < 0$, we have $0 < (-1)$ by (i), so $0 < (-1)(-1) = 1$ by (O4), a contradiction.
(iv): We have $x^{-1} \neq 0$. If $x^{-1} < 0$, then we have

$$1 = xx^{-1} < x0 = 0$$

by (O4), contradicting (iii).
(v): We have $0 < x^{-1}$ and $0 < y^{-1}$ by (iv), so $1 = x^{-1}x < x^{-1}y$ by (O4), and hence $y^{-1} = y^{-1}1 < x^{-1}yy^{-1} = x^{-1}$ again by (O4).
(vi): We have $x + z < y + z < y + w$ by two applications of (O3). The proof of (vii) is almost identical. 👌

III.1.2.4. Transitivity also holds with $\leq$, and with $>$ and $\geq$:

If $x \leq y$ and $y \leq z$, then $x \leq z$.

If $x < y$ and $y \leq z$, then $x < z$.

If $x \leq y$ and $y < z$, then $x < z$.

If $x > y$ and $y > z$, then $x > z$.

If $x \geq y$ and $y \geq z$, then $x \geq z$.

If $x > y$ and $y \geq z$, then $x > z$.

If $x \geq y$ and $y > z$, then $x > z$.

The results of III.1.2.3. with $<$ replaced by $\leq$ throughout also hold. The simple proofs, checking case-by-case, are left to the reader.

Note, however, that transitivity only holds for inequalities going the same direction: for example, if $x < y$ and $y > z$, then no relation between x and z can be concluded. Thus the direction of inequalities must always be carefully noted. ✋

III.1.2.5. Instead of using (Ord) as the axiomatization of the order structure on $\mathbb{R}$, one can equivalently use an axiom describing the positive numbers:

(Pos) There is a nonempty subset P of $\mathbb{R}$ such that
- (P1) If $x, y \in P$, then $x + y \in P$.
- (P2) If $x, y \in P$, then $xy \in P$.
- (P3) $0 \notin P$.
- (P4) If $x \in \mathbb{R}$, $x \neq 0$, then either $x \in P$ or $(-x) \in P$, but not both.

In the presence of the algebraic axioms, it is easy to derive the (Ord) axioms from (Pos), and vice versa. The arguments are good examples of how properties are proved from the axioms.

III.1.2.6. Suppose we assume the axiom (Pos), along with the algebraic axioms. We then define the relation $<$ by saying $x < y$ if (and only if) $(y - x) \in P$. We verify the axioms (O1)–(O4) from (Pos):
(O1): If $x, y \in \mathbb{R}$, then from (P3) and (P4) we conclude that exactly one of $y - x = 0$, $(y - x) \in P$, $-(y - x) = (x - y) \in P$ holds. Thus exactly one of $x = y$, $x < y$, or $y < x$ holds.
(O2): If $x < y$ and $y < z$, i.e. $(y - x) \in P$ and $(z - y) \in P$, then $(z - x) = (y - x) + (z - y) \in P$ by (P1).
(O3): $(y + z) - (x + z) = y - x$.
(O4): If $x < y$ and $0 < z$, i.e. $(y - x) \in P$ and $z \in P$, then $yz - xz = z(y - x) \in P$ by (P2).

III.1.2.7. Now assume (Ord), along with the algebraic axioms. Define $P = \{x \in \mathbb{R} : 0 < x\}$. We verify that axioms (P1)–(P4) hold for this P:
(P1): If $x, y \in P$, i.e. $0 < x$ and $0 < y$, then we have, using (O3) and the algebraic axioms,

$$0 = 0 + 0 < x + 0 < x + y$$

so $x + y \in P$.
(P2): This is very similar to (P1). If $x, y \in P$, i.e. $0 < x$ and $0 < y$, by (O4) and III.1.1.3.(iii) we have

$$0 = 0 \cdot y < x \cdot y \ .$$

(P3): By (O1), we do not have $0 < 0$.
(P4): If $x \neq 0$, then by (O1) we have $0 < x$ or $x < 0$, but not both. If $0 < x$, then $x \in P$. If $x < 0$, then $0 = x + (-x) < 0 + (-x) = -x$ by (O3) (and (A3)–(A4)), so $(-x) \in P$.

III.1.2.8. Thus we are free to assume either (Pos) or (Ord), along with the algebraic axioms; exactly the same results can be obtained in each case. We will sometimes use (Pos) and sometimes (Ord), depending on which is more convenient at the time.

III.1.2.9. A field additionally satisfying (Ord) (or, equivalently, (Pos)) is called an *ordered field*. Ordered fields look somewhat more like $\mathbb{R}$ than general fields, but there are still many ordered fields besides $\mathbb{R}$; see, for example, III.5.3.3.(d) and III.10.2.21.. The results of III.1.2.3. and III.1.2.4. hold in a general ordered field since they are just proved from the axioms.

The final axiom, the Completeness Axiom, is what distinguishes $\mathbb{R}$ from all other ordered fields. But before discussing this axiom, we describe some standard sets of numbers and their properties.

III.1.3. Absolute Value

Informally, the absolute value of a real number is the number "with the sign thrown away," i.e. with a negative sign changed to + if necessary. Formally:

III.1.3.1. DEFINITION. Let x be a real number. Then the *absolute value of* x, denoted $|x|$, is

$$|x| = \begin{cases} x & \text{if} \quad x \geq 0 \\ -x & \text{if} \quad x < 0 \end{cases} .$$

III.1.3.2. This definition may look a little jarring: how can there be a minus sign? The point is that if $x < 0$, then $-x > 0$. The statement is equivalent to the more natural looking

$$x = \begin{cases} |x| & \text{if} \quad x \geq 0 \\ -|x| & \text{if} \quad x < 0 \end{cases}$$

but this version is not a good form for a definition.

III.1.3.3. The absolute value of a number x can be profitably thought of as the distance between x and 0. (However, this is unsuitable as a definition of $|x|$; in fact, it can be taken as a definition of *distance*.) By translation, if x and y are real numbers, the number $|x - y| = |y - x|$ can be thought of as the distance between x and y on the number line.

Here are some simple but important properties of absolute value. These can be proved brute-force in a low-tech way, case-by-case.

III.1.3.4. PROPOSITION. (i) If $x \in \mathbb{R}$, then $|x| = |-x| = \sqrt{x^2} \geq 0$, and $|x| > 0$ if (and only if) $x \neq 0$.
(ii) If $x, y \in \mathbb{R}$, then $|xy| = |x||y|$.
(iii) If $x, y \in \mathbb{R}$, then $|x + y| \leq |x| + |y|$ (*triangle inequality*).
(iv) If $x, y \in \mathbb{R}$, then $||x| - |y|| \leq |x - y|$.
In (iii) and (iv), equality holds if and only if x and y have the same sign (or one is zero).

III.1.4. The Natural Numbers

III.1.4.1. The first set of numbers developed, going back to prehistoric times, is the *natural numbers* or *counting numbers*, also described as the positive integers. This set is usually called $\mathbb{N}$. Thus, as it is often described or defined,

$$\mathbb{N} = \{1, 2, 3, \dots\} .$$

Unfortunately, this "description" is totally inadequate as a definition of $\mathbb{N}$: it is entirely unclear what the "..." means (if you think you know, try giving a precise formulation of the meaning!)

Instead, we make the following careful definition (assuming we already have $\mathbb{R}$):

III.1.4.2. Definition. $\mathbb{N}$ is the smallest subset of $\mathbb{R}$ containing 1 and which is closed under adding 1, i.e. if $n \in \mathbb{N}$, then $(n+1) \in \mathbb{N}$.

To see that there is such a smallest subset of $\mathbb{R}$, we could alternately define $\mathbb{N}$ to be the intersection of all subsets S of $\mathbb{R}$ with the property that $1 \in S$ and, whenever $n \in S$, $n+1$ is also in S. There is at least one such subset S, namely $\mathbb{R}$ itself.

III.1.4.3. One mathematically minor point needs to be addressed. There is no general agreement among mathematicians about whether $\mathbb{N}$ includes 0; some mathematicians include it, and others do not. We will not. The question of whether to work with $\mathbb{N}$ or $\mathbb{N} \cup \{0\}$ as the basic set is strictly a matter of taste, since it is mathematically trivial to go from either to the other. See III.2.1.2. for further comments.

We now derive some of the basic properties of $\mathbb{N}$ directly from its definition.

III.1.4.4. Proposition. $\mathbb{N}$ is closed under addition, i.e. if $n, k \in \mathbb{N}$, then $(n+k) \in \mathbb{N}$.

Proof: Fix $k \in \mathbb{N}$, and let $S_k = \{n \in \mathbb{N} : n+k \in \mathbb{N}\}$. Then $1 \in S_k$ since $k+1 \in \mathbb{N}$. If $n \in S_k$, then $n \in \mathbb{N}$ and $n+k \in \mathbb{N}$, so $n+1 \in \mathbb{N}$ and $(n+1)+k = (n+k)+1 \in \mathbb{N}$, and hence $n+1 \in S_k$. Thus $S_k = \mathbb{N}$. 👍

III.1.4.5. Proposition. $\mathbb{N}$ is closed under multiplication, i.e. if $n, k \in \mathbb{N}$, then $nk \in \mathbb{N}$.

Proof: Fix $k \in \mathbb{N}$, and let $S_k = \{n \in \mathbb{N} : nk \in \mathbb{N}\}$. Then $1 \in S_k$ since $1 \cdot k = k \in \mathbb{N}$. If $n \in S_k$, then $n \in \mathbb{N}$ and $nk \in \mathbb{N}$. Then $n+1 \in \mathbb{N}$, and $(n+1)k = nk + n \in \mathbb{N}$ by III.1.4.4., so $n+1 \in S_k$. Thus $S_k = \mathbb{N}$. 👍

III.1.4.6. Proposition. If $n \in \mathbb{N}$, then either $n = 1$ or $n-1 \in \mathbb{N}$.

Proof: Let $S = \{n \in \mathbb{N} : n = 1 \text{ or } n-1 \in \mathbb{N}\}$. Then $1 \in S$, and if $n \in S$, then $n+1 \in \mathbb{N}$ and $(n+1)-1 = n \in \mathbb{N}$, so $n+1 \in S$. Thus $S = \mathbb{N}$. 👍

III.1.4.7. Proposition. If $m, n \in \mathbb{N}$ and $m < n$, then $n - m \in \mathbb{N}$.

Proof: Let S be the set of all $k \in \mathbb{N}$ such that $n \in \mathbb{N}$, $n > k$ implies $n - k \in \mathbb{N}$. Then $1 \in S$ by III.1.4.6.. If $k \in S$ and $n > k+1$, then $n > k$, so $n-k \in \mathbb{N}$, and $n-k > 1$ since $n > k+1$; thus $(n-k)-1 = n-(k+1) \in \mathbb{N}$ by III.1.4.6.. Thus $k+1 \in S$. So $S = \mathbb{N}$. 👍

III.1.4.8. Proposition. If $n \in \mathbb{N}$, then $n \geq 1$.

Proof: Let $S = \{n \in \mathbb{N} : n \geq 1\}$. Then $1 \in S$, and if $n \in S$, then $n+1 \in S$; so $S = \mathbb{N}$. 👍

III.1.4.9. PROPOSITION. If $m, n \in \mathbb{N}$ and $m \neq n$, then $|m - n| \geq 1$.

PROOF: We may assume $m < n$. Then $n - m \in \mathbb{N}$, so $n - m \geq 1$. 👍

III.1.4.10. DEFINITION. Let $n \in \mathbb{N}$. An $m \in \mathbb{N}$ is a *predecessor* of n if $m < n$.

III.1.4.11. PROPOSITION. Every $n \in \mathbb{N}$ has only finitely many predecessors in $\mathbb{N}$. (To avoid circularity, we interpret this statement to mean that the set of predecessors of any $n \in \mathbb{N}$ is a finite set in the sense of II.5.2.6..)

PROOF: Let S be the set of all $n \in \mathbb{N}$ with only finitely many predecessors in $\mathbb{N}$. Then $1 \in S$ since 1 has no predecessors in $\mathbb{N}$ (III.1.4.8.). Suppose $n \in S$. If $m \in \mathbb{N}$ is a predecessor of $n + 1$, i.e. $m < n + 1$, then $(n + 1) - m \in \mathbb{N}$, so $(n + 1) - m \geq 1$. If $(n + 1) - m = 1$, then $m = n$; otherwise $(n + 1) - m > 1$ (III.1.4.8.) and $n > m$, i.e. m is a predecessor of n. Thus the set of predecessors of $n + 1$ is $\{n\} \cup \{\text{predecessors of } n\}$, which is a finite set, so $n + 1 \in S$. Thus $S = \mathbb{N}$. 👍

III.1.4.12. PROPOSITION. Every nonempty subset of $\mathbb{N}$ has a smallest element ($\mathbb{N}$ is *well ordered* ()).

PROOF: Let S be a nonempty subset of $\mathbb{N}$, and let $n \in S$. The set A of all elements of S which are predecessors of $n + 1$ is finite by III.1.4.11., and nonempty since $n \in A$. But any finite nonempty subset of $\mathbb{R}$ has a smallest element (II.5.2.12.). So A has a smallest element, which must be the smallest element of S. 👍

III.1.5. Induction and Recursion

The properties of $\mathbb{N}$ we have developed lead to an important method of proof of statements involving natural numbers:

III.1.5.1. [PRINCIPLE OF INDUCTION] Suppose $P(n)$ is a statement for each $n \in \mathbb{N}$. Suppose the following two statements hold:

(i) $P(1)$ is true.

(ii) For every $n \in \mathbb{N}$, if $P(n)$ is true, then $P(n + 1)$ is true.

Then $P(n)$ is true for all $n \in \mathbb{N}$.

We could have called this a Theorem, but the statement is not precise enough to qualify. One reason is that we have not been precise about the meaning of the term "statement", since to be so would require an extensive excursion into set theory and mathematical logic. We have also been imprecise about what it means to say a statement is "true"; this should be interpreted to mean *provable*, i.e. a Theorem in the theory being considered, or occasionally "true in a given model." Many books on mathematical logic such as [**?**] contain careful discussions of these points. But any type of mathematical statement or interpretation of truth likely to be made by a working mathematician comes within the scope of the principle.

III.1.5.2. To see why the Principle of Induction holds, set

$$S = \{n \in \mathbb{N} : P(n) \text{ is true } \} \ .$$

We have that $1 \in S$ by (i), and if $n \in S$ we have $n+1 \in S$ by (ii). Thus $S = \mathbb{N}$.

III.1.5.3. Proofs by induction have been described as knocking over a sequence of dominoes. The "inductive step" (ii) in the procedure amounts to showing that, for each n, if the n'th domino falls, it knocks over the $(n+1)$'st also. But you still have to push over the first domino by showing (i).

III.1.5.4. Careful proofs of many algebraic formulas can, and should, be done by induction. One example, the subject of a famous (but possibly apocryphal) story, is a formula for the sum of the first n natural numbers.

According to the story, when C. F. Gauss, one of the greatest mathematicians in history, was a boy in the late 18'th century, his schoolteacher became annoyed with his class and told the children to add the (natural) numbers from 1 to 100. The teacher soon saw Gauss sitting quietly, not working, and told him to get busy. He replied that he had already finished, and the teacher found that he had the correct answer. When asked how he did it so quickly, Gauss replied that he had written the numbers in increasing order on one line, in decreasing order on the next line, and added them vertically first pairwise to obtain a sum with all terms the same, which can be computed by multiplication, giving twice the desired sum. Symbolically, he wrote

$$\begin{array}{ccccccccccc} 1 & + & 2 & + & \cdots & + & 99 & + & 100 & = & s \\ 100 & + & 99 & + & \cdots & + & 2 & + & 1 & = & s \\ \hline 101 & + & 101 & + & \cdots & + & 101 & + & 101 & = & 2s \end{array}$$

He then computed the sum in the last line to be $(100)(101) = 10100$, so $s = 10100/2 = 5050$.

This clever argument can be easily modified to develop a formula for the sum of the natural numbers from 1 to n, for any n:

$$\begin{array}{ccccccccccc} 1 & + & 2 & + & \cdots & + & (n-1) & + & n & = & s \\ n & + & (n-1) & + & \cdots & + & 2 & + & 1 & = & s \\ \hline (n+1) & + & (n+1) & + & \cdots & + & (n+1) & + & (n+1) & = & 2s \end{array}$$

and the last sum is $n(n+1)$, yielding the formula

$$s = \frac{n(n+1)}{2} \ .$$

However, although this argument is a good way to *derive* (or, more accurately, *guess*) the correct formula, it is not precise enough to constitute a rigorous proof, primarily because of the $\cdots$ in each row. To give a careful proof, we proceed by induction, taking, for each n,

$$P(n): \ \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \ .$$

To carry out the proof, we must (i) verify $P(1)$, which is easy: both sides of the equation are 1. Then we must show (ii), that $P(n)$ implies $P(n+1)$ for any n. So let n be a natural number, and *assume* that

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \ .$$

We then need to show that

$$\sum_{k=1}^{n+1} k = \frac{(n+1)([n+1]+1)}{2} = \frac{(n+1)(n+2)}{2} \ .$$

But we have

$$\sum_{k=1}^{n+1} k = \left[\sum_{k=1}^{n} k\right] + (n+1) = \frac{n(n+1)}{2} + (n+1)$$
$$= \frac{n(n+1)+2(n+1)}{2} = \frac{(n+2)(n+1)}{2}$$

which is what was to be shown. By the Principle of Induction, $P(n)$ is true for all n, i.e. the formula holds for all n.

Proof by induction is used in far more circumstances than just establishing algebraic formulas, however.

Complete Induction

A proof procedure which is a variation of ordinary induction is complete induction.

III.1.5.5. [Principle of Complete Induction] Suppose $P(n)$ is a statement for each $n \in \mathbb{N}$. Suppose, for each $n \in \mathbb{N}$, the following *inductive implication* holds:

"If $P(m)$ is true for all $m < n$, then $P(n)$ is true."

Then $P(n)$ is true for all $n \in \mathbb{N}$.

We could have again called this a Theorem, but the statement is also not precise enough to qualify because of the imprecision in the meaning of the terms "statement" and "true."

III.1.5.6. Complete induction is not quite the same as ordinary induction, where statements $P(n)$ indexed by $\mathbb{N}$ are proved for all n by proving $P(1)$ and $[P(n) \Rightarrow P(n+1)]$ for all n. Instead, in complete induction, we assume $P(m)$ is true for all $m \in \mathbb{N}$, $m < n$, and deduce $P(n)$. Ordinary induction is a special case of complete induction [if we have that $P(n)$ implies $P(n+1)$, then we certainly have that $P(m)$ for all $m < n+1$ implies $P(n+1)$]; on the other hand, complete induction can be easily deduced from ordinary induction. Indeed, suppose the statements $P(n)$ satisfy the hypotheses of complete induction. For each n, set

$$Q(n)\text{: "}P(m) \text{ is true for all } m < n\text{."}$$

Then, for each $n \in \mathbb{N}$, the statement "$Q(n)$ implies $P(n)$" is logically equivalent to "$Q(n)$ implies $Q(n+1)$." We have that $Q(1)$ is automatically true since there are no $m < 1$, so by ordinary induction we obtain that $Q(n)$ is true for all n, which implies that $P(n)$ is true for all n.

Complete induction is useful where there is no simple relationship between $P(n)$ and $P(n+1)$, but there is a relationship between $P(n)$ and $P(m)$ for certain m's less than n. The next simple example is a good illustration. (We are slightly ahead of ourselves here, since prime numbers are technically only defined in III.1.6.13., but the definition is well known.)

III.1.5.7. PROPOSITION. Every natural number can be written as a finite product of prime numbers. [We define the product of an empty set of primes to be 1, and a product of one prime p to be p.]

PROOF: For each $n \in \mathbb{N}$, let $P(n)$ be the statement "The integer n can be written as a finite product of prime numbers." Fix $n \in \mathbb{N}$, and assume $P(m)$ is true for all $m < n$; we must prove that $P(n)$ is true. There are three cases:
(1) $n = 1$. Then n is an empty product of prime numbers.
(2) $n > 1$ is prime. Then n is a product of one prime number.
(3) $n > 1$ is not prime. Then $n = n_1 n_2$ for some $n_1, n_2 \in \mathbb{N}$, $n_1, n_2 > 1$. Then $n_1, n_2 < n$, so by assumption n_1 and n_2 can each be written as a finite product of prime numbers. Hence so can $n_1 n_2 = n$. 👍

III.1.5.8. Note that in complete induction, it is not necessary in general to separately prove $P(1)$, since it is included in the general case. In the case $n = 1$, the assumption that $P(m)$ is true for all $m < n$ is automatically satisfied since there are no such m's (in this case, the hypothesis is said to be *vacuous*.) Of course, it is *allowed* to prove $P(1)$ as a special case, as was done in the proof of III.1.5.7..

Infinite Regress

There is an alternate way to phrase a proof by induction or complete induction, based on the fact (III.1.4.12.) that every nonempty subset of $\mathbb{N}$ has a smallest element.

III.1.5.9. [PRINCIPLE OF INFINITE REGRESS] Suppose $P(n)$ is a statement for each $n \in \mathbb{N}$. Suppose we have:

> "For each $n \in \mathbb{N}$, if $P(n)$ is false, then there is an $m \in \mathbb{N}$, $m < n$, such that $P(m)$ is false."

Then $P(n)$ is true for all n.

III.1.5.10. To prove this, let

$$S = \{n \in \mathbb{N} : P(n) \text{ is false } \} .$$

If S is nonempty, then it has a smallest element n. But if $n \in S$, the hypothesis insures that there is an $m \in S$ with $m < n$, a contradiction. Thus $S = \emptyset$.

III.1.5.11. Actually, the statement in the Principle of Infinite Regress is just the contrapositive of the statement of Complete Induction, so Infinite Regress is logically equivalent to Complete Induction (and hence to ordinary Induction).

Infinite Regress is sometimes used in a seemingly more perverse way: if for each n we have a statement $Q(n)$, and if we can show that if $Q(n)$ is *true*, then there is an $m < n$ such that $Q(m)$ is true, then it follows that $Q(n)$ cannot be true for any n! To show this, let $P(n)$ be the negation of $Q(n)$ for each n, and apply infinite regress to the $P(n)$.

We often phrase arguments using Infinite Regress in a slightly different way, effectively combining Complete Induction with a proof by contradiction: to prove that $P(n)$ is true for all n we assume $P(n)$ is false for some n, let n be the smallest natural number for which $P(n)$ is false, and obtain a contradiction by showing that $P(m)$ is false for some $m \in \mathbb{N}$, $m < n$. "True" and "false" can be interchanged in this statement.

III.1.5.12. As an example of the use of this principle, we redo the proof of III.1.5.7.. Suppose not every element of $\mathbb{N}$ can be written as a product of prime numbers, and let n be the smallest natural number which cannot. Then $n \neq 1$, since 1 can be written as the empty product of primes, and n is not prime, since a prime number can be written as a product of length 1. Thus n can be written as $n_1 n_2$, where $1 < n_1, n_2 < n$. Then at least one of n_1, n_2 cannot be written as a product of primes, contradicting that n is the smallest such number.

This argument is not really simpler when phrased this way, but some arguments do become much simpler or more natural (e.g. III.1.12.6.). In principle, any argument done by induction or complete induction can be rephrased as an infinite regress argument, and conversely.

Definitions by Induction or Recursion

Induction (or, technically, *recursion*) can also be used to define sequences. We will again not be completely precise in defining terms; see, for example, [**?**, p. 249-250] for a careful treatment.

III.1.5.13. Let X be a set. Suppose, for each $n \in \mathbb{N}$, we have a definite procedure or rule R_n for choosing an $x_n \in X$; the x_n may depend on the previous choice of the x_m for $m < n$ (but the rule R_n does not depend on the choice of the x_m for $m < n$). Then

III.1.5.14. [Principle of Recursion] There is a unique sequence (x_n) in X such that, for each $n \in \mathbb{N}$, x_n is determined by the specified procedure R_n applied to $\{x_m : m < n\}$.

Roughly speaking, if the sequence can be defined "one step at a time" by the rules R_n, then it can be defined "all at once." The procedure is often used when all the rules R_n (except R_1, which just amounts to specification of the first term of the sequence) are the same.

Here is a precise version of the Principle of Recursion:

III.1.5.15. Theorem. Let X be a set, and for each $n \in \mathbb{N}$ let f_n be a function from X^n to X. If x_1 is any element of X, then there is a unique sequence (x_n) in X with $x_{n+1} = f_n(x_1, \ldots, x_n)$ for all $n \in \mathbb{N}$.

Proof: Let S be the smallest subset of $\mathbb{N} \times X$ containing $(1, x_1)$, with the property that, for any $n \in \mathbb{N}$, whenever $\{(k, y_k) : k \in \mathbb{N}, k \leq n\} \subseteq S$ for $(y_1, \ldots, y_n) \in X^n$, then $(n+1, f_n(y_1, \ldots, y_n)) \in S$. ($S$ is the

intersection of all subsets of $\mathbb{N} \times X$ with these two properties; there is at least one such subset, $\mathbb{N} \times X$ itself.) Then let

$$A = \{n \in \mathbb{N} : (k, x_k) \in S \text{ for exactly one } x_k \in X \text{ for all } k \in \mathbb{N}, k \leq n\} .$$

First we show $1 \in A$. Set

$$S' = \{(1, x_1)\} \cup \{(n, x) \in S : n > 1\}$$

and note that $S' \subseteq S$, $(1, x_1) \in S'$, and $\{(k, y_k) : k \in \mathbb{N}, k \leq n\} \subseteq S'$ implies $(n+1, f_n(y_1, \ldots, y_n)) \in S'$; hence $S' = S$ and $1 \in A$. Now suppose $n \in A$, i.e. there is a unique $x_k \in X$ with $(k, x_k) \in S$ for all $k \in \mathbb{N}$, $k \leq n$. Set

$$S'' = \{(k, x_k) : k \in \mathbb{N}, k \leq n\} \cup \{(n+1, f_n(x_1, \ldots, x_n))\} \cup \{(m, x) \in S : m > n+1\}$$

and note that $S'' \subseteq S$, $(1, x_1) \in S''$, and $\{(k, y_k) : k \in \mathbb{N}, k \leq m\} \subseteq S''$ implies $(m+1, f_m(y_1, \ldots, y_m)) \in S''$. Thus $S'' = S$ and $n + 1 \in A$. So $A = \mathbb{N}$.

For each $n \in \mathbb{N}$, let x_n be the unique element of X for which $(n, x_n) \in S$. Let $B = \{n \in \mathbb{N} : x_{n+1} = f_n(x_1, \ldots, x_n)\}$. We have $1 \in B$ since $(2, f_1(x_1)) \in S$; and if $n \in B$, then $(n+2, f_{n+1}(x_1, \ldots, x_{n+1})) \in S$, so $n + 1 \in B$. Thus $B = \mathbb{N}$. 👍

Note that no form of the Axiom of Choice is needed in this proof; in fact, the AC is not needed even for the transfinite version II.4.2.14..

III.1.5.16. In other words, if a definite rule is specified for each n for determining the $(n+1)$'st term of the sequence from the first n terms, there is a unique sequence generated beginning with any first term. The rule f_n can vary with n, but all the f_n must be specified in advance. The most common situation is that there is one function $f : X \to X$ such that $f_n(x_1, \ldots, x_n) = f(x_n)$ for each n, i.e. the next term in the sequence is determined just from the previous term by the same rule each time.

The sequence is said to be *defined by induction* (or, more correctly, *defined by recursion*) from x_1 by the functions (f_n).

III.1.5.17. **Caution:** In definitions by induction or recursion, the rule R_n can be quite general, but it must be a definite rule giving a specific element as a function of $\{x_m : m < n\}$. Thus, for example, a "rule" like "Let x_n be any element of $X \setminus \{x_m : m < n\}$" is not permissible, even if it is proved that this set is always nonempty (say, by a cardinality argument.) But if X is well-ordered (), it is permissible to let R_n be the rule "Let x_n be the *first* element of $X \setminus \{x_m : m < n\}$" (provided this set is always nonempty.) ✋

III.1.6. The Integers

III.1.6.1. The next set of numbers we will consider is the *integers*, or *whole numbers* (although historically positive fractions predated the concept of 0 or negative integers). The letter $\mathbb{Z}$ (from the German *Zahl*, "number") is customarily used to denote the set of integers. Thus, as is often stated,

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\} .$$

As with $\mathbb{N}$, this is too imprecise to be even an informal definition of $\mathbb{Z}$.

Instead, we define $\mathbb{Z}$ to be the smallest subset of $\mathbb{R}$ containing $\mathbb{N}$ and closed under subtraction, i.e. with the property that if $m, n \in \mathbb{Z}$, then $m - n \in \mathbb{Z}$.

III.1.6.2. We could describe $\mathbb{Z}$ more precisely as

$$\mathbb{Z} = \{m - n : m, n \in \mathbb{N}\} .$$

Indeed, this set contains $\mathbb{N}$ and is closed under subtraction; on the other hand, every difference of natural numbers must be contained in $\mathbb{Z}$.

There is a conceptual problem with defining $\mathbb{Z}$ to be the set of differences of natural numbers, however. Although this description is technically correct, there is no uniqueness: an integer can be written as a difference of natural numbers in many different ways. Thus it is conceptually clearer to define $\mathbb{Z}$ the way we did.

III.1.6.3. There is another standard way to describe $\mathbb{Z}$: $\mathbb{Z}$ is the disjoint union of $\mathbb{N}$, $\{0\}$, and $-\mathbb{N} = \{-n : n \in \mathbb{N}\}$. It is clear that the union of these three sets is contained in $\mathbb{Z}$, and is closed under subtraction by checking case-by-case, so equals $\mathbb{Z}$. To show the sets are disjoint, just note that if $n \in \mathbb{N}$, then $n > 0$, so $-n < 0$ and $-n \notin \mathbb{N}$. Thus we observe

$$\mathbb{N} = \{x \in \mathbb{Z} : x > 0\} .$$

III.1.6.4. Since $\mathbb{Z}$ is closed under subtraction and taking negatives, it is also closed under addition. In fact, it is the smallest subset of $\mathbb{R}$ containing $\mathbb{N}$ and closed under addition and taking negatives. (It is not the smallest subset of $\mathbb{R}$ containing $\mathbb{N}$ and closed under taking negatives: that would be $\mathbb{N} \cup (-\mathbb{N})$.)

III.1.6.5. It also follows immediately from III.1.6.3. and III.1.1.3.(iv) and (vii) that $\mathbb{Z}$ is closed under multiplication. Thus $\mathbb{Z}$ satisfies all the field axioms (III.1.1.2.) except (M4). (In algebraic language, it is a *commutative unital ring*). In fact, (M4) fails spectacularly: the only elements of $\mathbb{Z}$ with multiplicative inverses in $\mathbb{Z}$ are 1 and -1.

Number Theory

III.1.6.6. All the algebraic properties of $\mathbb{N}$ have direct analogs in $\mathbb{Z}$. In fact, except for complications involved with keeping track of signs, the properties are identical. In particular, the Division Algorithm and the Fundamental Theorem of Arithmetic hold in $\mathbb{Z}$; the statements are obvious modifications of the ones in $\mathbb{N}$. (In algebraic language, $\mathbb{Z}$ is a *Euclidean Domain*.)

III.1.6.7. Theorem. [Division Algorithm] Let $n, d \in \mathbb{Z}$, $d \neq 0$. Then there are unique $q, r \in \mathbb{Z}$ such that $0 \leq r < |d|$ and $n = qd + r$.

See III.2.9.4. for the proof if $n, d \in \mathbb{N}$. The other cases are nearly identical. We say d *divides* n, written $d|n$, if $r = 0$, i.e. if $n = qd$ for some $q \in \mathbb{Z}$; q is called the *quotient* of n by d, denoted $\frac{n}{d}$. If d divides n, then d also divides $-n$, and $-d$ divides $\pm n$. Every nonzero integer divides 0. If $n, d \in \mathbb{N}$, then $q \in \mathbb{N}$, and $q, d \leq n$, with equality for one if and only if the other is 1. If m and n each divide the other, then $m = \pm n$. If d divides both m and n, say $m = ad$ and $n = bd$ for $a, b \in \mathbb{Z}$, then $m + n = ad + bd = (a + b)d$, so d divides $m + n$. Conversely, if d divides m and $m + n$, then d divides $(m + n) - m = n$.

With $\mathbb{Z}$ at our disposal, there is a more efficient approach to the theory of prime numbers and factorization than can be done staying strictly within $\mathbb{N}$.

III.1.6.8. DEFINITION. If $m, n \in \mathbb{N}$, the number $d \in \mathbb{N}$ is the *greatest common divisor* for m, n if $d|m$, $d|n$, and $c|d$ for any $c \in \mathbb{N}$ which divides both m and n. Write $d = gcd(m, n)$ or $(q) = (m, n)$ (*ideal notation*).

This definition can be made for general nonzero elements of $\mathbb{Z}$, but is cleaner if we stick to $\mathbb{N}$. It is clear that if the greatest common divisor of m and n exists, it is unique and is indeed the largest natural number that divides both m and n, but it is not so clear that the largest such divisor is a greatest common divisor in the sense of the definition, and hence it is unclear that a greatest common divisor exists. However, this turns out to be true:

III.1.6.9. THEOREM. If m and n be natural numbers. Then m and n have a (unique) greatest common divisor, which is of the form $d = am + bn$ for some (nonunique) $a, b \in \mathbb{Z}$.

PROOF: Set

$$J = \{mx + ny : x, y \in \mathbb{Z}\}$$

(J is an *ideal* in $\mathbb{Z}$). Then $J \cap \mathbb{N} \neq \emptyset$ since $m, n \in J$. Let d be the smallest element of $J \cap \mathbb{N}$, and write $d = am + bn$ for $a, b \in \mathbb{Z}$.

We show that d divides m and n. In fact, d divides any $z \in J$: suppose $z = mx + ny \in J$, and write $z = qd + r$ by the Division Algorithm with $q \in \mathbb{Z}$ and $0 \leq r < d$. Then

$$mx + ny = q(ax + by) + r$$

$$r = (x - qa)m + (y - qb)n \in J$$

and by minimality of d we have that $r = 0$.

If c divides both m and n, then c clearly divides $am + bn = d$. 👍

III.1.6.10. There is an effective and numerically efficient algorithm (the *Euclidean algorithm* ()) for finding $gcd(m, n)$ for natural numbers m and n.

III.1.6.11. If $m, n \in \mathbb{N}$, we say m and n are *relatively prime* if $gcd(m, n) = 1$. (This definition, like the notion of greatest common divisor itself, can be extended to pairs of nonzero integers.)

III.1.6.12. PROPOSITION. Let $d, m, n \in \mathbb{N}$. If d divides mn and d and m are relatively prime, then d divides n.

PROOF: If d and m are relatively prime, then there are integers a and b with $ad + bm = 1$. Then

$$n = 1 \cdot n = (ad + bm)n = adn + bmn$$

so n is divisible by d. 👍

III.1.6.13. A $p \in \mathbb{N}$ is a *prime number* if $p > 1$ and the only natural numbers which divide p are 1 and p, i.e. the only integers dividing p are ± 1 and $\pm p$. (To be consistent with the language of abstract algebra, we should also call $-p$ prime, and we should use the term *irreducible* instead of *prime* until III.1.6.14. is proved.)

If $p, n \in \mathbb{N}$ and p is prime, then either p divides n or p and n are relatively prime. In particular, if p and q are distinct prime numbers, then p and q are relatively prime. We also have the following corollary of III.1.6.12. (the last part of the statement is proved by induction from the first part):

III.1.6.14. Corollary. Let $p, m, n \in \mathbb{N}$ with p prime. If p divides mn, then p divides m or p divides n. If $n_1, \ldots, n_r \in \mathbb{N}$ and p divides $n_1 n_2 \cdots n_r$, then p divides at least one of the n_k.

We can then state the Fundamental Theorem of Arithmetic:

III.1.6.15. Theorem. [Fundamental Theorem of Arithmetic] Every natural number can be written as a (finite) product of prime numbers, uniquely up to the order of the factors. (The "empty product" 1 and products of length 1 are included. Repeated factors are, of course, allowed.)

Proof: It has already been shown (III.1.5.7.) that every natural number can be written as a product of prime numbers, so we only need to show uniqueness. This can be done by induction on the length of the product, but we phrase the argument alternately by infinite regress.

Suppose

$$n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$$

is a nonunique factorization with r the smallest possible length for one of the factorizations (hence $r \leq s$). Then $n > 1$, so both r and s are at least 1, and neither can be 1 since the p_j and q_k are prime. Thus $r, s > 1$. We have that p_1 divides $q_1 \cdots q_s$, so it divides q_k for some k, and since q_k is prime we have $p_1 = q_k$. By reordering the q's we may assume $k = 1$. Thus

$$\frac{n}{p_1} = p_2 \cdots p_r = q_2 \cdots q_s$$

is a nonunique factorization with one product of length $r - 1$, contradicting minimality of r. Thus nonunique factorizations do not exist. 👍

We conclude our brief introduction to the algebraic theory of $\mathbb{N}$ by noting that there are an infinite number of primes. This observation and its slick proof are due to Euclid. (Actually Euclid never stated that there are infinitely many primes; the ancient Greeks had no notion of infinity.)

III.1.6.16. Theorem. There are infinitely many prime numbers. If $n \in \mathbb{N}$, there is a prime number p with $p > n$.

Proof: Suppose $p_1, \ldots, p_r$ is a finite set of prime numbers. Set $m = 1 + p_1 p_2 \cdots p_r$. Then no p_k divides m, so if p is a prime dividing m, p is not one of the p_k and no finite list of primes is complete.

The last statement follows from the first and the fact that the set of predecessors of n is finite. As an alternate proof, the number $n! + 1$ (cf. Exercise III.1.14.4.) is not divisible by any natural number $\leq n$ except 1, so any prime number dividing it is greater than n. 👍

III.1.7. The Rational Numbers

The next step historically in building the number system was the rational numbers, usually denoted $\mathbb{Q}$ (for "quotient"). Actually at least some simple rational numbers probably predated negative numbers historically.

III.1.7.1. $\mathbb{Q}$ is the smallest subset of $\mathbb{R}$ containing $\mathbb{Z}$ and closed under quotients of nonzero elements.

III.1.7.2. If $a, b \in \mathbb{R}$ and $b \neq 0$, then $\frac{a}{b}$ is defined to be ab^{-1}. The formulas

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}; \quad \frac{a}{b}\frac{c}{d} = \frac{ac}{bd}; \quad \left(\frac{b}{d}\right)^{-1} = \frac{d}{b}; \quad \frac{a/b}{c/d} = \frac{ad}{bc}$$

for $a, b, c, d \in \mathbb{R}$, $b \neq 0$, $d \neq 0$ (and, in the last, $c \neq 0$), provable from the field axioms, show that

$$\mathbb{Q} = \left\{ \frac{m}{n} : m, n \in \mathbb{Z}, n \neq 0 \right\}$$

and that $\mathbb{Q}$ is the smallest subset of $\mathbb{R}$ containing $\mathbb{Z}$ and closed under multiplication and reciprocals of nonzero elements; the first formula also shows that $\mathbb{Q}$ is closed under addition. Since $-\frac{m}{n} = \frac{-m}{n} = \frac{m}{-n}$ for any m, n ($n \neq 0$), we also have that $\mathbb{Q}$ is closed under subtraction, and that any element of $\mathbb{Q}$ can be written as $\frac{m}{n}$, where $m \in \mathbb{Z}$, $n \in \mathbb{N}$.

III.1.7.3. Rational numbers are often described as "fractions," but this is incorrect. Rational numbers are not the same thing as fractions: many fractions such as $\frac{1}{2}$ and $\frac{2}{4}$ can represent the same rational number. Instead, a rational number is a number which can be represented as a fraction. Although each rational number does have a standard representation as a fraction in "lowest terms" (III.1.7.4.), when a rational number appears in mathematics it often does not have any *natural* representation as a fraction (e.g. it could arise as a repeating decimal).

III.1.7.4. Since for $n, r \neq 0$ we have $\frac{mr}{nr} = \frac{m}{n}\frac{r}{r}$ and by definition $\frac{r}{r} = 1$, we have $\frac{mr}{nr} = \frac{m}{n}$. Thus, if $m, n \in \mathbb{N}$ and $d = gcd(m, n)$, and $m = ad$, $n = bd$, we have $\frac{m}{n} = \frac{a}{b}$, and a and b are relatively prime. By the Fundamental Theorem of Arithmetic, the representation of a positive rational number as a quotient of relatively prime natural numbers is unique. (Actually it just follows immediately from III.1.4.12. that any positive rational number has a unique representation as $\frac{a}{b}$ with $b \in \mathbb{N}$ minimal, and this must be in lowest terms, i.e. a and b have no common factors. The fact that this is the unique representation as $\frac{m}{n}$ with m and n relatively prime is not quite as elementary.) Negative rational numbers have a unique representation as $\frac{-a}{b}$, where a and b are relatively prime natural numbers.

III.1.7.5. The set $\mathbb{Q}$ of rational numbers with its algebraic operations, and the ordering induced as a subset of $\mathbb{R}$, satisfies all the axioms of $\mathbb{R}$ we have discussed so far, i.e. it is an ordered field. (In fact it is the "smallest" ordered field: any ordered field contains a smallest subfield, which is isomorphic to $\mathbb{Q}$ as an ordered field.) It is, in fact, not so easy to see (but not terribly difficult; see e.g. III.1.12.2.–III.1.12.6.) that $\mathbb{Q}$ is not all of $\mathbb{R}$, i.e. that there exist irrational numbers. The axiom which distinguishes $\mathbb{R}$ from $\mathbb{Q}$ (and all other ordered fields) is the Completeness Axiom.

III.1.8. The Number Line

The geometric representation of the real numbers as a line is familiar, at least informally, to anyone who has studied analytic geometry or even high-school algebra. The number line should be viewed merely as a convenient picture of the real numbers and not as a rigorous mathematical construction, since it is logically either baseless or circular depending on whether, or how, the relevant geometric objects are defined.

III.1.8.1. DESCARTES was the one to formalize the notion that points on a line can be identified with [real] numbers, but, at least in retrospect, the basic idea really goes back at least to EUCLID (and probably to predecessors such as EUDOXUS). One of the essential features of Euclidean geometry (the name in Greek refers to "measuring the earth") is that any two points on a line, or in space, have a numerical distance between them, at least once a fixed unit distance has been specified. EUCLID actually refers to "magnitudes" of lines [line segments] instead of numerical distances between points, but discusses equality, addition, subtraction, multiplication, and division of magnitudes, so his magnitudes correspond naturally to [constructible] nonnegative real numbers; indeed, the principal conceptual advance of DESCARTES' time over ancient Greece was the abstraction of the notion of "magnitude" or "quantity" into "number". (However, the conception of a line as a collection of points is a subtle one, and was not widely accepted in ancient Greece, or indeed universally accepted even today.) The only thing missing from EUCLID then, besides the abstract concept of number (and perhaps the conception of a line as a collection of points), is the idea of arbitrarily fixing one point on the line and calling it 0, and a unit distance and positive direction, which can be simultaneously specified by giving one additional point on the line to be called +1; once this is done, identifying all other points on the line with numbers, and vice versa, via distance and direction from 0, is almost immediate. (Identifying points in a plane or three-dimensional space with ordered pairs or triples of numbers is not quite so straightforward: not only must an origin and unit distance be chosen, but also a set of perpendicular axes and an orientation.)

III.1.8.2. To do the identification carefully (or at least systematically), we fix a line L and arbitrarily fix a point on L which we call 0. Then fix another point on L which we call 1. We usually draw L as a horizontal line, with 1 to the right of 0. Successively mark off equally spaced points to the right of 1 (i.e. the same direction from 0 as 1) and call these 2, 3, ... Do the same to the left (the opposite direction from 1) of 0, and call these numbers -1, -2, ... In this way, the numbers in $\mathbb{Z}$ are identified with points on L. See Figure (). (We often mark points on the number line L by short cross segments, but these are of course not part of L.)

III.1.8.3. Now, for each $m \in \mathbb{N}$, choose $m-1$ equally spaced points between 0 and 1, and call them $\frac{1}{m}$, $\frac{2}{m}$, ..., $\frac{m-1}{m}$ from left to right. Similarly, for each $n \in \mathbb{Z}$ choose $m-1$ equally spaced points between n and $n+1$, and call them $n+\frac{1}{m}$, ..., $n+\frac{m-1}{m}$ from left to right. Thus each rational number is associated with a point on L, and the ordering on $\mathbb{Q}$ is the same as the geometric ordering on L from left to right. We can thus picture $\mathbb{Q}$ as a subset of L. Although it is not immediately obvious whether every point of L corresponds to a rational number (we will see that actually "most" do not), the subset corresponding to $\mathbb{Q}$ is *dense* in L, i.e. every subinterval of L, no matter how short, contains a point of $\mathbb{Q}$, in fact infinitely many. See Figure (). (Actually, density of $\mathbb{Q}$ in the number line, which seems geometrically obvious, takes some proof; see ().)

III.1.9. Suprema, Infima, and the Completeness Axiom

III.1.9.1. One plausible assumption about a line is that it has no "holes". A way of expressing this is that any time a line L is split into two disjoint subsets A and B, with every point in A to the left of (or, in the usual order, less than) every point of B, then there should be a point of L exactly at the boundary of A and B, i.e. either A should have a largest (rightmost) point or B should have a smallest (leftmost) point. (Both things cannot happen simultaneously, for then the largest point of A and the smallest point of B would be consecutive points on L, which clearly is intuitively impossible.) See Figure (). The term used for this property, called the *completeness property*, is that a line is *connected*, or a *continuum*. It should be emphasized that the completeness property of a line cannot be proved – it is just a reasonable assumption, although a rather subtle one.

To rephrase this property in terms of the real numbers, we need some terminology:

III.1.9.2. Definition. If A is a nonempty subset of $\mathbb{R}$, a number $b \in \mathbb{R}$ is an *upper bound* for A if $a \leq b$ for all $a \in A$. If A has an upper bound, then A is *bounded above*. A number $c \in \mathbb{R}$ is a *lower bound* for A if $c \leq a$ for all $a \in A$. If A has a lower bound, then A is *bounded below*. A is *bounded* if A is bounded above and bounded below.

The real number s is the *least upper bound*, or *supremum*, of the nonempty subset A, written $s = \sup(A)$ (or sometimes $s = \text{lub}(A)$), if s is an upper bound for A and $s \leq b$ for any upper bound b of A. The *greatest lower bound*, or *infimum*, of A, written $\inf(A)$ or $\text{glb}(A)$, is defined analogously.

III.1.9.3. Note that an upper or lower bound for a set A is not necessarily an element of A. Not every subset of $\mathbb{R}$ is bounded above or below: for example, $A = \mathbb{R}$ has no upper or lower bound in $\mathbb{R}$. If A is bounded above, there is nothing unique about an upper bound for A: if b is an upper bound for A, so is b' for any $b' > b$. Similarly, lower bounds are not unique. But the supremum of A, if it exists, is unique; similarly, the infimum of A is unique if it exists. But note that it is not obvious that a set which is bounded above has a supremum, or that a set which is bounded below has an infimum.

We do not define upper or lower bounds, or the supremum or infimum, of the empty set $\emptyset$. (Technically, if Definition III.1.9.2. were applied to $A = \emptyset$, every real number would be both an upper bound and a lower bound for $\emptyset$.)

III.1.9.4. Examples. (i) If a subset A of $\mathbb{R}$ has a maximum b, then $b = \sup(A)$ and A is bounded above. If A has a minimum c, then $c = \inf(A)$ and A is bounded below. In particular, every finite subset of $\mathbb{R}$ has a supremum and an infimum, which are its largest and smallest elements respectively.

(ii) Let A be the open interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$. Then A is bounded above, and $1 = \sup(A)$ (to see this, note that if $b < 1$, then there is an $x \in \mathbb{R}$ with $b < x < 1$); but A has no maximum element, since $1 \notin A$. Similarly, $0 = \inf(A)$, but A has no minimum.

(iii) It is intuitively clear that $0 = \inf\{\frac{1}{n} : n \in \mathbb{N}\}$, but this requires proof; cf. ().

(iv) If $A \subseteq \mathbb{R}$, set $-A = \{-x : x \in A\}$. Then $-b$ is a lower bound for $-A$ if and only if b is an upper bound for A. So $-A$ is bounded below if and only if A is bounded above, and $-A$ has an infimum if and only if A has a supremum, with $\inf(-A) = -\sup(A)$. Thus the upper and lower situations are completely symmetric under taking negatives.

III.1.9.5. PROPOSITION. Let A be a nonempty subset of $\mathbb{R}$, and $s \in \mathbb{R}$. The following are equivalent:

(i) $s = \sup(A)$.

(ii) $x \leq s$ for all $x \in A$, and for any $b \in \mathbb{R}$, $b < s$, there is an $x \in A$ with $x > b$.

(iii) $x \leq s$ for all $x \in A$, and for every $\epsilon \in \mathbb{R}$, $\epsilon > 0$, there is an $x \in A$ with $x > s - \epsilon$.

PROOF: (i) $\Rightarrow$ (ii): If $s = \sup(A)$ and $b < s$, then b is not an upper bound for A.
(ii) $\Rightarrow$ (i): Under (ii), s is an upper bound for A, and no $b < s$ is an upper bound for A.
(ii) $\Longleftrightarrow$ (iii): Given b as in (ii), set $\epsilon = s - b$; given ϵ as in (iii), set $b = s - \epsilon$. 👍

III.1.9.6. PROPOSITION. Let A be a nonempty subset of $\mathbb{R}$, and $c \in \mathbb{R}$. Set $c + A = \{c + x : x \in A\}$. Then $c + A$ is bounded above if and only if A is bounded above, and $\sup(c + A) = c + \sup(A)$. Similarly, $c + A$ is bounded below if and only if A is bounded below, and $\inf(c + A) = c + \inf(A)$.

PROOF: If b is an upper bound for A, then $c + b$ is an upper bound for $c + A$; conversely, if d is an upper bound for $c + A$, then $d - c$ is an upper bound for A. If $s = \sup(A)$, then it is clear that $c + s$ is the least upper bound of $c + A$, and conversely. 👍

We can then phrase the Completeness Axiom as follows. This turns out to be equivalent to the completeness property of the real line as described above, but is easier to work with.

III.1.9.7. **Completeness Axiom:** Every nonempty subset of $\mathbb{R}$ which is bounded above has a supremum.

We will show that the Completeness Axiom is equivalent to the completeness property of the real line, as described in III.1.9.1.. We first make a more general observation:

III.1.9.8. PROPOSITION. Let A and B be nonempty subsets of $\mathbb{R}$, with $a \leq b$ for all $a \in A$, $b \in B$. Then A is bounded above, B is bounded below, and $\sup(A) \leq \inf(B)$.

PROOF: Every $b \in B$ is an upper bound for A. So A is bounded above, and so has a supremum $\sup(A)$ which satisfies $\sup(A) \leq b$ for each $b \in B$ since $\sup(A)$ is the *least* upper bound of A. This means $\sup(A)$ is a lower bound for B (as is each $a \in A$), so B is bounded below, and $\inf(B)$ exists and satisfies $\inf(B) \geq \sup(A)$ since $\inf(B)$ is the *greatest* lower bound of B. 👍

We can then show using the Completeness Axiom that the real line has the completeness property. This is sometimes (cf. [Lan51]) called *Dedekind's Fundamental Theorem*, although in our approach it hardly merits Theorem status since it is almost immediate (in other approaches to the real numbers it is a real theorem).

III.1.9.9. THEOREM. If A and B are nonempty disjoint subsets of $\mathbb{R}$ with $A \cup B = \mathbb{R}$ and $x < y$ for all $x \in A$ and $y \in B$, then either A has a largest element or B has a smallest element, but not both. Alternatively, there is a unique real number a such that $\{x \in \mathbb{R} : x < a\} \subseteq A$ and $\{y \in \mathbb{R} : a < y\} \subseteq B$ (a may be either the largest element of A or the smallest element of B).

PROOF: If such an a exists, it is both an upper bound for A and a lower bound for B, and conversely, a number a which is both an upper bound for A and a lower bound for B satisfies the conclusions. So we need to show that there is a unique a with these properties.

By III.1.9.8., $\sup(A)$ and $\inf(B)$ exist and satisfy $\sup(A) \leq \inf(B)$. We cannot have $\sup(A) < \inf(B)$, since then $c = (\sup(A) + \inf(B))/2$ would satisfy $\sup(A) < c < \inf(B)$ and c could not be in either A or B. Thus $\sup(A) = \inf(B)$; call this real number a. If b is an upper bound for A, then $a \leq b$, and if d is a lower bound for B, then $d \leq a$. Thus a is the unique number with both properties. 👍

III.1.9.10. Conversely, if the real line is assumed to have the completeness property of III.1.9.1., then the Completeness Axiom can be proved, and thus the Completeness Axiom is equivalent to the completeness property. For suppose C is a nonempty subset of $\mathbb{R}$ which is bounded above. We wish to show from the completeness property that C has a supremum. If C has a largest element, the conclusion is obvious. If C has no largest element, i.e. if no element of C is an upper bound for C, let B be the set of all upper bounds for C in $\mathbb{R}$, and let A be the set of real numbers which are not upper bounds for C. Then $\mathbb{R}$ is the disjoint union of A and B. B is nonempty by assumption, and A is nonempty since $C \subseteq A$. We clearly have $x < y$ for all $x \in A$, $y \in B$. Let a be the unique element of $\mathbb{R}$ given by the completeness property. If $a \in A$, then a would be an upper bound for A and hence for C, contradicting the definition of A; hence $a \in B$ and a is an upper bound for C. If $x < a$, then $x \in A$, so x is not an upper bound for C; hence a is the least upper bound of C.

III.1.9.11. The Completeness Axiom is really the cornerstone of analysis. Most of the basic concepts of analysis, e.g. convergence of sequences, limits and continuity of functions, and derivatives, can be equally well defined in the same way in any ordered field such as $\mathbb{Q}$, and many of the standard results about them (e.g. algebraic combinations of limits, rules of differentiation) hold in this setting with the same proofs as in $\mathbb{R}$. But the objects have very different properties from the real case. For example, working on $\mathbb{Q}$, define

$$f(x) = \begin{cases} -1 & \text{if} \quad x^2 < 2 \\ 1 & \text{if} \quad x^2 > 2 \end{cases} .$$

Then f is continuous and even differentiable everywhere, and f' is identically 0, but f is not constant and does not satisfy the Intermediate Value Theorem. If

$$g(x) = \begin{cases} x & \text{if} \quad x^2 < 2 \\ 0 & \text{if} \quad x^2 > 2 \end{cases}$$

then g is continuous on $[-2, 2]$, but does not satisfy the Extreme Value Theorem. Thus these properties of continuous or differentiable functions on $\mathbb{R}$, which may seem "obvious," are in fact rather subtle since they depend on the Completeness Axiom.

These points are argued persuasively in [Kör04]. The book begins by asking, "Why do we bother?" and then states:

> It is surprising how many people think that analysis consists in the difficult proofs of obvious theorems. All we need know, they say, is what a limit is, the definition of continuity and the definition of the derivative. All the rest is 'intuitively clear.' "

This notion is then well refuted. There is also an interesting discussion of the question, "Is the Intermediate Value Theorem obvious?" In that book, the working definition of analysis is "mathematics based on the Completeness Axiom" (called the fundamental axiom); everything else (of the mathematics covered there) is regarded as "mere algebra."

Intervals in $\mathbb{R}$

Using the Completeness Axiom, we can give a clean description of intervals in $\mathbb{R}$.

III.1.9.12. DEFINITION. A subset J of $\mathbb{R}$ is an *interval* if

(i) J contains more than one real number.

(ii) If $x, y \in J$, $x < y$, and $z \in \mathbb{R}$, $x < z < y$, then $z \in J$.

Informally, an interval is a subset of $\mathbb{R}$ containing all numbers between any two of its elements. We do not regard $\emptyset$ or singleton sets to be intervals, although they technically satisfy (ii), but we sometimes call these sets *degenerate intervals*.

III.1.9.13. EXAMPLES. Here are standard examples of intervals, and the standard interval notation used:

Let $a, b \in \mathbb{R}$, $a < b$. Set

$$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$$
$$[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$$
$$(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$$
$$(a, b) = \{x \in \mathbb{R} : a < x < b\}$$

If $a \in \mathbb{R}$, set

$$[a, \infty) = [a, +\infty) = \{x \in \mathbb{R} : x \geq a\}$$
$$(a, \infty) = (a, +\infty) = \{x \in \mathbb{R} : x > a\}$$

If $b \in \mathbb{R}$, set

$$(-\infty, b] = \{x \in \mathbb{R} : x \leq b\}$$
$$(-\infty, b) = \{x \in \mathbb{R} : x < b\}$$

Finally, set $(-\infty, \infty) = (-\infty, +\infty) = \mathbb{R}$.

Intervals of the form $[a, b]$, $[a, \infty)$, and $(-\infty, b]$ are called *closed intervals*, and intervals of the form (a, b), (a, ∞), and $(-\infty, b)$ are called *open intervals*. The interval $(-\infty, \infty)$ is regarded as both a closed interval and an open interval. An interval of the form $[a, b)$ or $(a, b]$ is called a *half-open interval*.

III.1.9.14. PROPOSITION. Every interval J in $\mathbb{R}$ is of one of the forms from III.1.9.13.. More specifically, J is of the form $[a,b]$, $[a,b)]$, $(a,b]$, or (a,b) if J is bounded, of the form $[a,\infty)$ or (a,∞) if J is bounded below but not bounded above, of the form $(-\infty,b]$ or $(-\infty,b)$ if J is bounded above but not bounded below, and $J=(-\infty,\infty)$ if J is not bounded above or below.

PROOF: Suppose J is bounded. Set $a=\inf(J)$ and $b=\sup(J)$. Then $a<b$ since J is not a singleton, and $J\subseteq[a,b]$. If $z\in\mathbb{R}$, $a<z<b$, then z is neither an upper or lower bound for J, so there are $x,y\in J$ with $x<z<y$. But then $z\in J$ by (ii). So $(a,b)\subseteq J$. The argument if J is unbounded is almost identical. 👍

III.1.9.15. Without the Completeness Axiom, we could not assert that every interval is of such a simple form. For example, we can formally define intervals in $\mathbb{Q}$ the same way. But then

$$\{x\in\mathbb{Q}:x^2<2\}$$

can be easily shown to be an interval in $\mathbb{Q}$; but even though it is a nonempty proper subset of $\mathbb{Q}$, it has no endpoints (in $\mathbb{Q}$) and so cannot be described as either an open or closed interval in the usual way.

Lengths of Intervals

An important consequence of the characterization of intervals in $\mathbb{R}$, sometimes not properly appreciated, is the fact that each interval has a well-defined numerical length:

III.1.9.16. DEFINITION. Let I be an interval in $\mathbb{R}$. If I is not bounded, set $\ell(I)=+\infty$. If I is bounded, then I is of the form $[a,b]$, $[a,b)]$, $(a,b]$, or (a,b) for some $a,b\in\mathbb{R}$, $a<b$, and we set $\ell(I)=b-a$. Then $\ell(I)$, called the *length of* I, satisfies $0<\ell(I)\leq+\infty$.

III.1.9.17. Without the characterization of intervals, it would not be possible to give a numerical length to intervals in general. For example, if our number system were $\mathbb{Q}$, the interval described in III.1.9.15. would not have a well-defined numerical length within our number system.

The following two properties of interval length seem like (and really are) almost no-brainers, but are surprisingly powerful, eventually making possible a nontrivial theory of Lebesgue measure on $\mathbb{R}$ ().

III.1.9.18. PROPOSITION. (i) Let I and J be intervals in $\mathbb{R}$. If $J\subseteq I$, then $\ell(J)\leq\ell(I)$.
(ii) Let I and J be intervals in $\mathbb{R}$. If $I\cup J$ is an interval (e.g. if $I\cap J\neq\emptyset$), then

$$\ell(I\cup J)\leq\ell(I)+\ell(J)\ .$$

III.1.9.19. THEOREM. [FUNDAMENTAL THEOREM OF INTERVAL LENGTH] Let $I_1, \ldots, I_n, J$ be intervals in $\mathbb{R}$. If $J \subseteq I_1 \cup \cdots \cup I_n$, then

$$\ell(J) \leq \sum_{k=1}^{n} \ell(I_k) \ .$$

PROOF: By induction on n. The case $n = 1$ is III.1.9.18.(i). Suppose the statement is true for n, and let $J \subseteq I_1 \cup \cdots \cup I_{n+1}$. We may assume I_1 is bounded, $I_1 \cap J \neq \emptyset$, and $J \not\subseteq I_1$. Let a_1 and b_1 be the left and right endpoints of I_1. Then either a_1 or b_1 is in J. Suppose $a_1 \in J$. (1) If $a_1 \notin I_1$, then there is an I_k with $a_1 \in I_k$ and $I_1 \cup I_k$ is an interval. (2) If $a_1 \in I_1$ and J contains a $c < a_1$, then there is one of the intervals I_k with endpoints a_k, b_k with $a_k < a_1$ and $b_k \geq a_1$ since there are only finitely many I_k [if not, $d = \max\{b_k : a_k < a_1\} < a_1$ and $x \in J \setminus \cup I_k$ for $d < x < a_1$]; then also $I_1 \cup I_k$ is an interval. The case where $b_1 \in J$ is similar. We must be in either case (1) or (2) for at least one of a_1 and b_1 since $J \not\subseteq I_1$.

Thus there must be an I_k such that $I_1 \cup I_k$ is an interval. By renumbering, we may assume $k = n + 1$. Set $I_1' = I_1 \cup I_{n+1}$. Then $J \subseteq I_1' \cup I_2 \cup \cdots \cup I_n$, and $\ell(I_1') \leq \ell(I_1) + \ell(I_{n+1})$ by III.1.9.18.(ii). Thus, by the inductive hypothesis,

$$\ell(J) \leq \ell(I_1') + \sum_{k=2}^{n} \ell(I_k) \leq \sum_{k=1}^{n+1} \ell(I_k) \ .$$



III.1.10. The Archimedean Property and Density of $\mathbb{Q}$

Using the Completeness Axiom, we now show that the Archimedean property (cf. ()) also holds in $\mathbb{R}$:

III.1.10.1. THEOREM. Let $a, b \in \mathbb{R}$ with $a > 0$. Then there is an $n \in \mathbb{N}$ with $b < na$.

PROOF: Suppose $na \leq b$ for all $n \in \mathbb{N}$, i.e. b is an upper bound for $A = \{na : n \in \mathbb{N}\}$. Then A is bounded above, so by the Completeness Axiom A has a supremum. Set $s = \sup(A)$. Then $s - a$ is not an upper bound for A since $a > 0$, so there is an $n \in \mathbb{N}$ with $na > s - a$. But then $(n + 1)a \in A$ and $(n + 1)a = na + a > (s - a) + a = s$, contradicting the fact that s is an upper bound for A. Thus the assumption that $na \leq b$ for all $n \in \mathbb{N}$ is false.

III.1.10.2. COROLLARY. Let $b \in \mathbb{R}$. Then there is an $n \in \mathbb{N}$ with $b < n$. ($\mathbb{N}$ is not bounded above in $\mathbb{R}$.)

PROOF: Apply III.1.10.1. with $a = 1$.

III.1.10.3. COROLLARY. $\inf\left\{\frac{1}{n} : n \in \mathbb{N}\right\} = 0$, i.e. if $\epsilon \in \mathbb{R}$, $\epsilon > 0$, there is an $n \in \mathbb{N}$ with $\frac{1}{n} < \epsilon$.

PROOF: 0 is a lower bound for $\left\{\frac{1}{n} : n \in \mathbb{N}\right\}$, so $\inf\left\{\frac{1}{n} : n \in \mathbb{N}\right\}$ exists. If $\epsilon \in \mathbb{R}$, $\epsilon > 0$, then there is an $n \in \mathbb{N}$ with $\frac{1}{\epsilon} < n$. For this n, $\frac{1}{n} < \epsilon$. Thus ϵ is not a lower bound for $\left\{\frac{1}{n} : n \in \mathbb{N}\right\}$. 

There is a tightening of III.1.10.2. which is an important observation:

III.1.10.4. PROPOSITION. Let $x \in \mathbb{R}$. then there is a unique $n \in \mathbb{Z}$ with $n \leq x < n+1$. The integer n is called the *integer part* of x, denoted $\lfloor x \rfloor$.

PROOF: Suppose first $x \geq 0$. Then $\{n \in \mathbb{N} : x < n\}$ is nonempty; let m be the smallest element of this set, and set $n = m - 1$. Then $n \leq x < n + 1$. If $x < 0$, there is an $r \in \mathbb{N}$ such that $r \leq -x < r + 1$, so $-(r+1) < x \leq -r$. If $x = -r$, set $n = -r$; otherwise set $n = -(r+1)$.

The n is unique since if $n \leq x < n+1$ and $m \leq x < m+1$, we have $-(m+1) < -x \leq -m$, so $n - (m+1) < 0 < (n+1) - m$, $-1 < m - n < 1$, $m - n = 0$. 👍

We then get the density of $\mathbb{Q}$ in $\mathbb{R}$:

III.1.10.5. THEOREM. Let $a, b \in \mathbb{R}$, $a < b$. Then there is an $r \in \mathbb{Q}$ with $a < r < b$.

In fact, by applying the theorem repeatedly to the intervals (a, b), (a, r), (r, b), etc., it is readily shown that there are infinitely many rational numbers in the interval (a, b).

PROOF: By III.1.10.3., there is an $n \in \mathbb{N}$ with $\frac{1}{n} < b - a$. For this n, there is an $m \in \mathbb{Z}$ with $m - 1 \leq na < m$. We have $a < \frac{m}{n}$. Also, $m - na \leq 1$, $m \leq na + 1$, so

$$\frac{m}{n} \leq a + \frac{1}{n} < a + (b - a) = b \ .$$

👍

III.1.10.6. COROLLARY. Let $a \in \mathbb{R}$, and set $A = \{r \in \mathbb{Q} : r < a\}$. Then $\sup(A) = a$.

PROOF: A is nonempty ($-n \in A$ for some $n \in \mathbb{N}$), and a is an upper bound for A, so $s = \sup(A)$ exists and satisfies $s \leq a$. If $s < a$, then there is an $r \in \mathbb{Q}$ with $s < r < a$. But then $r \in A$, contradicting that s is an upper bound for A. Thus $s = a$. 👍

III.1.11. Uniqueness of $\mathbb{R}$

We can now prove that the real numbers are uniquely determined by the axioms, i.e. that any two structures satisfying the axioms for $\mathbb{R}$ are the same up to a renaming of the elements. Specifically:

III.1.11.1. THEOREM. Let $\mathbb{R}$ and $\mathbb{R}'$ be sets with addition, multiplication, and order satisfying the algebraic, order, and completeness axioms. Then there is a unique function $\phi : \mathbb{R} \to \mathbb{R}'$ which is one-to-one and onto, such that for all $x, y \in \mathbb{R}$,

$$\phi(x + y) = \phi(x) + \phi(y)$$

$$\phi(xy) = \phi(x)\phi(y)$$

$$\phi(x) < \phi(y) \text{ if and only if } x < y$$

(Such a ϕ is called an *isomorphism* of the ordered fields $\mathbb{R}$ and $\mathbb{R}'$, and will automatically satisfy $\phi(0) = 0'$ and $\phi(1) = 1'$.)

The proof will be done in a number of steps. Although the argument is long, the steps are quite straightforward. To justify the last comment, and motivate the beginning of the construction, note that any such ϕ must satisfy

$$\phi(0) + \phi(1) = \phi(0 + 1) = \phi(1) = 0' + \phi(1)$$

and hence $\phi(0) = 0'$ by additive cancellation in $\mathbb{R}'$. Similarly,

$$\phi(1) \cdot \phi(1) = \phi(1 \cdot 1) = \phi(1) = 1' \cdot \phi(1)$$

and, since $\phi(1) \neq \phi(0) = 0'$ because ϕ is one-to-one, we have $\phi(1) = 1'$ by multiplicative cancellation.

III.1.11.2. Now note that $\mathbb{R}$ and $\mathbb{R}'$ (in fact, any ordered field) each contain a "copy" (model) of the rational numbers. Let $\mathbb{N}$ be the smallest subset of $\mathbb{R}$ containing its multiplicative identity 1 and closed under adding 1, $\mathbb{Z}$ the smallest subset of $\mathbb{R}$ containing $\mathbb{N}$ and closed under subtraction, and $\mathbb{Q}$ the smallest subset of $\mathbb{R}$ containing $\mathbb{Z}$ and closed under division by nonzero elements; similarly, let $\mathbb{N}'$ be the smallest subset of $\mathbb{R}'$ containing its multiplicative identity $1'$ and closed under adding $1'$, and $\mathbb{Z}'$ and $\mathbb{Q}'$ the corresponding subsets of $\mathbb{R}'$ closed under subtraction and division.

III.1.11.3. We will define $\phi : \mathbb{R} \to \mathbb{R}'$ and its inverse $\psi : \mathbb{R}' \to \mathbb{R}$ in stages. We begin by defining ϕ and ψ inductively (recursively) on $\mathbb{N}$ and $\mathbb{N}'$. Define $\phi(1) = 1'$, and if $\phi(n)$ has been defined, define $\phi(n + 1) = \phi(n) + 1'$. (We should write $+'$, $\cdot'$, $<'$ for the algebraic operations and order relation in $\mathbb{R}'$, but we will not do so to avoid becoming overprimed.) This defines ϕ on all of $\mathbb{N}$, with range contained in $\mathbb{N}'$. Similarly, define $\psi : \mathbb{N}' \to \mathbb{N}$ inductively by $\psi(1') = 1$ and $\psi(n' + 1') = \psi(n') + 1$.

III.1.11.4. Lemma. We have $\psi(\phi(n)) = n$ and $\phi(\psi(n')) = n'$ for all $n \in \mathbb{N}$, $n' \in \mathbb{N}'$. Thus ϕ and ψ are mutually inverse bijections between $\mathbb{N}$ and $\mathbb{N}'$.

Proof: This is an easy proof by induction. We have $\psi(\phi(1)) = \psi(1') = 1$. If $\psi(\phi(n)) = n$, then

$$\psi(\phi(n + 1)) = \psi(\phi(n) + 1') = \psi(\phi(n)) + 1 = n + 1$$

so $\psi(\phi(n)) = n$ for all $n \in \mathbb{N}$. The other formula is proved similarly by induction on $\mathbb{N}'$. 👍

III.1.11.5. Lemma. If $n, m \in \mathbb{N}$, then $\phi(n + m) = \phi(n) + \phi(m)$ and $\phi(nm) = \phi(n)\phi(m)$.

Proof: For a fixed $n \in \mathbb{N}$, we prove the formula by induction on m. We have $\phi(n + 1) = \phi(n) + 1' = \phi(n) + \phi(1)$, so the formula holds for $m = 1$. If $\phi(n + m) = \phi(n) + \phi(m)$, then

$$\phi(n+(m+1)) = \phi((n+m)+1) = \phi(n+m)+1' = (\phi(n)+\phi(m))+1' = \phi(n)+(\phi(m)+1') = \phi(n)+\phi(m+1)$$

so the formula holds for all m. The formula for multiplication is proved similarly. 👍

III.1.11.6. LEMMA. If $n, m \in \mathbb{N}$, then $\phi(n) < \phi(m)$ if and only if $n < m$.

PROOF: We have $n < m$ if and only if there is a $p \in \mathbb{N}$ with $m = n + p$. Then

$$\phi(m) = \phi(n + p) = \phi(n) + \phi(p)$$

so $\phi(n) < \phi(m)$. The argument also works for ψ, giving the converse. 👍

III.1.11.7. We now extend ϕ and ψ to $\mathbb{Z}$ and $\mathbb{Z}'$. If $x \in \mathbb{Z}$, write $x = n - m$ for $n, m \in \mathbb{N}$, and define $\phi(x) = \phi(n) - \phi(m)$. This is well defined since if $n - m = q - p$, i.e. $n + p = m + q$, we have

$$\phi(n) + \phi(p) = \phi(n + p) = \phi(m + q) = \phi(m) + \phi(q)$$

and thus $\phi(q) - \phi(p) = \phi(n) - \phi(m)$. Alternatively, write $\mathbb{Z}$ as the disjoint union of $\mathbb{N}$, $\{0\}$ and $-\mathbb{N}$, and $\mathbb{Z}'$ as the disjoint union of $\mathbb{N}'$, $\{0'\}$, and $-\mathbb{N}'$, and define $\phi(0) = 0'$ and $\phi(-n) = -\phi(n)$ for $n \in \mathbb{N}$. Define $\psi : \mathbb{Z}' \to \mathbb{Z}$ in the same manner. It is obvious, especially from the second definition, that ψ is the inverse for ϕ, so ϕ and ψ are bijections between $\mathbb{Z}$ and $\mathbb{Z}'$.

The next result is just a simple but somewhat tedious calculation checking case-by-case:

III.1.11.8. LEMMA. If $x, y \in \mathbb{Z}$, then $\phi(x+y) = \phi(x)+\phi(y)$ and $\phi(xy) = \phi(x)\phi(y)$. We have $\phi(x) < \phi(y)$ if and only if $x < y$.

III.1.11.9. Now we extend ϕ and ψ to $\mathbb{Q}$ and $\mathbb{Q}'$. If $x \in \mathbb{Q}$, write $x = \frac{m}{n}$ for $m, n \in \mathbb{Z}$, $n \neq 0$, and set $\phi(x) = \frac{\phi(m)}{\phi(n)} \in \mathbb{Q}'$. This makes sense since $\phi(n) \neq 0$. Then ϕ is well defined since, if $\frac{m}{n} = \frac{p}{q}$, i.e. $mq = np$, we have

$$\phi(m)\phi(q) = \phi(mq) = \phi(np) = \phi(n)\phi(p)$$

so $\frac{\phi(m)}{\phi(n)} = \frac{\phi(p)}{\phi(q)}$. Similarly define $\psi : \mathbb{Q}' \to \mathbb{Q}$. It is clear that ψ is the inverse of ϕ, so ϕ and ψ are bijections.

III.1.11.10. LEMMA. If $x, y \in \mathbb{Q}$, then $\phi(x + y) = \phi(x) + \phi(y)$ and $\phi(xy) = \phi(x)\phi(y)$. We have $\phi(x) < \phi(y)$ if and only if $x < y$.

PROOF: Write $x = \frac{m}{n}$, $y = \frac{p}{q}$ with $m, p \in \mathbb{Z}$ and $n, q \in \mathbb{N}$. Then

$$\phi(x + y) = \phi\left(\frac{m}{n} + \frac{p}{q}\right) = \phi\left(\frac{mq + np}{nq}\right) = \frac{\phi(m)\phi(q) + \phi(n)\phi(p)}{\phi(n)\phi(q)} = \frac{\phi(m)}{\phi(n)} + \frac{\phi(p)}{\phi(q)} = \phi(x) + \phi(y)$$

$$\phi(xy) = \phi\left(\frac{m}{n} \cdot \frac{p}{q}\right) = \phi\left(\frac{mp}{nq}\right) = \frac{\phi(mp)}{\phi(nq)} = \frac{\phi(m)\phi(p)}{\phi(n)\phi(q)} = \frac{\phi(m)}{\phi(n)} \cdot \frac{\phi(p)}{\phi(q)} = \phi(x)\phi(y) \ .$$

If $x < y$, then $mq < np$, so

$$\phi(m)\phi(q) = \phi(mq) < \phi(np) = \phi(n)\phi(p)$$

so we have $\phi(x) = \frac{\phi(m)}{\phi(n)} < \frac{\phi(p)}{\phi(q)} = \phi(y)$, and the argument is reversible. 👍

III.1.11.11. It is clear from the requirement that $\phi(0) = 0'$ and $\phi(1) = 1'$ that the definition of ϕ and its inverse ψ between $\mathbb{N}$ and $\mathbb{N}'$ can only be made in the manner of III.1.11.3., i.e. there are *unique* functions ϕ and ψ between $\mathbb{N}$ and $\mathbb{N}'$ with the specified properties. The extensions to $\mathbb{Z}$ and $\mathbb{Q}$ are then forced by the properties of ϕ needed, i.e. the uniqueness statement applies to functions from $\mathbb{Q}$ to $\mathbb{Q}'$ also.

III.1.11.12. So far, we have only used the algebraic and order axioms, so the arguments work for any ordered fields: if F and F' are ordered fields, then the smallest subfields of F and F' are naturally isomorphic (and isomorphic to $\mathbb{Q}$). Thus we may naturally regard $\mathbb{Q}$ as a subfield of any ordered field.

Now we use the Completeness Axiom and density of $\mathbb{Q}$ in $\mathbb{R}$, and density of $\mathbb{Q}'$ in $\mathbb{R}'$, to extend ϕ and ψ to $\mathbb{R}$ and $\mathbb{R}'$.

III.1.11.13. Definition. Let $x \in \mathbb{R}$. Let $\phi(x) = \sup\{\phi(r) : r \in \mathbb{Q}, r < x\}$. If $x' \in \mathbb{R}'$, define $\psi(x') = \sup\{\psi(r') : r' \in \mathbb{Q}', r' < x'\}$.
It follows immediately from III.1.10.6. applied in both $\mathbb{R}$ and $\mathbb{R}'$ that if $x \in \mathbb{Q}$, $x' \in \mathbb{Q}'$, this definition of $\phi(x)$ and $\psi(x')$ agrees with the previous one, and that ψ is the inverse of ϕ in general, so ϕ and ψ are mutually inverse bijections between $\mathbb{R}$ and $\mathbb{R}'$.

III.1.11.14. Lemma. If $x, y \in \mathbb{R}$, then $x < y$ if and only if $\phi(x) < \phi(y)$.

Proof: Suppose $x < y$. Let $t \in \mathbb{Q}$, $x < t < y$. Then $\phi(t) \leq \phi(y)$ by definition of $\phi(y)$, and $\phi(t) \neq \phi(y)$ since ϕ is injective, so $\phi(t) < \phi(y)$. Also, $\phi(r) < \phi(t)$ for every $r < x$, so $\phi(t)$ is an upper bound for $\{\phi(r) : r \in \mathbb{Q}, r < x\}$ and therefore $\phi(x) \leq \phi(t)$ (even $\phi(x) < \phi(t)$ since ϕ is injective). So $\phi(x) < \phi(y)$ by transitivity. The same argument works for ψ, giving the converse. 👍

III.1.11.15. Corollary. If $x \in \mathbb{R}$, then $\phi(x) = \inf\{\phi(r) : r \in \mathbb{Q}, x < r\}$.

Proof: Set $b' = \inf\{\phi(r) : r \in \mathbb{Q}, x < r\}$. Since $\phi(x)$ is a lower bound for this set, $\phi(x) \leq b'$. If $\phi(x) < b'$, let $r' \in \mathbb{Q}'$ with $\phi(x) < r' < b'$. Then $r' = \phi(r)$ for some $r \in \mathbb{Q}$, $x < r$, contradicting that b' is a lower bound. Thus $b' = \phi(x)$. 👍

III.1.11.16. Corollary. If $x \in \mathbb{R}$, then $\phi(-x) = -\phi(x)$.

It remains to show that ϕ preserves addition and multiplication.

III.1.11.17. LEMMA. Let $x, y \in \mathbb{R}$. Then $\phi(x+y) = \phi(x) + \phi(y)$ and $\phi(xy) = \phi(x)\phi(y)$.

PROOF: Suppose $a \in \mathbb{Q}$. Then, if $r \in \mathbb{Q}$, we have $a + r < a + y$ if and only if $r < y$, so

$$\phi(a+y) = \sup\{\phi(a+r) : r \in \mathbb{Q}, r < y\} = \sup\{\phi(a) + \phi(r) : r \in \mathbb{Q}, r < y\} \ .$$

By III.1.9.6., the last supremum equals

$$\phi(a) + \sup\{\phi(r) : r \in \mathbb{Q}, r < y\} = \phi(a) + \phi(y) \ .$$

Then we have

$$\phi(a) + \phi(y) = \phi(a+y) < \phi(x+y)$$

for every $a \in \mathbb{Q}$, $a < x$, and

$$\phi(x+y) < \phi(b+y) = \phi(b) + \phi(y)$$

for every $b \in \mathbb{Q}$, $x < b$. But by III.1.11.15. and III.1.9.6., we have

$$\phi(x) + \phi(y) = \sup\{\phi(a) + \phi(y) : a \in \mathbb{Q}, a < x\} = \inf\{\phi(b) + \phi(y) : b \in \mathbb{Q}, x < b\}$$

and so $\phi(x+y) = \phi(x) + \phi(y)$.

The proof that $\phi(xy) = \phi(x)\phi(y)$ is very similar if x and y are positive. The proof for general x, y then follows from this and III.1.11.16.. 

III.1.11.18. By the density of $\mathbb{Q}$ in $\mathbb{R}$ and of $\mathbb{Q}'$ in $\mathbb{R}'$, the values of ϕ and ψ are completely determined by their values on $\mathbb{Q}$ and $\mathbb{Q}'$ respectively. Since the functions between $\mathbb{Q}$ and $\mathbb{Q}'$ are unique (III.1.11.11.), we obtain the uniqueness of the function ϕ (and ψ) with the specified properties.

This completes the proof of Theorem III.1.11.1.. 

III.1.12. Existence of Roots

Although we will later get more general results, we show here using the Completeness Axiom that every positive real number has arbitrary roots. We first need an estimate:

III.1.12.1. PROPOSITION. Let $a, b \in \mathbb{R}$. Set $c = \max(|a|, |b|)$. If $n \in \mathbb{N}$, then

$$|b^n - a^n| \leq nc^{n-1}|b - a| \ .$$

PROOF: This is easily proved using the Mean Value Theorem, but there is an elementary proof: it follows immediately from the identity (for $n > 1$)

$$b^n - a^n = (b-a)\sum_{k=0}^{n-1} a^k b^{n-k-1}$$

which is easily verified by an algebraic calculation. Then we have, by the triangle inequality,

$$|b^n - a^n| = |b-a| \left| \sum_{k=0}^{n-1} a^k b^{n-k-1} \right| \leq |b-a| \sum_{k=0}^{n-1} |a|^k |b|^{n-k-1} \leq |b-a| \cdot nc^{n-1} \ .$$

III.1.12.2. THEOREM. Let $a \in \mathbb{R}$, $a \geq 0$, and $n \in \mathbb{N}$. Then there is a unique $x \in \mathbb{R}$, $x \geq 0$, with $x^n = a$.

PROOF: If there is a solution, it is unique since by the identity in the proof of III.1.12.1., if $0 < x < y$, then $x^n < y^n$. If $a = 0$, the unique solution is $x = 0$. For existence, it suffices to consider the case $0 < a < 1$ since if $a > 1$, so $\frac{1}{a} < 1$, and $x^n = \frac{1}{a}$, then $\left(\frac{1}{x}\right)^n = a$, and the case $a = 1$ is obvious.

Set

$$A = \{r \in \mathbb{R} : r > 0, r^n < a\} \ .$$

A is nonempty since there is an $k \in \mathbb{N}$ with $\frac{1}{k} < a$, and for this k we have $\frac{1}{k^n} \leq \frac{1}{k} < a$. Also, A is bounded: 1 is an upper bound for A. Thus $x = \sup(A)$ exists, and $x \leq 1$.

We claim $x^n = a$. First suppose $x^n > a$, and choose $k \in \mathbb{N}$ with $\frac{1}{k} < x$ and $\frac{1}{k} < \frac{x^n - a}{n}$. Then, by III.1.12.1.,

$$x^n - \left(x - \frac{1}{k}\right)^n \leq nx^{n-1}\frac{1}{k} \leq n\frac{1}{k} < x^n - a$$

and so $\left(x - \frac{1}{k}\right)^n > a$ and $x - \frac{1}{k}$ is an upper bound for A, contradicting that x is the least upper bound of A. Thus $x^n \leq a$.

So $x < 1$ since $a < 1$. If $x^n < a$, choose $k \in \mathbb{N}$ so that $x + \frac{1}{k} < 1$ and $\frac{1}{k} < \frac{a - x^n}{n}$. Then, by III.1.12.1.,

$$\left(x + \frac{1}{k}\right)^n - x^n \leq n\left(x + \frac{1}{k}\right)^{n-1} \cdot \frac{1}{k} < n\frac{1}{k} < a - x^n$$

so $\left(x + \frac{1}{k}\right)^n < a$, contradicting that x is an upper bound for A. Thus $x^n = a$.

If n is odd, then negative real numbers have n'th roots too:

III.1.12.3. COROLLARY. Let $n \in \mathbb{N}$ be odd, and $a \in \mathbb{R}$. Then there is a unique $x \in \mathbb{R}$ with $x^n = a$.

PROOF: If $a \geq 0$, existence comes from III.1.12.2.. If $a < 0$, then by III.1.12.2. there is an $x \in \mathbb{R}$ with $x^n = -a$. Then, since n is odd, $(-x)^n = -x^n = a$. Uniqueness follows as in the beginning of the proof of III.1.12.2..

III.1.12.4. If n is even and $a < 0$, there is no solution to $x^n = a$ in $\mathbb{R}$, since if $n = 2m$, then $x^n = (x^m)^2 \geq 0$ for any $x \in \mathbb{R}$ (). If $a > 0$, there are two solutions, one positive and one negative. As usual, we will denote the positive solution as $\sqrt[n]{a}$. If n is odd, then for any a we denote the unique solution by $\sqrt[n]{a}$. When $n = 2$ (square root), we just write $\sqrt{a}$ for the nonnegative root when $a \geq 0$.

III.1.12.5. So every positive real number has a square root. In particular, there is a real number $\sqrt{2}$ whose square is 2. Since there is no rational number whose square is 2 (I.6.5.8.), $\sqrt{2}$ is a real number which is not rational (an *irrational* number). It turns out that "most" real numbers are irrational ().

The classical proof that $\sqrt{2}$ is irrational is given in I.6.5.8.. Here is a more general result; the special case $d = 2$ gives another pretty proof for $\sqrt{2}$ which is even simpler than the classic one.

III.1.12.6. THEOREM. If $d \in \mathbb{N}$, then $\sqrt{d}$ is either a natural number or irrational.

PROOF: Suppose $\sqrt{d}$ is not a natural number. Since $\sqrt{d} > 1$ (because $(\sqrt{d})^2 > 1^2$; cf. III.1.2.3.(vii)), there is a $q \in \mathbb{N}$ with $q < \sqrt{d} < q + 1$, i.e. $0 < \sqrt{d} - q < 1$ (III.1.10.4.). (If $d = 2$, then $q = 1$.)

Now suppose $\sqrt{d}$ is rational. Then there is a natural number n such that $n\sqrt{d}$ is an integer, i.e.

$$A = \{n \in \mathbb{N} : n\sqrt{d} \in \mathbb{Z}\}$$

is nonempty. By III.1.4.12., A has a smallest element n. But then

$$m := (\sqrt{d} - q)n = n\sqrt{d} - nq \in \mathbb{Z}$$

and $m > 0$ since $\sqrt{d} - q > 0$; hence $m \in \mathbb{N}$. Also,

$$m\sqrt{d} = (d - q\sqrt{d})n = dn - q(n\sqrt{d}) \in \mathbb{Z}$$

so $m \in A$. But $m < n$ since $\sqrt{d} - q < 1$, contradicting that n is the smallest element of A. 👍

This proof is due to DEDEKIND. See III.1.14.7. for another version of the proof, and III.9.5.4. for a generalization.

III.1.12.7. One can significantly extend III.1.12.3. using the notions of limits and continuity, and the Intermediate Value Theorem. For example, it can be easily shown (V.2.3.6.) that every polynomial of odd degree with real coefficients has a root in $\mathbb{R}$ (III.1.12.3. is the special case of the polynomial $x^n - a$).

III.1.12.8. An ordered field in which every positive element has a square root and every polynomial of odd degree has a root is called a *real-closed field*. There are many other equivalent characterizations of real-closed fields. The theory of real-closed fields turns out to be exactly the first-order theory of $\mathbb{R}$, i.e. any statement true in $\mathbb{R}$ which has a first-order formulation (not allowing quantification over subsets) is true in every real-closed field. There are many real-closed fields other than $\mathbb{R}$, including countable examples (actually, the field of real algebraic numbers is a countable example). Other examples include any model for the hyperreal numbers of nonstandard analysis (), and the Levi-Civita numbers. See () for a description of the theory of real-closed fields.

III.1.13. The Extended Real Line

It is often convenient to add "points at infinity" at each end of the real line. We can think of these as "ideal points" which are not really there, or as actual objects which are just not part of the real line itself.

III.1.13.1. We denote by $+\infty$ and $-\infty$ two additional points which are added at the right and left "end" of the real line respectively. The *extended real line* (or *extended real numbers*) is the set $\mathbb{R} \cup \{+\infty, -\infty\}$.

It is common to just write ∞ for $+\infty$, and we will occasionally do this, but the symbol ∞ is somewhat ambiguous (in more advanced mathematics, a line is often completed into a circle by adding a single point at infinity; cf. also ()), so we will try to stick to $+\infty$ for clarity, with a few exceptions, even when it seems to be overkill.

Ordering on the Extended Real Line

III.1.13.2. We extend the ordering on $\mathbb{R}$ to the extended real line by setting $-\infty < x < +\infty$ for all $x \in \mathbb{R}$ (and, to make the ordering transitive, $-\infty < +\infty$).

With this ordering, we can talk about intervals in the extended real numbers. In fact, we have already used $\pm\infty$ in our standard notation for unbounded intervals in $\mathbb{R}$ (III.1.9.13.). We can now also consider intervals of the form $[a, +\infty]$, $[-\infty, b]$, etc., in the extended real line. We can even write $[-\infty, +\infty]$ for the entire extended real line.

III.1.13.3. One important property of the extended real numbers as an ordered set is that every nonempty subset has a supremum and an infimum: if A is a nonempty subset of $[-\infty, +\infty]$, then

$\sup(A) = +\infty$ if $+\infty \in A$ or if $A \cap \mathbb{R}$ is nonempty and not bounded above (in $\mathbb{R}$).

$\sup(A) = b$ if $+\infty \notin A$ and $A \cap \mathbb{R}$ is nonempty and bounded above, with least upper bound $b \in \mathbb{R}$.

$\sup(A) = -\infty$ if $A = \{-\infty\}$.

We similarly define $\inf(A)$. Thus, for example, $\sup(\mathbb{N}) = \sup(\mathbb{Z}) = +\infty$ and $\inf(\mathbb{Z}) = -\infty$.

Algebraic Operations and Infinity

III.1.13.4. It must *always* be kept in mind that $\pm\infty$ are *not* numbers and that ordinary algebraic operations with them are not defined. But we can extend some (not all!) algebraic operations to expressions involving $\pm\infty$, although not all properties of the operations are valid for the extensions.

III.1.13.5. We make the following extensions of addition, multiplication, and division:

$a + (+\infty) = (+\infty) + a = +\infty$ for any $a \in \mathbb{R}$.

$(+\infty) + (+\infty) = +\infty$.

$a + (-\infty) = (-\infty) + a = -\infty$ for any $a \in \mathbb{R}$.

$(-\infty) + (-\infty) = -\infty$.

$a \cdot (+\infty) = (+\infty) \cdot a = +\infty$ for all $a \in \mathbb{R}$, $a > 0$.

$a \cdot (+\infty) = (+\infty) \cdot a = -\infty$ for all $a \in \mathbb{R}$, $a < 0$.

$a \cdot (-\infty) = (-\infty) \cdot a = -\infty$ for all $a \in \mathbb{R}$, $a > 0$.

$a \cdot (-\infty) = (-\infty) \cdot a = +\infty$ for all $a \in \mathbb{R}$, $a < 0$.

$(+\infty) \cdot (+\infty) = +\infty$.

$(+\infty) \cdot (-\infty) = (-\infty) \cdot (+\infty) = -\infty$.

$(-\infty) \cdot (-\infty) = +\infty$.

$\frac{+\infty}{a} = +\infty$ for all $a \in \mathbb{R}$, $a > 0$.

$\frac{+\infty}{a} = -\infty$ for all $a \in \mathbb{R}$, $a < 0$.

$\frac{-\infty}{a} = -\infty$ for all $a \in \mathbb{R}$, $a > 0$.

$\frac{-\infty}{a} = +\infty$ for all $a \in \mathbb{R}$, $a < 0$.

These extended operations on the extended real line are associative and commutative, and distributive when defined. However, cancellation laws do *not* hold, e.g. $a + (+\infty) = b + (+\infty)$ does not imply $a = b$.

III.1.13.6. The following operations are *not* defined:

$(+\infty) + (-\infty)$ and $(-\infty) + (+\infty)$ (usually called $\infty - \infty$).

$0 \cdot \pm\infty$ and $\pm\infty \cdot 0$.

$\frac{\pm\infty}{\pm\infty}$.

One can run into serious problems trying to perform or even make sense of these expressions. Some expressions from calculus seem to have these forms in the limit; these expressions are called *indeterminate forms* and careful calculus methods must be used to make sense of them. See ().

One can define $\frac{a}{\pm\infty} = 0$ for all $a \in \mathbb{R}$ without getting in too much trouble. But there are some pitfalls which must be avoided with such expressions, and they are of limited use, so we will generally refrain from using them.

III.1.14. Exercises

III.1.14.1. Show that in the field axioms for $\mathbb{R}$ (III.1.1.2.), in the presence of axioms (A1) and (A2) (and an assumption that $\mathbb{R} \neq \emptyset$), axioms (A3) and (A4) can be replaced by

(A5) For every $x, y \in \mathbb{R}$, there is a $z \in \mathbb{R}$ with $x + z = y$.

Suppose (A1), (A2), and (A5) hold. If $x \in \mathbb{R}$, there is a $z \in \mathbb{R}$ such that $x + z = x$. For this z, $y + z = y$ for every $y \in \mathbb{R}$. [Write $y = x + w$ for some $w \in \mathbb{R}$ and use (A1) and (A2).] Conclude that (A3) holds. Then show (A4) holds. Conversely, show that if (A1)–(A4) hold, then (A5) holds.

Using (A5) in place of (A3) and (A4) is arguably a more efficient axiomatization for fields. But the axioms are conventionally written using (A3) and (A4) since there are other algebraic structures satisfying (A3) but not (A4) (and sometimes not (A2)), and it is nice to have a uniform set of algebraic axioms. Similarly, (M3)-(M4) can be replaced by a single axiom (which is not quite as clean), but this is even less desirable since in rings (M3) often holds without (M4) (and often without (M2) either).

III.1.14.2. Prove by induction on n that if $\{x_k, y_k : 1 \leq k \leq n\}$ is a set of real numbers, then

$$\sum_{k=1}^{n}(x_k + y_k) = \sum_{k=1}^{n} x_k + \sum_{k=1}^{n} y_k \ .$$

III.1.14.3. (a) Let $x \in \mathbb{R}$. Show how to define x^n for each $n \in \mathbb{N}$ by recursion.
(b) Prove by induction that if $x, y \in \mathbb{R}$ and $m, n \in \mathbb{N}$, then $(xy)^n = x^n y^n$, $x^m x^n = x^{m+n}$, $(x^m)^n = x^{mn}$, and, if $y \neq 0$, $(y^n)^{-1} = (y^{-1})^n$, $\left(\frac{x}{y}\right)^n = \frac{x^n}{y^n}$ and $\frac{y^m}{y^n} = y^{m-n}$. Extend these rules of exponents to arbitrary integer powers, with appropriate hypotheses.

III.1.14.4. Show how to carefully define $n! = 1 \cdot 2 \cdot 3 \cdots n$ for all $n \in \mathbb{N}$ by recursion.

III.1.14.5. Define the *Fibonacci sequence* (x_n) by $x_1 = 1$, $x_2 = 1$, $x_n = x_{n-1} + x_{n-2}$ for $n > 2$.
(a) Show that this definition can be phrased as a proper recursive definition under II.4.2.16..
(b) Prove by induction that any two consecutive terms in the sequence are relatively prime.
(c) Rephrase the argument of (b) as an Infinite Regress argument. Which way is simpler?

III.1.14.6. Carefully prove the assertions in III.1.7.2. and III.1.7.4..

III.1.14.7. [CG96] Give an alternate proof of III.1.12.6., as follows. Compare with the given proof of III.1.12.6..
(a) Suppose $\sqrt{d}$ is not an integer, but is rational. Write $\sqrt{d} = \frac{a}{b}$ in lowest terms (e.g. with b minimal in $\mathbb{N}$), with $a, b \in \mathbb{N}$. Then also $\sqrt{d} = \frac{d}{\sqrt{d}} = \frac{db}{a}$.
(b) Since $\frac{db}{a} = \frac{a}{b}$, the fractional parts (which are nonzero since $\sqrt{d} \notin \mathbb{N}$) are equal. Thus there are $p, q \in \mathbb{N}$ with $p < a$ and $q < b$ and $\frac{p}{a} = \frac{q}{b}$.
(c) From $\frac{p}{a} = \frac{q}{b}$ we get $\frac{p}{q} = \frac{a}{b} = \sqrt{d}$, contradicting that $\frac{a}{b}$ is in lowest terms.
(d) If r is a positive rational number and $r = \frac{a}{b}$ with b minimal in $\mathbb{N}$, then the fraction $\frac{a^n}{b}$ is also the representation of $a^{n-1}r$ with minimal denominator in $\mathbb{N}$ (cf. III.1.7.4.).
(e) Using (d), prove that if $d, n \in \mathbb{N}$, then $\sqrt[n]{d}$ is either an integer or irrational (cf. III.9.5.4.). [Use that $\sqrt[n]{d} = \frac{d}{(\sqrt[n]{d})^{n-1}}$.]

III.1.14.8. The real numbers are characterized up to isomorphism by the field, order, and completeness axioms (III.1.11.1.), and have the Archimedean property (III.1.10.1.). Conversely, prove the following (cf. III.10.1.4.):

THEOREM. (i) Let $\mathbb{F}$ be an Archimedean ordered field. Then there is a unique order-preserving isomorphism of $\mathbb{F}$ with a subfield of $\mathbb{R}$.
(ii) Let $\mathbb{F}$ be a maximal Archimedean ordered field (i.e. $\mathbb{F}$ cannot be embedded in a strictly larger Archimedean ordered field). Then there is a unique isomorphism of $\mathbb{F}$ with $\mathbb{R}$.

Thus $\mathbb{R}$ is characterized by the field, order, and Archimedean axioms, along with a maximality axiom.

III.2. The Natural Numbers

We now begin the construction of the number systems from scratch, beginning with the natural numbers.

III.2.1. Introduction

III.2.1.1. The first set of numbers developed, going back to prehistoric times, is the *natural numbers* or *counting numbers*, also described as the positive integers. This set is usually called $\mathbb{N}$. Thus, as it is often described or defined,

$$\mathbb{N} = \{1, 2, 3, \dots\} .$$

Unfortunately, this "description" is totally inadequate as a definition of $\mathbb{N}$, even an informal one, for two reasons. Most importantly, it is entirely unclear what the "..." means (if you think you know, try giving a precise formulation of the meaning!) And even if this point is specified, this description says nothing about the operations of arithmetic which are an essential part of the structure.

III.2.1.2. One mathematically minor point needs to be addressed. There is no general agreement among mathematicians about whether $\mathbb{N}$ includes 0; some mathematicians include it, and others do not. We will not. Set theorists generally include it, since from the point of view of the set-theoretic construction of $\mathbb{N}$ it is natural to do so. However, in my opinion 0 should *not* be called a "natural number." The positive integers historically far predate the emergence of the concept of 0, and there are important societies, and even some modern-day societies, that never developed 0; the notion of 0 is far more abstract and subtle than that of numbers such as 1 and 2. (On the other hand, I acknowledge that it is also historically questionable whether "large" positive integers should be regarded as "natural numbers," since there is a practical limit to the size of numbers which occur "naturally," and in addition the development of notation which allows one to even contemplate numbers such as $10^{10^{10}}$ is relatively recent compared even to the concept of 0.) However, it is useful to have a notation for the set $\mathbb{N} \cup \{0\}$ of nonnegative integers; we will use the notation $\mathbb{N}_0$ for this set. The question of whether to work with $\mathbb{N}$ or $\mathbb{N}_0$ as the basic set is strictly a matter of taste, since it is mathematically trivial to go from either to the other.

III.2.2. Axiomatic Characterization of $\mathbb{N}$

III.2.2.1. Rather than trying to define what $\mathbb{N}$ *is*, it is more useful to list a set of basic properties that $\mathbb{N}$ has, from which all the usual structure can be derived, and then to show there is an essentially unique mathematical structure having these properties. The basic axiom scheme was first developed by G. Peano in 1889, although the actual construction of $\mathbb{N}$ along these lines which motivated the axiomatization had already been done by R. Dedekind.

Peano Axioms for $\mathbb{N}$:

(P1) There is a natural number 1.

(P2) Every natural number n has a *successor* $S(n)$.

(P3) 1 is not the successor of any natural number.

(P4) If $m, n \in \mathbb{N}$ and $S(m) = S(n)$, then $m = n$.

(P5) (Axiom of Induction) If $A \subseteq \mathbb{N}$, $1 \in A$, and $n \in A$ implies $S(n) \in A$, then $A = \mathbb{N}$.

The Axiom of Induction can be equivalently rephrased: "The smallest subset of $\mathbb{N}$ containing 1 and closed under taking successors is all of $\mathbb{N}$." To make precise the notion of "smallest," note that if $\{A_i\}$ is a collection of subsets of $\mathbb{N}$ each containing 1 and closed under taking successors, then $\cap_i A_i$ also contains 1 and is closed under taking successors. Thus the intersection of all subsets which contain 1 and are closed under taking successors is clearly the smallest such set.

III.2.2.2. This theory is technically a *second-order theory* since in the Axiom of Induction, there is quantification over subsets of $\mathbb{N}$. There is a corresponding first-order theory, but it is not as strong and has infinitely many axioms replacing the Axiom of Induction. See books on mathematical logic and model theory, such as (), for an explanation of first-order Peano arithmetic and the difference between first-order and second-order theories.

Here are some simple observations. The first is a "converse" to axiom (P3):

III.2.2.3. PROPOSITION. If $n \in \mathbb{N}$, $n \neq 1$, then there is an $m \in \mathbb{N}$ with $n = S(m)$.

PROOF: Let A be the set of all $n \in \mathbb{N}$ such that either $n = 1$ or there is an $m \in \mathbb{N}$ with $n = S(m)$. We trivially have $1 \in A$. If $n \in A$ (actually, if just $n \in \mathbb{N}$), then obviously $S(n) \in A$. By axiom (P5), $A = \mathbb{N}$. 👍

III.2.2.4. PROPOSITION. $\mathbb{N}$ is infinite (in the sense of Dedekind (II.5.2.1.).)

PROOF: The function $S : \mathbb{N} \to \mathbb{N}$ is injective by axiom (P4), but not surjective by axiom (P3). 👍

Sets and Models

> "The definitions of number are very numerous and of great variety ... We must not be surprised that there are so many. If any one of them was satisfactory we should not get any new ones."
>
> *H. Poincaré*[3]

There are some logical subtleties in regarding $\mathbb{N}$ as a "set", which we now discuss briefly.

III.2.2.5. DEFINITION. A *model* for Peano arithmetic is a triple (N, e, s), where N is a set, $e \in N$, and s is a function from N to N satisfying

(P3) $s(x) \neq e$ for every $x \in N$.

(P4) If $x, y \in N$ and $s(x) = s(y)$, then $x = y$ (i.e. s is injective).

(P5) If $A \subseteq N$, $e \in A$, and $x \in A$ implies $s(x) \in A$, then $A = N$.

(Of course, Peano axioms (P1) and (P2), suitably renamed, also hold tautologically in (N, e, s).)

[3] [?, p. 155].

III.2.2.6. There is a "standard" model for Peano arithmetic in ZF set theory or the equivalent, the set ω of finite ordinals. Ordinals are discussed more carefully in (), but here we briefly outline the construction of ω, which is originally due to J. VON NEUMANN. Begin with $\emptyset$, which plays the role of 1 (or, more commonly, 0 in $\mathbb{N}_0$; set theorists often call this set 0). Define the *successor* of α to be $S(\alpha) = \alpha \cup \{\alpha\}$. Thus the successor of $\emptyset$ is $S(\emptyset) = \{\emptyset\} = \{0\}$ (note that $\{\emptyset\} \neq \emptyset$!) Set theorists usually call this set 1. Then $S(1) = S(S(\emptyset)) = \{\emptyset, \{\emptyset\}\} = \{0, 1\}$ (called 2 in set theory). Continuing,

$$S(2) = S(S(S(\emptyset))) = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = \{0, 1, 2\}$$

(called 3 in set theory), etc. The Axiom of Infinity then gives the set ω.

III.2.2.7. This is a model for either first- or second-order Peano arithmetic. There is a logical subtlety in this model: it is open to question whether second-order Peano arithmetic can be constructed in ZF set theory, but it can. (There is a further logical subtlety in that it apparently cannot be proved that ZF set theory is consistent; cf. ().) We do not discuss these questions further, but simply refer the reader to books on mathematical logic such as (), and we agree not to be bothered by them from now on and accept ω as a valid model for $\mathbb{N}$.

III.2.2.8. Set theorists generally *define* $\mathbb{N}$ (or $\mathbb{N}_0$) to be ω. However, this leads to logical difficulties: technically, the natural number 1, the integer 1, the rational number 1, the real number 1, and the complex number 1 are all different objects (sets) as defined in set theory, and they must somehow, and somewhat awkwardly, be identified together (cf. III.2.2.9.).

> "If numbers are sets, then they must be *particular sets*, for each set is some particular set. But if the number 3 is really one set rather than another, it must be possible to give some cogent reason for thinking so; for the position that this is an unknowable truth is hardly tenable."
>
> *Paul Benacerraf*[4]

It is hard to conceive that the principles of logic entail that we must inevitably conclude that the number 3 (whether regarded as natural, integer, rational, real, or complex) must be the set $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}$, or any other specific set.

Another approach to defining natural numbers is frequently considered, reflecting their use as counting numbers (cardinal numbers): the number n is defined to be the class of all sets with exactly n elements (this can be carefully phrased in a non-circular manner, e.g. by fixing a model for $\mathbb{N}$ such as ω and taking n to be the class of all sets equipotent with the corresponding element of the model). Thus, for example, the number 1 is defined to be the class of all singleton sets. This approach has logical difficulties, since these objects cannot themselves be sets (e.g. there is an obvious one-one correspondence between the class of all singleton sets and the class of all sets), and anyway does not avoid the problem that there is no obvious way to extend the definition coherently to negative integers or rational numbers, let alone irrational numbers (attempts to do this logically, such as in [**?**], become unmanageably complicated and convoluted).

It is simpler, although perhaps less satisfying logically, not to precisely define $\mathbb{N}$ as a set. This approach will carry through also when the number system is extended, i.e. we will not define individual numbers to be any particular sets. This will not cause any difficulties in the actual use of the number systems in analysis.

[4] "What Numbers Could Not Be," *Philosophical Review* 74(1965), 47-73.

Anyone who is uncomfortable with this approach can be comforted by the knowledge that the number system *can*, in a somewhat arbitrary and awkward way, be precisely constructed within set theory.

F. Waismann [**?**, p. 116] compares the question, "What is a number?" to the question, "What is time?" and quotes St. Augustine:

> "If I am not asked, I know it; if I am asked, I do not know it."

However, the question of what a number actually *is* does not generally concern or even greatly interest mathematicians:

> "[Q]uestions about the true 'nature' of mathematical objects – such as whether $\sqrt{2}$ is a set – are always irrelevant to mathematical practice. In mathematics one is exclusively concerned with abstract relationships rather than with the 'internal constitution' of objects."
>
> *Martin Davis*[5]

III.2.2.9. Technically, what we can do is assume that within set theory we have a sufficient supply of atoms (), and whenever each new collection of numbers is defined as a set, a collection of additional atoms is designated to name those numbers not already named. The operations of set theory and arithmetic can then be transferred to this set of atoms. Thus we will regard $\mathbb{N}$ (and, later, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$) as sets, with $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$, but we will not regard individual numbers as sets.

The (second-order) Peano axioms uniquely determine the structure of $\mathbb{N}$, up to a simple renaming of its elements, in the following sense:

III.2.2.10. Theorem. Let (N, e, s) and (N', e', s') be models of the Peano axioms. Then there is a unique bijection $\phi : N' \to N$ with $\phi(e') = e$ and $\phi(s'(x)) = s(\phi(x))$ for all $x \in N'$. (Such a ϕ is called an *isomorphism* of (N', e', s') and (N, e, s).)

Proof: It suffices to assume that $(N', e', s') = (\omega, 0, S)$, where S is the ordinal successor function $S(\alpha) = \alpha \cup \{\alpha\}$ (). [If $\phi : \omega \to N$ and $\psi : \omega \to N'$ are successor-preserving bijections, then so is $\phi \circ \psi^{-1} : N' \to N$.] Inductively define a function ϕ from ω to N by $\phi(0) = e$ and $\phi(S(\alpha)) = s(\phi(\alpha))$ (). This function ϕ preserves successors by construction. If $R \subseteq N$ is the range of ϕ, then $e \in R$, and if $x = \phi(\alpha) \in R$, then so is $s(x) = \phi(S(\alpha))$. Thus $R = N$ by axiom (P5), and ϕ is surjective.

To show that ϕ is injective, let A be the set of all $x \in N$ such that x has a unique preimage. If $\phi(\alpha) = e$, then α must be 0, since otherwise $\alpha = S(\beta)$ for some β (III.2.2.3.), and $e = \phi(\alpha) = \phi(S(\beta)) = s(\phi(\beta))$, contradicting axiom (P3). Thus $e \in A$. Now suppose $x \in A$, and suppose $\alpha, \beta \in \omega$ with $\phi(\alpha) = \phi(\beta) = s(x)$. Then $\alpha = S(\sigma)$ and $\beta = S(\tau)$ for some $\sigma, \tau \in \omega$. We have

$$s(\phi(\sigma)) = \phi(S(\sigma)) = \phi(\alpha) = s(x) = \phi(\beta) = \phi(S(\tau)) = s(\phi(\tau))$$

and so $\phi(\sigma) = x = \phi(\tau)$ by axiom (P4). Thus $\sigma = \tau$ since x has unique preimage, and hence $\alpha = \beta$, i.e. $s(x) \in A$. Thus $A = N$ by axiom (P5) and ϕ is injective. 👍

There is no such theorem for first-order Peano arithmetic; in fact, by general facts about first-order logic (the Lowenheim-Skolem Theorem), there are models of first-order Peano arithmetic of arbitrary infinite cardinality (the nonstandard natural numbers ${}^*\mathbb{N}$ (III.10.3.7.) are one example).

[5] [Dav77, p. 11]

III.2.2.11. In particular, every model for Peano arithmetic is isomorphic to ω. Thus, by slight abuse of language, we can talk about *the* set of natural numbers even though we technically do not define $\mathbb{N}$ precisely within set theory (i.e. to *be* any specific model).

Since we have not defined $\mathbb{N}$ as a particular set, we must technically identify $\mathbb{N}$ with a specific model to define functions in or out; such definitions are independent of the choice of model in an evident way, so we will freely use $\mathbb{N}$ as a domain or range for functions as though it were an honest set (this is logically justifiable by the procedure of III.2.2.9.).

III.2.3. Induction and Recursion

We will not reiterate here the full content of () (indeed we cannot yet). Note that with the careful construction of $\mathbb{N}$, we can give a more direct and complete justification for the principles of induction and recursion. We do not yet have addition on $\mathbb{N}$, so we cannot phrase things using "$n + 1$" and we must instead use the successor function. We also do not yet have the ordering on $\mathbb{N}$ so we cannot phrase Complete Induction or Infinite Regress.

III.2.3.1. [Principle of Induction] Suppose $P(n)$ is a statement for each $n \in \mathbb{N}$. Suppose the following two statements hold:

(i) $P(1)$ is true.

(ii) For every $n \in \mathbb{N}$, if $P(n)$ is true, then $P(S(n))$ is true.

Then $P(n)$ is true for all $n \in \mathbb{N}$.

III.2.3.2. To see why the Principle of Induction holds, set

$$A = \{n \in \mathbb{N} : P(n) \text{ is true } \} .$$

We have that $1 \in A$ by (i), and if $n \in A$ we have $S(n) \in A$ by (ii). Thus $A = \mathbb{N}$.

Definitions by Induction or Recursion

We will write sequences, which are really functions on $\mathbb{N}$, using functional notation. The following theorem is not quite as general as II.4.2.16., but is the version we will use at present. The proof is similar to the proof of II.4.2.16., but somewhat more complicated since we have less machinery available.

III.2.3.3. Theorem. Let X be a set, and let g be a function from X to X. If x_1 is any element of X, then there is a unique function $f : \mathbb{N} \to X$ with $f(1) = x_1$ and $f(S(n)) = g(f(n))$ for all $n \in \mathbb{N}$.

Proof: Let B be the smallest subset of $\mathbb{N} \times X$ containing $(1, x_1)$, with the property that, whenever $(n, x) \in B$, then $(S(n), g(x)) \in B$. (B is the intersection of all subsets of $\mathbb{N} \times X$ with these two properties; there is at least one such subset, $\mathbb{N} \times X$ itself.)

We first show that for every $n \in \mathbb{N}$, there is at least one $x \in X$ with $(n, x) \in B$. Set

$$A = \{n \in \mathbb{N} : \exists x \in X \text{ with } (n, x) \in B\} .$$

Then $1 \in A$ since $(1, x_1) \in B$. If $n \in A$, and $(n, x) \in B$, then $(S(n), g(x)) \in B$ so $S(n) \in A$. Thus $A = \mathbb{N}$.

Next we show that B is a function, i.e. if (n,x) and (n,y) are in B, then $x=y$. Let

$$A' = \{n \in \mathbb{N} : (n,x) \in B, (n,y) \in B \text{ imply } x = y\} .$$

First we show $1 \in A'$. Set

$$B' = \{(1,x_1)\} \cup \{(n,x) \in B : n \neq 1\}$$

and note that $B' \subseteq B$, $(1,x_1) \in B'$, and $(n,x) \in B'$ implies $(S(n), g(x)) \in B'$; hence $B' = B$ and $1 \in A'$. Now suppose $n \in A'$, i.e. there is a unique $x_n \in X$ with $(n,x_n) \in B$. Set

$$B'' = \{(S(n), g(x_n))\} \cup \{(m,x) \in B : m \neq S(n)\}$$

and note that $B'' \subseteq B$, $(1,x_1) \in B''$, and $(m,x) \in B''$ implies $(S(m), g(x)) \in B''$ [if $m \neq n$ we have $S(m) \neq S(n)$ so the result is obvious; if $m = n$ then $x = x_n$ so the result holds too]. Thus $B'' = B$ and $S(n) \in A'$. So $A' = \mathbb{N}$ and B is a function, i.e. for each $n \in \mathbb{N}$ we can set $f(n) = x_n$, where x_n is the unique element of X such that $(n,x_n) \in B$.

We need to show that f is unique. If f' is another function from $\mathbb{N}$ to X with $f'(1) = x_1$ and $f'(S(n)) = g(f'(n))$ for all $n \in \mathbb{N}$, set

$$A'' = \{n \in \mathbb{N} : f(n) = f'(n)\} .$$

We have $1 \in A''$ since $f(1) = f'(1) = x_1$. Suppose $n \in A''$. Then

$$f(S(n)) = g(f(n)) = g(f'(n)) = f'(S(n))$$

so $S(n) \in A''$, and thus $A'' = \mathbb{N}$, $f = f'$. 👍

A version of this theorem is used in defining algebraic operations on $\mathbb{N}$:

III.2.3.4. Corollary. Let g be a function from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$, and h a function from $\mathbb{N}$ to $\mathbb{N}$. Then there is a unique function f from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$ satisfying $f(m,1) = h(m)$ and $f(m,S(n)) = g(m, f(m,n))$ for all $m,n \in \mathbb{N}$.

Proof: Fix $m \in \mathbb{N}$, and let g_m be the function from $\mathbb{N}$ to $\mathbb{N}$ given by $g_m(n) = g(m,n)$. Define a function $f_m : \mathbb{N} \to \mathbb{N}$ recursively by $f_m(1) = h(m)$ and $f_m(S(n)) = g_m(f_m(n))$. Set $f(m,n) = f_m(n)$ for each n. This gives a well-defined function $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$, and we have $f(m,1) = f_m(1) = h(m)$ and

$$f(m,S(n)) = f_m(S(n)) = g_m(f_m(n)) = g(m, f(m,n))$$

for $m,n \in \mathbb{N}$. Uniqueness is also obvious from the uniqueness of the f_m. 👍

III.2.4. Algebraic Structure of $\mathbb{N}$

There is no explicit mention of algebraic operations on $\mathbb{N}$ in the Peano axioms. But addition and multiplication can be defined inductively in terms of the successor function:

III.2.4.1. DEFINITION. Let $(\mathbb{N}, 1, S)$ be the natural numbers (or, technically, a model for the natural numbers). Define a function $+ : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ inductively (recursively) as follows:

(a) $+(m, 1) = S(m)$.

(b) If $m \in \mathbb{N}$, then $+(m, S(n)) = S(+(m, n))$.

This is done from III.2.3.4. by setting $h(m) = S(m)$ and $g(m, n) = S(n)$ for all m, n.
Then define a function $\cdot : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ by

(c) If $m \in \mathbb{N}$, then $\cdot(m, 1) = m$.

(d) If $m \in \mathbb{N}$, then $\cdot(m, S(n)) = +(m, \cdot(m, n))$ for all $n \in \mathbb{N}$.

This is done from III.2.3.4. by setting $h(m) = m$ and $g(m, n) = +(m, n)$ for all m, n.

We normally write $m+n$ for $+(m, n)$, and mn instead of $\cdot(m, n)$. We will suppress the notations $+(m, n)$ and $\cdot(m, n)$ from now on, and use $m + n$ and mn instead. We also adopt the standard convention that multiplication is grouped before addition, i.e. that $mn + r$ means $(mn) + r$, etc.

III.2.4.2. PROPOSITION. The functions $+$ and $\cdot$ are well defined on all of $\mathbb{N} \times \mathbb{N}$.

PROOF: This is an immediate consequence of III.2.3.4.. 👍

We can then prove the associative and commutative laws for addition.

III.2.4.3. THEOREM. The binary operation $+$ is associative and commutative on $\mathbb{N}$, i.e.
(i) [ASSOCIATIVE LAW] We have $(x + y) + z = x + (y + z)$ for all $x, y, z \in \mathbb{N}$.
(ii) [COMMUTATIVE LAW] We have $x + y = y + x$ for all $x, y \in \mathbb{N}$.

PROOF: (i): Fix $m, n \in \mathbb{N}$. Let $A_{m,n}$ be the set of all $r \in \mathbb{N}$ such that $(m + n) + r = m + (n + r)$. By (a) and (b),

$$(m + n) + 1 = S(m + n) = m + S(n) = m + (n + 1)$$

so $1 \in A_{m,n}$. Now suppose $r \in A_{m,n}$. We have, by (b),

$$(m + n) + S(r) = S((m + n) + r) = S(m + (n + r)) = m + S(n + r) = m + (n + S(r))$$

and thus $S(r) \in A_{m,n}$. So $A_{m,n} = \mathbb{N}$, i.e. $(m + n) + r = m + (n + r)$ for all r. Since m and n are arbitrary, (i) is proved.
(ii): Let C be the set of $n \in \mathbb{N}$ such that $n + 1 = 1 + n$. Trivially $1 \in C$. If $n \in C$, then by (a) and (b)

$$1 + S(n) = S(1 + n) = S(n + 1) = S(n) + 1$$

so $S(n) \in C$. Thus $C = \mathbb{N}$. Now fix $m \in \mathbb{N}$, and let D_m be the set of $n \in \mathbb{N}$ such that $m + n = n + m$. We have $1 \in D_m$ by the first part of the proof. Now suppose $n \in D_m$. Then

$$m + S(n) = S(m + n) = (m + n) + 1 = 1 + (m + n) = 1 + (n + m) = (1 + n) + m = (n + 1) + m = S(n) + m$$

by (i) and the first part of the proof, so $S(n) \in D_m$. Thus $D_m = \mathbb{N}$, i.e. $m + n = n + m$ for all n. Since m is arbitrary, this is true for all $m, n \in \mathbb{N}$. 👍

We now show that multiplication is associative, commutative, and distributive over addition.

III.2.4.4. THEOREM.

(i) [ASSOCIATIVE LAW] We have $(mn)r = m(nr)$ for all $m, n, r \in \mathbb{N}$.

(ii) [COMMUTATIVE LAW] We have $mn = nm$ for all $m, n \in \mathbb{N}$.

(iii) [DISTRIBUTIVE LAW] We have $m(n + r) = mn + mr$ (and $(m + n)r = mr + nr$) for all $m, n, r \in \mathbb{N}$.

PROOF: The proof is similar in form to the previous ones. We first prove the first (left) distributive law (iii). First note that if $m, n \in \mathbb{N}$, we have, by (c), (d), and commutativity of addition,

$$m(n + 1) = mS(n) = m + mn = mn + m = mn + m1 \ .$$

Fix $m, n \in \mathbb{N}$, and let $A_{m,n}$ be the set of all $r \in \mathbb{N}$ such that $m(n + r) = mn + mr$. We have $1 \in A_{m,n}$ by the first part of the proof. Now let $r \in A_{m,n}$. Then

$$m(n + S(r)) = mS(n + r) = m + m(n + r) = m + (mn + mr)$$

$$= (mn + mr) + m = mn + (mr + m1) = mn + m(r + 1) = mn + mS(n)$$

by (c), (d), associativity and commutativity of addition, and the first part of the proof applied to m and r. Thus $A_{m,n} = \mathbb{N}$ and the first distributive law is proved.

The right distributive law $(m+n)r = mr+nr$ for all $m, n, r \in \mathbb{N}$ will follow from the left distributive law and commutativity of multiplication. But we will need a special case of the right distributive law to prove commutativity of multiplication: $(1 + n)r = 1r + nr$ for all $n, r \in \mathbb{N}$. To prove this, fix $n \in \mathbb{N}$ and let B_n be the set of all $r \in \mathbb{N}$ such that the formula holds. Since

$$(1 + n)1 = 1 + n = 11 + n1$$

by (c), $1 \in B_n$. If $r \in B_n$, then

$$(1 + n)S(r) = (1 + n) + (1 + n)r = (1 + n) + (1r + nr) = (1r + 1) + (n + nr)$$

$$= (1r + 11) + (n + nr) = 1(r + 1) + nS(r) = 1S(r) + nS(r)$$

by (d), the associative and commutative laws of addition, and the left distributive law. Thus $S(r) \in B_n$, $B_n = \mathbb{N}$, and $(1 + n)r = 1r + nr$ for all $n, r \in \mathbb{N}$.

(i): To show that $\cdot$ is associative, fix $m, n \in \mathbb{N}$ and let $C_{m,n}$ be the set of all $r \in \mathbb{N}$ such that $(mn)r = m(nr)$. Since $(mn)1 = mn = m(n1)$ by (c), $1 \in B_{m,n}$. If $r \in A_{m,n}$, then

$$(mn)S(r) = mn + (mn)r = mn + m(nr) = m(n + nr) = m(nS(r))$$

by (d) and the (left) distributive law. So $S(r) \in C_{m,n}$. Thus $C_{m,n} = \mathbb{N}$ and the associative law is proved.

(ii): For the commutative law, let D be the set of n for which $1n = n$. We have $1 \in D$ by (c). If $n \in D$, then

$$1S(n) = 1 + 1n = 1 + n = n + 1 = S(n)$$

by (d), (a), and commutativity of addition. So $S(n) \in D$ and $D = \mathbb{N}$, i.e. $1n = n$ for all $n \in \mathbb{N}$.

Now fix $m \in \mathbb{N}$, and let E_m be the set of all $n \in \mathbb{N}$ such that $mn = nm$. By (c) and the previous argument, $1 \in E_m$. If $n \in E_m$, then

$$mS(n) = m + mn = m + nm = 1m + nm = (1 + n)m = (n + 1)m = S(n)m$$

by (d), the previous calculation, and the special case of the right distributive law. Thus $E_m = \mathbb{N}$ and the commutative law is proved. 👍

Among the many algebraic properties of $\mathbb{N}$ which can be proved from the Peano axioms, here are two which are particularly important:

III.2.4.5. PROPOSITION. There do not exist $m, n \in \mathbb{N}$ with $m + n = m$.

PROOF: Let A be the set of all $m \in \mathbb{N}$ such that there is no $n \in \mathbb{N}$ with $m + n = n$. We have $1 \in A$, since $1 + n = 1$ gives $1 = S(n)$, contradicting axiom (P3). Let $m \in A$, and suppose there is $n \in \mathbb{N}$ with $S(m) + n = S(m)$. Then

$$S(m + n) = S(n + m) = n + S(m) = S(m) + n = S(m)$$

by III.2.4.1.(b) and commutativity of addition. Thus $m + n = m$ by Axiom (P4), contradicting that $m \in A$. So $S(m) \in A$, $A = \mathbb{N}$. 👍

III.2.4.6. PROPOSITION. [CANCELLATION] Let $m, n, r \in \mathbb{N}$. If $m + r = n + r$, then $m = n$.

PROOF: The proof follows the usual scheme. Let A be the set of all $r \in \mathbb{N}$ such that $m + r = n + r$ implies $m = n$ for all $m, n \in \mathbb{N}$. We have $1 \in A$ by axiom (P4). If $r \in A$, let $m, n \in \mathbb{N}$ with $m + S(r) = n + S(r)$. Then

$$S(m + r) = m + S(r) = n + S(r) = S(n + r)$$

by III.2.4.1.(b). Thus $m + r = n + r$ by axiom (P4), hence $m = n$ since $r \in A$; so $S(r) \in A$ and $A = \mathbb{N}$. 👍

There is also cancellation for multiplication, but it takes more work to prove (III.2.6.7.).

III.2.5. Algebraic Axiomatization of $\mathbb{N}$

III.2.5.1. An alternate axiomatization of $\mathbb{N}$ can be given which is much closer to the way $\mathbb{N}$ is thought of and used. We have shown that $\mathbb{N}$ satisfies the following set of axioms:

- ($\mathbb{N}$1) $\mathbb{N}$ contains an element 1.
- ($\mathbb{N}$2) $\mathbb{N}$ has an associative binary operation $+$.
- ($\mathbb{N}$3) There is no $n \in \mathbb{N}$ with $n + 1 = 1$.
- ($\mathbb{N}$4) For any $m, n \in \mathbb{N}$, $m + 1 = n + 1$ implies $m = n$.
- ($\mathbb{N}$5) If $A \subset \mathbb{N}$, $1 \in A$, and A is closed under adding 1 (i.e. if $n \in A$ implies $n + 1 \in A$), then $A = \mathbb{N}$.

III.2.5.2. THEOREM. Let $(N, e, +)$ be a model for axioms (ℕ1)–(ℕ5), i.e. $e \in N$, $+$ is a binary operation on N, there is no $x \in N$ with $x + e = e$, $x + e = y + e$ implies $x = y$, and a subset A of N containing e and closed under adding e is all of N. For $x \in N$, define $s(x) = x + e$. Then (N, e, s) is a model for Peano arithmetic.

PROOF: This is quite obvious, since axioms (ℕ3), (ℕ4), and (ℕ5) are simple rephrasings of axioms (P3), (P4), and (P5) respectively. 👍

If we start with a model (N, e, S) for Peano arithmetic, define $+$ on N as in III.2.4.1., and then define s as in III.2.5.2. from $+$, it is obvious that s coincides with S. One can go the other way too:

III.2.5.3. PROPOSITION. Let $(N, e, +)$ be a model for axioms (ℕ1)–(ℕ5). Define s as in III.2.5.2., and define $+'$ on N from s as in III.2.4.1.. Then $+'$ coincides with $+$.

PROOF: We need only check that $e + e = s(e)$, $s(x) + e = s(x + e)$ for all $x \in N$, and $x + s(y) = s(x + y)$ for all $x, y \in N$, since these formulas with $+$ replaced by $+'$ characterize $+'$ by III.2.4.1.. We have $e + e = s(e)$ by definition of s. For any $x \in N$, we have

$$s(x) + e = (x + e) + e = s(x + e)$$

also by definition of s. Finally, if $x, y \in N$, we have

$$x + s(y) = x + (y + e) = (x + y) + e = s(x + y) \ .$$

👍

Putting together III.2.5.3., III.2.5.2., and III.2.2.10., we obtain:

III.2.5.4. COROLLARY. If $(N, e, +)$ and $(N', e', +')$ are models for (ℕ1)–(ℕ5), then there is a bijection $\phi : N \to N'$ such that $\phi(e) = e'$ and $\phi(x + y) = \phi(x) +' \phi(y)$ for all $x, y \in N$.

PROOF: Let ϕ be as in III.2.2.10. for the successor operations s and s'. Let A be the set of all $y \in N$ such that $\phi(x + y) = \phi(x) +' \phi(y)$ for all $x \in N$. We have that $e \in A$ by III.2.2.10.. If $y \in A$, then

$$\phi(x + s(y)) = \phi(s(x + y)) = s'(\phi(x + y)) = s'(\phi(x) +' \phi(y))$$

$$= \phi(x) +' s'(\phi(y)) = \phi(x) +' \phi(s(y))$$

and so $s(y) \in A$, $A = N$. 👍

III.2.5.5. COROLLARY. If $(N, e, +)$ is a model for axioms (ℕ1)–(ℕ5), then the binary operation $+$ is commutative.

III.2.5.6. Remarks. (i) Note that multiplication does not explicitly enter into this algebraic characterization of $\mathbb{N}$. This is because multiplication is completely determined by addition and property (d) of III.2.4.1.. In fact, if $*$ is an associative binary operation on $\mathbb{N}$ which distributes over addition (i.e. $x*(y+z)=(x*y)+(x*z)$ and $(x+y)*z=(x*z)+(y*z)$) and for which $1*1=1$, then $*$ must agree with multiplication since it must satisfy III.2.4.1.(a)–(c).

(ii) An examination of the proof of III.2.5.2. shows that associativity of $+$ is not used. Thus, if we modify ($\mathbb{N}2$) to ($\mathbb{N}2'$) by deleting the word "associative," and take ($\mathbb{N}1'$), ($\mathbb{N}3'$), ($\mathbb{N}4'$), ($\mathbb{N}5'$) to be identical to ($\mathbb{N}1$), ($\mathbb{N}3$), ($\mathbb{N}4$), ($\mathbb{N}5$) respectively, and if $(N,e,+)$ is a model for ($\mathbb{N}1'$)–($\mathbb{N}5'$) with $+$ nonassociative, we can still define s as in III.2.5.2. and obtain a model of Peano arithmetic. But the conclusion of III.2.5.3. will fail for this model, and hence the corollaries will also fail. There are such models with nonassociative $+$: see Exercise ().

III.2.5.7. All of the theorems of elementary number theory (e.g. unique factorization of natural numbers into products of prime numbers) can be proved directly from the Peano axioms. See books on number theory, e.g. ().

III.2.6. Order Structure of $\mathbb{N}$

There is also a natural order structure on $\mathbb{N}$. This order structure can either be transferred from the set containment order on ω (), or defined directly from the addition. We take the latter approach since it is more elementary.

III.2.6.1. Definition. Let $m,n\in\mathbb{N}$. Set $m<n$ if there is an $r\in\mathbb{N}$ with $m+r=n$.

III.2.6.2. Proposition. Let $m,n,r\in\mathbb{N}$. If $m<n$ and $n<r$, then $m<r$ (the relation $<$ is *transitive*).

Proof: We have $m+s=n$ and $n+t=r$ for some $s,t\in\mathbb{N}$. Then $m+(s+t)=r$. 👍

The next simple fact is essentially a rephrasing of III.2.2.3.:

III.2.6.3. Proposition. For any $n\in\mathbb{N}$, $n\neq 1$, we have $1<n$.

Proof: If $n\in\mathbb{N}$, $n\neq 1$, then by III.2.2.3. $n=S(m)$ for some $m\in\mathbb{N}$, so $1+m=n$ and $1<n$. 👍

The next result is the basic fact about $<$:

III.2.6.4. THEOREM. If $m, n \in \mathbb{N}$, then exactly one of the following is true: $m < n$, $m = n$, $n < m$.

PROOF: If $m = n$, we cannot have $m < n$ by III.2.4.5.. Similarly, if $m < n$ and $n < m$, we have $m + r = n$ and $n + s = m$ for some $r, s \in \mathbb{N}$; then $m + (r + s) = m$, contradicting III.2.4.5.. Thus, for any m and n, at most one of $m < n$, $m = n$, $n < m$ is true.

To show that at least one is true, let A be the set of $m \in \mathbb{N}$ such that, for every $n \in \mathbb{N}$, either $m < n$, $m = n$, or $n < m$ is true. We have $1 \in A$ by III.2.6.3.. If $m \in A$, and $n \in \mathbb{N}$, then either $n < m$, $n = m$, or $m < n$. If $n = m$, then $S(m) = n + 1$ so $n < S(m)$. Similarly, if $n < m$, i.e. $n + r = m$ for some $r \in \mathbb{N}$, then $S(m) = n + (r + 1)$, so $n < S(m)$. If $m < n$, then $n = m + r$ for some $r \in \mathbb{N}$. If $r = 1$, then $n = S(m)$; otherwise $r = 1 + s$ for some $s \in \mathbb{N}$, in which case $n = (m+1)+s = S(m)+s$, so $S(m) < n$. Thus $S(m) \in A$, $A = \mathbb{N}$. 👍

III.2.6.5. Combining III.2.6.4. and III.2.6.2., we obtain that the relation $<$ is a *strict total order* () on $\mathbb{N}$. As is customary, we will write $m \leq n$ if $m < n$ or $m = n$. We will also write $n > m$ if $m < n$ and $n \geq m$ if $m \leq n$.

Another important property of $<$ is *translation-invariance*:

III.2.6.6. THEOREM. Let $m, n \in \mathbb{N}$.
(i) If $m < n$, then $m + r < n + r$ for all $r \in \mathbb{N}$. If $m + r < n + r$ for some $r \in \mathbb{N}$, then $m < n$.
(ii) If $m < n$, then $mr < nr$ for all $r \in \mathbb{N}$. If $mr < nr$ for some $r \in \mathbb{N}$, then $m < n$.

PROOF: (i): If $m < n$, i.e. there is an $s \in \mathbb{N}$ with $m + s = n$, then for any $r \in \mathbb{N}$ we have $(m + r) + s = (m + s) + r = n + r$, i.e. $m + r < n + r$. Conversely, suppose $m + r < n + r$ for some r, i.e. $(m + r) + s = n + r$ for some $s \in \mathbb{N}$. Then $m + s = n$ by cancellation (III.2.4.6.), i.e. $m < n$.

(ii): If $m < n$, write $n = m + s$ for some $s \in \mathbb{N}$. Then for any $r \in \mathbb{N}$ we have $mr + sr = (m + s)r = nr$, i.e. $mr < nr$. For the converse, we cannot use multiplicative cancellation which is not yet proved, so we need a different argument (this argument could also have been used for addition). Suppose $mr < nr$ for some r. We have either $m < n$, $m = n$, or $m > n$. If $m = n$, we have $mr = nr$, contradicting III.2.6.4., and if $n < m$ we would have $nr < mr$ by the first part of the proof, again contradicting III.2.6.4.. Thus we must have $m < n$. 👍

As a corollary, we get cancellation for multiplication:

III.2.6.7. COROLLARY. Let $m, n, r \in \mathbb{N}$. If $mr = nr$, then $m = n$.

PROOF: We have either $m < n$, $m = n$, or $n < m$. If $m < n$, we would have $mr < nr$ by III.2.6.6., contradicting III.2.6.4.. Similarly, we cannot have $n < m$. 👍

III.2.6.8. Because of cancellation, we can replace $<$ by $\leq$ throughout III.2.6.6..

We also obtain:

III.2.6.9. COROLLARY. If $m, n \in \mathbb{N}$ and $mn = 1$, then $m = n = 1$.

PROOF: Suppose $n \neq 1$. Then $n > 1$, so $mn \geq m1 = m \geq 1$, a contradiction. So $n = 1$, and hence $m = 1$ too. 👍

There is one more important basic fact about the ordering on $\mathbb{N}$. If $m, n \in \mathbb{N}$, we say m is a *predecessor* of n if $m < n$. By III.1.4.11., every $n \in \mathbb{N}$ has only finitely many predecessors in $\mathbb{N}$.

III.2.7. Subtraction and Division in $\mathbb{N}$

III.2.7.1. Subtraction and division can be partially defined on $\mathbb{N}$. If $m, n \in \mathbb{N}$, we say $m - n$ is defined in $\mathbb{N}$ if there is an $r \in \mathbb{N}$ with $m = n + r$; if such a r exists, it is unique by cancellation, and we set $m - n = r$. We have that $m - n$ is defined in $\mathbb{N}$ if and only if $n < m$, and we have $m - n = (m + r) - (n + r)$ for any $r \in \mathbb{N}$. If $m, n, r, s \in \mathbb{N}$ and $m - n$ and $r - s$ are defined, then $(m + r) - (n + s)$ is defined, and $(m + r) - (n + s) = (m - n) + (r - s)$.

Similarly, if $m, n \in \mathbb{N}$, we say m is *divisible* by n in $\mathbb{N}$ if there is a (necessarily unique) $r \in \mathbb{N}$ with $m = nr$. We then say r is the *quotient* of m divided by n, written $r = \frac{m}{n}$. We have, for all $n \in \mathbb{N}$, that n is divisible by 1 and $\frac{n}{1} = n$, and n is divisible by n and $\frac{n}{n} = 1$. If m is divisible by n, we necessarily have $n \leq m$. If m is divisible by n and r is divisible by s, then mr is divisible by ns and $\frac{mr}{ns} = \frac{m}{n}\frac{r}{s}$. In particular, if m is divisible by n, then mr is divisible by nr for any $r \in \mathbb{N}$, and $\frac{mr}{nr} = \frac{m}{n}$.

III.2.7.2. The fact that subtraction is only partially defined on $\mathbb{N}$ leads to the expansion of $\mathbb{N}$ to a larger number system $\mathbb{Z}$ in which subtraction is everywhere defined. Similarly, the fact that division is only partially defined leads to the expansion of $\mathbb{N}$ (or $\mathbb{Z}$) to the system of rational numbers $\mathbb{Q}$. There is a choice of which order to do the extensions: in the following sections we will first expand to $\mathbb{Z}$ to get subtraction defined everywhere, and then expand $\mathbb{Z}$ to $\mathbb{Q}$ to get division (except division by 0) everywhere defined. We could instead proceed by expanding $\mathbb{N}$ to the positive rational numbers $\mathbb{Q}_+$ to get division defined everywhere, and then construct $\mathbb{Q}$ from $\mathbb{Q}_+$ to make subtraction everywhere defined.

III.2.8. Axiomatization of $\mathbb{N}_0$

III.2.8.1. We can also axiomatize $\mathbb{N}_0$. Actually, the Peano axioms (P1)–(P5) give an axiomatization of $\mathbb{N}_0$ if 1 is replaced by 0 throughout. But the addition (and multiplication) must be defined differently (Exercise ()). We can give an algebraic axiomatization of $\mathbb{N}_0$:

- ($\mathbb{N}_0$1) $\mathbb{N}_0$ contains elements 0 and 1.
- ($\mathbb{N}_0$2) $\mathbb{N}_0$ has an associative binary operation $+$.
- ($\mathbb{N}_0$3) If $n \in \mathbb{N}_0$, then $n + 1 \neq 0$.
- ($\mathbb{N}_0$4) $m, n \in \mathbb{N}_0$, $m + 1 = n + 1$ implies $m = n$.
- ($\mathbb{N}_0$5) If $A \subset \mathbb{N}_0$, $0 \in A$, and A is closed under adding 1 (i.e. if $n \in A$ implies $n + 1 \in A$), then $A = \mathbb{N}_0$.
- ($\mathbb{N}_0$6) $0 \neq 1$.

III.2.8.2. There is a model (in the obvious sense) for ($\mathbb{N}_0$1)–($\mathbb{N}_0$6): take $\mathbb{N}$ (or, more precisely, any model for $\mathbb{N}$), let 0 be any object not in $\mathbb{N}$, let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and extend + from $\mathbb{N}$ to $\mathbb{N}_0$ by setting $0+x = x+0 = 0$ for all $x \in \mathbb{N}_0$. The axioms are easily checked.

Conversely, let $(N_0, 0, 1, +)$ be a model for ($\mathbb{N}_0$1)–($\mathbb{N}_0$5) (($\mathbb{N}_0$6) is not needed here). If we define $S(x) = x + 1$ for $x \in N_0$, then $(N_0, 0, S)$ is a model for Peano arithmetic.

III.2.8.3. Lemma. Let $(N_0, 0, 1, +)$ be a model for ($\mathbb{N}_0$1)–($\mathbb{N}_0$6). Then $0 + 1 = 1$.

Proof: This proof is somewhat more complex than the others in this section. Define a sequence in N_0 (function from $\mathbb{N}$ to N_0) by setting $x_1 = 0 + 1$ and $x_{S(n)} = x_n + 1$ (we have avoided writing $n + 1$ for $S(n)$ since in this proof + will denote the operation in N_0). We have that $x_n \neq 0$ for all n by ($\mathbb{N}_0$3). The set $\{0\} \cup \{x_n : n \in \mathbb{N}\}$ contains 0 and is closed under adding 1, hence equals N_0 by ($\mathbb{N}_0$5). Thus $1 = x_n$ for some $n \in \mathbb{N}$ (since $1 \neq 0$ by ($\mathbb{N}_0$6)). Suppose $n = S(m)$ for some $m \in \mathbb{N}$. Then we have

$$(0 + x_m) + 1 = 0 + (x_m + 1) = 0 + x_n = 0 + 1$$

and hence $0 + x_m = 0$ by ($\mathbb{N}_0$4). If $m = S(r)$ for some $r \in \mathbb{N}$, then

$$0 = 0 + x_m = 0 + (x_r + 1) = (0 + x_r) + 1$$

which contradicts ($\mathbb{N}_0$3); thus $m = 1$, i.e. $0 + x_1 = 0$. But then

$$(0 + 0) + 1 = 0 + (0 + 1) = 0 + x_1 = 0$$

which again contradicts ($\mathbb{N}_0$3). So the assumption $n = S(m)$ for some $m \in \mathbb{N}$ is false, i.e. $n = 1$, $1 = x_1 = 0+1$. 👍

III.2.8.4. We then get $0 + 1 = 0 + (0 + 1) = (0 + 0) + 1$, so $0 = 0 + 0$ by ($\mathbb{N}_0$4). Then $(1 + 0) + 1 = 1 + (0 + 1) = 1 + 1$, so $1 + 0 = 1$ by ($\mathbb{N}_0$4). If A is the set of $x \in N_0$ for which $x + 0 = x$, we have $0 \in A$, and if $x \in A$, then $(x + 1) + 0 = x + (1 + 0) = x + 1$, so $x + 1 \in A$; thus $A = N_0$. Similarly, if B is the set of $x \in N_0$ such that $0 + x = x$, we have $0 \in B$, and if $x \in B$ then $0 + (x + 1) = (0 + x) + 1 = x + 1$, so $x + 1 \in B$; thus $B = N_0$. So $0 + x = x + 0 = x$ for all $x \in N_0$.

It follows from ($\mathbb{N}_0$3) by induction (in the usual way) that $x + y \neq 0$ for any $x, y \in N_0$, $y \neq 0$. Thus $(N_0 \setminus \{0\}, 1, +)$ (we need ($\mathbb{N}_0$6) to have $1 \in N_0 \setminus \{0\}$) is a model for ($\mathbb{N}$1)–($\mathbb{N}$5), so this set may be identified with $\mathbb{N}$. From III.2.2.10. we then obtain:

III.2.8.5. Theorem. Let $(N, 0, 1, +)$ and $(N', 0', 1', +')$ be models for ($\mathbb{N}_0$1)–($\mathbb{N}_0$6). Then there is a bijection $\phi : N' \to N$ with $\phi(0') = 0$, $\phi(1') = 1$, and $\phi(x +' y) = \phi(x) + \phi(y)$ for all $x, y \in N'$.

III.2.8.6. So we may (again with slight abuse of language) talk about *the* structure $\mathbb{N}_0$, and regard $\mathbb{N}$ as a subset of $\mathbb{N}_0$. It follows that + is commutative on $\mathbb{N}_0$. We may also define multiplication on $\mathbb{N}_0$ by extending the multiplication on $\mathbb{N}$ by setting $n0 = 0n = 0$ for all $n \in \mathbb{N}_0$. This definition is forced by the distributive law and cancellation: if $n \in \mathbb{N}$, then $0 + n = n = 1n = (0 + 1)n = 0n + 1n = 0n + n$ for any n, and the argument for $n0$ is similar. It is easily checked that the extended multiplication is associative, commutative, and distributive over addition (), and $mn = 0$ if and only if $m = 0$ or $n = 0$. We also have:

III.2.8.7. PROPOSITION. Let $m, n \in \mathbb{N}_0$. Then there are unique $r, s \in \mathbb{N}_0$ such that $m + r = n + s$ and at least one of r, s is 0.

III.2.8.8. We can extend the ordering on $\mathbb{N}$ to $\mathbb{N}_0$ by setting $0 < n$ for all $n \in \mathbb{N}$. It is easily checked that III.2.6.4. and III.2.6.6.(i) continue to hold, and that III.2.6.6.(ii) holds with $r \neq 0$.

III.2.8.9. It should be noted that we could have replaced axiom ($\mathbb{N}_0 6$) by

($\mathbb{N}_0 6'$) $0 + 1 = 1$.

Indeed, this axiom along with ($\mathbb{N}_0 3$) implies that $0 \neq 1$. The proof of Theorem III.2.8.5. is streamlined using this axiom since Lemma III.2.8.3. is not needed. Note that either ($\mathbb{N}_0 6$) or ($\mathbb{N}_0 6'$) (or some substitute) is necessary: $(\mathbb{N}, 1, 1, +)$ is a model for ($\mathbb{N}_0 1$)–($\mathbb{N}_0 5$).

III.2.9. Further Properties of $\mathbb{N}$

Many more detailed properties of $\mathbb{N}$ can be proved. Much of the detailed algebraic theory of $\mathbb{N}$ is beyond the scope of an analysis text, but we will cover a few of the essentials here.

Complete Induction, Infinite Regress, and Recursion

III.2.9.1. First, now that we have the necessary algebraic and order structure on $\mathbb{N}$, we can formulate and justify the principles of Complete Induction, Infinite Regress, and Definition by Recursion as described in (). We will not repeat them here.

Divisibility and the Division Algorithm

If $m, n, r \in \mathbb{N}$ and r divides both m and n, then an application of the distributive law shows that r divides $m + n$. There is a partial converse:

III.2.9.2. PROPOSITION. Let $m, n, r \in \mathbb{N}$. If r divides both m and $m + n$, then r also divides n.

PROOF: Write $m = ar$ and $m + n = br$ for $a, b \in \mathbb{N}$. We have $ar = m < m + n = br$, so $a < b$ by III.2.6.6.(ii). Thus $b - a = c$ is defined in $\mathbb{N}$. We have $a + c = b$, so

$$m + cr = ar + cr = br = m + n$$

so $cr = n$ by cancellation and r divides n. 👍

III.2.9.3. Now let $n, d \in \mathbb{N}$. We want to "divide" n by d and get a "quotient" and a "remainder." Since the numbers $\{md : m \in \mathbb{N}\}$ are all distinct, only finitely many are less than or equal to n, hence there is a largest such m; call it q. (There may not be any such m, i.e. we could have $d > n$; in this case set $q = 0$.) We then have $q \in \mathbb{N}_0$ and $qd \leq n < (q+1)d$. Then q is clearly the unique $m \in \mathbb{N}_0$ such that $md \leq n < (m+1)d$. We call q the *quotient* of n by d; if d divides n in the sense of III.2.7.1., this definition of quotient agrees with the definition there.

We now define the remainder. If $qd = n$, i.e. d divides n, set $r = 0$. If $qd < n$, then $r = n - qd$ is defined in $\mathbb{N}$. In either case we have $n = qd + r$. We have $0 \leq r$ since $r \in \mathbb{N}_0$, and we have

$$qd + r = n < (q+1)d = qd + d$$

so we obtain $r < d$ by III.2.6.6.(i). Then r is called the *remainder* of n divided by d. (Also, d is often called the *divisor* and n the *dividend*.)

We summarize the construction:

III.2.9.4. Theorem. [Division Algorithm] Let $n, d \in \mathbb{N}$. Then there are unique $q, r \in \mathbb{N}_0$ such that $0 \leq r < d$ and $n = qd + r$.

Proof: We only need to observe uniqueness. If we also have $n = q'd + r'$ with $0 \leq r' < d$, then

$$q'd \leq n = q'd + r' < q'd + d = (q'+1)d$$

and so $q' = q$ by the uniqueness of q with this property, and hence $r' = r$ by cancellation. 👍

Prime Numbers and Factorization

III.2.9.5. A number $p \in \mathbb{N}$ is normally defined to be *prime* if $p > 1$ and p cannot be written as a product mn of natural numbers except by taking m or n to be 1.

Although this is the standard definition of a prime number, and we will use it, it should be noted that this definition is not the same as the usual definition in abstract algebra of a prime element of a ring (the abstract algebra definition of a prime element is one satisfying the conclusion of III.2.9.6.; an element which cannot be factored in a nontrivial way is called an *irreducible* element. In $\mathbb{N}$, or $\mathbb{Z}$ (or, more generally, a Unique Factorization Domain), irreducible elements and prime elements turn out to be the same thing, but this is false in general rings.)

By III.1.5.7., every natural number can be written as a product of prime numbers. We will eventually show (III.2.9.8.) that this product is unique.

III.2.9.6. Theorem. Let p be a prime number, and $m, n \in \mathbb{N}$. If p divides mn, then p divides m or p divides n.

Proof: This can be proved somewhat more easily after $\mathbb{Z}$ is constructed (cf. III.1.6.12.), but we give a direct proof from the Peano axioms and the machinery we have already constructed from them. The proof is a multiple infinite regress.

Suppose there is a prime p for which there are $m, n \in \mathbb{N}$ not divisible by p for which p divides mn. Let p be the smallest such prime. For this p, let m be the smallest natural number not divisible by p for which

there is an $n \in \mathbb{N}$, not divisible by p, such that mn is divisible by p, and for this m let n be the smallest natural number satisfying these conditions.

Write $m = ap + r$, $n = bp + s$ with $a, b, r, s \in \mathbb{N}_0$ and $r, s < p$ by the Division Algorithm. We have $r \neq 0$ and $s \neq 0$ since p does not divide m or n; hence $r, s \in \mathbb{N}$. We have

$$mn = (ap + r)(bp + s) = abp^2 + aps + bpr + rs$$

and since mn and $abp^2 + aps + bpr$ are divisible by p, we have that rs is divisible by p by (). But r and s are not divisible by p since they are less than p. Since $r \leq m$ and $s \leq n$, we have $r = m$ and thus $s = n$ by minimality of m and n. Thus $m < p$ and $n < p$.

Write $mn = kp$ for some $k \in \mathbb{N}$. We have $k < p$ since $mn < p^2$. If $k = 1$, we have $p = mn$, contradicting that p is prime. If $k > 1$, let q be a prime number dividing k. Then $q < p$, and q divides mn, so q divides m or q divides n by minimality of p. Suppose q divides m, and write $m = qd$ and $k = qt$. Then $qdn = qtp$, so $dn = tp$ and p divides dn. But p does not divide d since it does not divide m, and $d < m$, contradicting minimality of m. If q divides n, a similar contradiction is obtained. 👍

A simple induction proof then shows:

III.2.9.7. Corollary. Let p be a prime number, and $n_1, \ldots, n_r \in \mathbb{N}$. If p divides $n_1 n_2 \cdots n_r$, then p divides at least one of the n_k.

We then obtain one of the most important facts about the natural numbers:

III.2.9.8. Theorem. [Fundamental Theorem of Arithmetic] Every natural number can be written as a (finite) product of prime numbers, uniquely up to the order of the factors. (The "empty product" 1 and products of length 1 are included. Repeated factors are, of course, allowed.)

Proof: It has already been shown (III.1.5.7.) that every natural number can be written as a product of prime numbers, so we only need to show uniqueness. This can be done by induction on the length of the product, but we phrase the argument alternately by infinite regress.

Suppose n is the smallest natural number for which there are two distinct factorizations

$$n = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s \ .$$

Then $n > 1$ by III.2.6.9., so both r and s are at least 1, and neither can be 1 since the p_j and q_k are prime. Thus $r, s > 1$. We have that p_1 divides $q_1 \cdots q_s$, so it divides q_k for some k, and since q_k is prime we have $p_1 = q_k$. By reordering the q's we may assume $k = 1$. Thus

$$\frac{n}{p_1} = p_2 \cdots p_r = q_2 \cdots q_s$$

and $\frac{n}{p_1}$ has two distinct factorizations. But $\frac{n}{p_1} < n$, contradicting minimality of n. Thus there is no such n. 👍

We conclude our brief introduction to the algebraic theory of $\mathbb{N}$ by noting that there are an infinite number of primes. This observation and its slick proof are due to Euclid.

III.2.9.9. THEOREM. There are infinitely many prime numbers. If $n \in \mathbb{N}$, there is a prime number p with $p > n$.

PROOF: Suppose $p_1, \ldots, p_r$ is a finite set of prime numbers. Set $m = 1 + p_1 p_2 \cdots p_r$. Then no p_k divides m, so if p is a prime dividing m, p is not one of the p_k and no finite list of primes is complete.

The last statement follows from the first and the fact that the set of predecessors of n is finite. As an alternate proof, the number $n!+1$ is not divisible by any number $\leq n$ except 1, so any prime number dividing it is greater than n. 👍

III.2.10. Summary

III.2.10.1. The Peano axioms for the natural numbers, like all axiom systems in mathematics, cannot be proved. They are accepted as a set of basic *assumptions* mathematicians make about the natural numbers, for several reasons:

(1) All of the axioms, with the arguable exception of P5, are statements which are "obviously" true about the natural numbers we learn about and use starting as children.

(2) Assuming consistency of some appropriate form of set theory, e.g. ZF, it can be rigorously shown that there is a mathematical structure, and (up to an obvious renaming of elements) only one, which satisfies the axioms, and which furthermore has uniquely determined arithmetic operations, definable directly from the axioms, which capture all the properties and facts we would expect about the natural numbers and no obviously false ones.

(3) All the standard properties of the natural numbers can be systematically and rigorously (if somewhat tediously) proved from the axioms.

(4) From the resulting mathematical structure, a rigorous construction can be made of larger number systems, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, which again have all the properties one would expect in these number systems. The constructed system $\mathbb{R}$ also allows a rigorous development of analysis. (Analysis is complicated enough that it cannot be said that there are no "surprises"; however, there are no known contradictions or paradoxes resulting from the development from the Peano axioms, and even the surprises make sense and become unsurprising when thoroughly thought through.)

(1)–(3) have been discussed in this section, and part of (4) will be treated in succeeding sections. The part of (4) about analysis, however, will occupy the rest of the book (and more)!

III.2.11. Exercises

III.2.11.1. Show that the ordering on $\mathbb{N}$ defined in III.2.6.1. coincides with the set containment order on ω (using the standard identification of $\mathbb{N}$ with ω).

III.3. The Integers

> Two mathematicians were standing talking on the sidewalk in front of an apparently empty building. As they talked two people walked by and went into the building. A few minutes later, three people came out of the building and walked away. The mathematicians looked at each other and one said, "Well, now, if one more person goes into the building it will be empty again."

Negative numbers came to be accepted as actual existing numbers only relatively recently (zero was accepted a little earlier); even in medieval times they were generally thought to be an absurd concept. In some contexts, as above, they are indeed absurd, but in other contexts they have a natural interpretation (e.g. as a loss in finance). It was eventually realized that there were no logical contradictions arising from them and that they were crucial to obtaining a number system with well-behaved arithmetic. We now use them comfortably, for example in the common Celsius and Fahrenheit temperature scales (but not Kelvin!), and when motion on a line is considered when motion in either direction is possible. Our counting of years is another example, with years BCE counting as "negative," although with a curious twist: there is no Year 0, throwing off some duration calculations by one (counting days and months as fractional years is also problematic past the divide).



III.3.1. Introduction and Axiomatization of $\mathbb{Z}$

III.3.1.1. The next set of numbers we will consider is the *integers*, or *whole numbers* (although historically positive fractions predated the concept of 0 or negative integers). The letter $\mathbb{Z}$ (from the German *Zahl*, "number") is customarily used to denote the set of integers. Thus, as is often stated,

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\} .$$

As with $\mathbb{N}$, this is too imprecise to be even an informal definition of $\mathbb{Z}$.

III.3.1.2. There are several equivalent ways to axiomatize $\mathbb{Z}$, and unlike with $\mathbb{N}$ there is no generally accepted standard axiomatization. We will describe a few of the possibilities here. One efficient way is the following set of axioms:

(G1) $\mathbb{Z}$ has an associative binary operation $+$.

(G2) There is an element $0 \in \mathbb{Z}$ such that $x + 0 = 0 + x = x$ for all $x \in \mathbb{Z}$.

(G3) For every $x \in \mathbb{Z}$ there is an element $(-x) \in \mathbb{Z}$ with $x + (-x) = (-x) + x = 0$.

(Cyc) There is an element $1 \in \mathbb{Z}$ such that, if $A \subseteq \mathbb{Z}$, $1 \in A$, and $x \in A$ implies $x + 1 \in A$ and $(-x) \in A$, then $A = \mathbb{Z}$.

(Fr) If P is the smallest subset of $\mathbb{Z}$ containing 1, such that $x \in P$ implies $x+1 \in P$, then $0 \notin P$.

A model for (G1)–(G3) is called a *group*. A group satisfying (Cyc) is called a *cyclic group*, and a cyclic group satisfying (Fr) is a *free cyclic group*. There is a vast theory of groups, which is a large portion of the modern subject of abstract algebra.

III.3.2. Models for $\mathbb{Z}$

There is a standard model for $\mathbb{Z}$, which can be constructed from either $\mathbb{N}$ or $\mathbb{N}_0$. The construction is slightly simpler notationally starting from $\mathbb{N}_0$, so we will do it this way. The motivation for the construction is that every integer can be written as a difference of natural numbers. However, the representation is not unique, so it takes some work to handle the nonuniqueness: the idea is that $x_1 - y_1 = x_2 - y_2$ if and only if $x_1 + y_2 = x_2 + y_1$.

III.3.2.1. Definition. Define a relation on $\mathbb{N}_0 \times \mathbb{N}_0$ by setting $(x_1, y_1) \sim (x_2, y_2)$ if $x_1 + y_2 = x_2 + y_1$.

III.3.2.2. Proposition. (i) The relation $\sim$ is an equivalence relation () on $\mathbb{N}_0 \times \mathbb{N}_0$.
(ii) If $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, then $(x_1 + z_1, y_1 + w_1) \sim (x_2 + z_2, y_2 + w_2)$.

Proof: (i): If $(x_1, y_1) \sim (x_2, y_2)$, it is trivial that $(x_2, y_2) \sim (x_1, y_1)$. It is also trivial that $(x, y) \sim (x, y)$ for any x, y. If $(x_1, y_1) \sim (x_2, y_2)$ and $(x_2, y_2) \sim (x_3, y_3)$, we have

$$(x_1 + y_3) + (x_2 + y_2) = (x_1 + y_2) + (x_2 + y_3) = (x_2 + y_1) + (x_3 + y_2) = (x_3 + y_1) + (x_2 + y_2)$$

by the associative and commutative laws, and hence $x_1+y_3 = x_3+y_1$ by cancellation. Thus $(x_1, y_1) \sim (x_3, y_3)$.
(ii): By the associative and commutative laws, we have

$$(x_1 + z_1) + (y_2 + w_2) = (x_1 + y_2) + (z_1 + w_2) = (x_2 + y_1) + (z_2 + w_1) = (x_2 + z_2) + (y_1 + w_1) \ .$$



We will write $[(x, y)]$ for the equivalence class of the pair (x, y). We will think of $[(x, y)]$ as representing the integer $x - y$; the equivalence relation $\sim$ is designed so that every representative of the equivalence class represents the same integer.

III.3.2.3. Definition. Let $\mathcal{Z}$ be the set of equivalence classes of ordered pairs. For $[(x, y)], [(z, w)] \in \mathcal{Z}$, define

$$[(x, y)] + [(z, w)] = [(x + z, y + w)] \ .$$

Note that $+$ is well defined on $\mathcal{Z}$ by III.3.2.2.(ii).

III.3.2.4. THEOREM. If $\mathcal{N}_0 = \{[(x,0)] : x \in \mathbb{N}_0\}$, then $(\mathcal{N}_0, [(0,0)], [(1,0)], +)$ is a model for $\mathbb{N}_0$, and $(\mathcal{Z}, +, [(0,0)], [(1,0)])$ is a model for $\mathbb{Z}$. Addition in $\mathcal{Z}$ is commutative.

PROOF: It is obvious that once we know that addition of equivalence classes is well defined, it inherits associativity (and also commutativity) from addition in $\mathbb{N}_0$. It is also obvious that $\mathcal{N}_0$ is closed under addition, and the map $\phi : \mathbb{N}_0 \to \mathcal{N}_0$ defined by $\phi(x) = [(x,0)]$ satisfies $\phi(x+y) = \phi(x) + \phi(y)$ for all $x, y \in \mathbb{N}_0$. Also, $\phi(x) = \phi(y)$ if and only if $(x,0) \sim (y,0)$, i.e. if and only if $x = y$. Thus $\mathcal{N}_0$ is a model for $\mathbb{N}_0$, with 0-element $\phi(0) = [(0,0)]$ and 1-element $\phi(1) = [(1,0)]$.

We now check the $\mathbb{Z}$-axioms for $\mathcal{Z}$. (G1) is already checked. We have $[(0,0)] + [(x,y)] = [(0+x, 0+y)] = [(x,y)]$ for all $x, y \in \mathbb{N}_0$, and similarly $[(x,y)] + [(0,0)] = [(x,y)]$ for all x, y; thus (G2) holds. For $x, y \in \mathbb{N}_0$, we have $[(x,y)] + [(y,x)] = [(x+y, y+x)] = [(0,0)]$, and similarly $[(y,x)] + [(x,y)] = [(0,0)]$, so (G3) holds with $-[(x,y)] = [(y,x)]$.

Suppose $A \subseteq \mathcal{Z}$ contains $[(0,0)]$, is closed under adding $[(1,0)]$ and under negations. Then $A \cap \mathcal{N}_0 = \mathcal{N}_0$ since $\mathcal{N}_0$ is a model for $\mathbb{N}_0$, so $[(x,0)] \in A$ for all $x \in \mathbb{N}_0$. Then also $-[(x,0)] = [(0,x)] \in A$ for all $x \in \mathbb{N}_0$. If $x, y \in \mathbb{N}_0$, either $x = y$, or there is a $z \in \mathbb{N}_0$ with $x + z = y$, or there is a $w \in \mathbb{N}_0$ with $x = y + w$. In the first case $[(x,y)] = [(0,0)] \in A$. If $x + z = y$, then $[(x,y)] = [(0,z)] \in A$. And if $x = y + w$, then $[(x,y)] = [(w,0)] \in A$. Thus (Cyc) holds.

It is obvious that (Fr) holds, since we have shown that $[(x,0)] \neq [(0,0)]$ for any $x \in \mathbb{N}_0$, $x \neq 0$. 👍

III.3.2.5. Note that each equivalence class in $\mathcal{Z}$ contains exactly one pair (x,y), where either x or y (or both) is 0. Thus we could have taken this set of ordered pairs as our model of $\mathbb{Z}$. This model has the advantage that the elements are more concrete; they are specific ordered pairs rather than equivalence classes. But defining addition and proving associativity are somewhat messy in this approach.

III.3.2.6. We could have done the whole construction using $\mathbb{N}$ instead of $\mathbb{N}_0$. In this case, the 0 element is $[(1,1)]$, and the 1-element is $[(1+1,1)]$. The natural number n is represented by $[(n+1,1)]$. The approach of III.3.2.5. can also be done using $\mathbb{N}$; details are left to the reader.

III.3.3. Uniqueness of $\mathbb{Z}$

III.3.3.1. Here are some basic consequences of the axioms (G1)–(G3).

(i) [CANCELLATION] If $x, y, z \in \mathbb{Z}$ and $x + z = y + z$, then $x = y$: we have

$$x = x + 0 = x + (z + (-z)) = (x + z) + (-z) = (y + z) + (-z) = y + (z + (-z)) = y + 0 = y \ .$$

Similarly, if $z + x = z + y$, then $x = y$. Since $0 + 0 = 0 = 0 + (-0)$ by (G2), we have $-0 = 0$.

(ii) If $x, y \in \mathbb{Z}$ and $x + y = 0$, then $y = -x$. In particular, $-(-x) = x$.

(iii) If $x, y \in \mathbb{Z}$, then $-(x+y) = (-y) + (-x)$: streamlining manipulations with the associative law, we have

$$(x + y) + [(-y) + (-x)] = x + (y + (-y)) + (-x) = x + 0 + (-x) = x + (-x) = 0 \ .$$

III.3.3.2. One property of $\mathbb{Z}$ which does not follow just from (G1)–(G3) is commutativity of addition: there are noncommutative groups (). Commutativity of addition does follow from (Cyc), along with (G1)–(G3). We will not prove this (cf. ()) since we will deduce commutativity of addition in another way.

III.3.3.3. THEOREM. Let $(Z, +, 0, 1)$ be a model for $\mathbb{Z}$, and let $P \subseteq Z$ be the set defined in (Fr). Then

(i) P is closed under addition, i.e. if $x, y \in P$, then $x + y \in P$.

(ii) $(P, 1, +)$ is a model for $\mathbb{N}$. Thus $+$ is commutative on P.

(iii) If $-P = \{-x : x \in P\}$, then Z is the disjoint union of P, $-P$, and $\{0\}$.

PROOF: (i): Let A be the set of all $y \in P$ such that $x + y \in P$ for all $x \in P$. Since $x \in P$ implies $x + 1 \in P$, we have $1 \in A$. If $y \in A$, then for any $x \in P$, we have $x + (y + 1) = (x + y) + 1 \in P$, so $y + 1 \in A$. So A is a subset of Z containing 1 and closed under adding 1; thus $P \subseteq A$ since P is the smallest such set. Since $A \subseteq P$, we have $A = P$.

(ii): Axiom (ℕ1) holds by (i), and axiom (ℕ2) is trivial. Since $0 + 1 = 1$ and $0 \notin P$, we have $x + 1 \neq 1$ for all $x \in P$ by cancellation, so axiom (ℕ3) holds. Axiom (ℕ4) holds by cancellation. Axiom (ℕ5) also holds by an argument similar to the proof of (i). It follows from III.2.5.5. that $+$ is commutative on P.

(iii) We have $\{0\} \cap P = \emptyset$ by (Fr), and similarly $\{0\} \cap (-P) = \emptyset$ since $-0 = 0$. If $x \in P \cap (-P)$, so $x = (-y)$ for some $y \in P$, then $0 = x + y \in P$ by (i), a contradiction; thus $P \cap (-P) = \emptyset$. If $x, y \in P$, then by (ii) and III.2.6.4. either $x = y$, or there is a $z \in P$ with $x + z = y$, or there is a $w \in P$ with $x = y + w$. Now let $A = P \cup (-P) \cup \{0\}$. We have $0 \in A$, and if $x \in A$, then $(-x) \in A$. If $x \in A$, we show $x + 1 \in A$. If $x = 0$ or $x \in P$, this is obvious. If $x \in (-P)$, then $x = (-y)$ for some $y \in P$. If $y = 1$, then $x + 1 = 0 \in A$; if $y = z + 1 = 1 + z$ for some $z \in P$, then $x = (-y) = (-z) + (-1)$, so $x + 1 = (-z) \in A$. Thus $A = \mathbb{Z}$ by (Cyc). 👍

III.3.3.4. PROPOSITION. If $(Z, +, 0, 1)$ is a model for $\mathbb{Z}$, then $+$ is commutative.

PROOF: Addition is commutative on P by III.3.3.3.(ii), and we have $0 + x = x + 0$ for any $x \in Z$ by axiom (G2). It remains to show that $x + (-y) = (-y) + x$ for $x, y \in P$. Fix $x, y \in P$. We have

$$y + [x + (-y)] = (y + x) + (-y) = (x + y) + (-y) = x + [y + (-y)] = x = [y + (-y)] + x = y + [(-y) + x]$$

and so $x + (-y) = (-y) + x$ by cancellation. 👍

III.3.3.5. THEOREM. If $(Z, +, 0, 1)$ and $(Z', +', 0', 1')$ are models for $\mathbb{Z}$, then there is a bijection $\phi : Z \to Z'$ such that $\phi(0) = 0'$, $\phi(1) = 1'$, and $\phi(x + y) = \phi(x) +' \phi(y)$ for all $x, y \in Z$.

PROOF: Let P and P' be the subsets of Z and Z' respectively defined in (Fr). Since P and P' are models for Peano arithmetic, there is a bijection $\phi : P \to P'$ with $\phi(1) = 1'$ and $\phi(x + y) = \phi(x) +' \phi(y)$ for all $x, y \in P$ (III.2.5.4.). Extend ϕ by setting $\phi(0) = 0'$ and $\phi(-x) = -\phi(x)$ for $x \in P$. By III.3.3.3., ϕ is well defined and bijective. We need only check that ϕ preserves addition. Fix $x, y \in P$. We know that $\phi(x+y) = \phi(x) +' \phi(y)$. We also have

$$\phi((-x) + (-y)) = \phi(-(y + x)) = -\phi(y + x) = -(\phi(y) +' \phi(x)) = [-\phi(x)] + [-\phi(y)] \ .$$

Finally, we either have $x = y$, or $x + z = y$ for some $z \in P$, or $x = y + w$ for some $w \in P$. If $x = y$, we have

$$\phi(x + (-y)) = \phi(0) = \phi(x) +' [-\phi(y)] = \phi(x) +' \phi(-y) \ .$$

If $x = y + w$, we have $x + (-y) = w$ (using commutativity of addition). Then

$$\phi(x) +' \phi(-y) = \phi(x) +' [-\phi(y)] = \phi(y + w) +' [-\phi(y)] = \phi(y) +' \phi(w) +' [-\phi(y)] = \phi(w) = \phi(x + (-y)) \ .$$

The argument if $x + z = y$ is almost identical. 👍

III.3.3.6. So, as with $\mathbb{N}_0$ and $\mathbb{N}$, we may talk about *the* set (or structure) $\mathbb{Z}$ without defining it precisely as a set. We will identify $\mathbb{N}$ with the subset P, called the set of *positive integers*, and hence identify $\mathbb{N}_0$ with the set $\mathbb{N} \cup \{0\}$ of *nonnegative integers*.

III.3.3.7. Note that the element 0 in $\mathbb{Z}$ is uniquely determined by the algebraic structure: since $x+0 = x$ for any x, by cancellation 0 is the unique $x \in \mathbb{Z}$ with the property that $x+x = x$. However, 1 is not completely determined by the algebraic (additive) structure of $\mathbb{Z}$: if $(Z, +, 0, 1)$ is a model for $\mathbb{Z}$, then $(Z, +, 0, -1)$ is also a model for $\mathbb{Z}$. In group-theoretic language, a free cyclic group has two generators. The axioms for $\mathbb{Z}$ as we have stated them implicitly require one of the two generators to be specified as the "positive" one, and the rest of the structure then follows uniquely.

III.3.4. Ordering and Subtraction in $\mathbb{Z}$

III.3.4.1. We may also define a total ordering on $\mathbb{Z}$ extending the order on $\mathbb{N}_0$ by setting $(-n) < 0$ for $n \in \mathbb{N}$, and $(-n) < (-m)$ if $n, m \in \mathbb{N}$ and $m < n$. It is easily checked that this is a total order and that $x < y$ implies $x + z < y + z$ for $x, y, z \in \mathbb{Z}$. We will also use the symbols $>$, $\leq$, $\geq$ with their usual (obvious) meaning.

III.3.4.2. We define the *absolute value* of an $x \in \mathbb{Z}$ by $|x| = x$ if $x \geq 0$ and $|x| = -x$ if $x < 0$, i.e. $|0| = 0$ and $|x| = |-x| = x$ for $x \in P$. We obviously then have $|x| \geq 0$ and $|-x| = |x|$ for all $x \in \mathbb{Z}$.

We also define the *signum* $\operatorname{sgn}(x)$ to be 1 if $x \in P$ and -1 if $x \in -P$; $\operatorname{sgn}(0)$ is not defined.

III.3.4.3. Proposition. [Triangle Inequality] If $x, y \in \mathbb{Z}$, then $||x| - |y|| \leq |x + y| \leq |x| + |y|$.

The proof is a straightforward but tedious case-by-case check. The result is called the "triangle inequality" in analogy with the vector version ().

III.3.4.4. We may define *subtraction* in $\mathbb{Z}$ by $x - y = x + (-y)$ for $x, y \in Z$. This operation extends the partially defined subtraction in $\mathbb{N}$. Subtraction is a binary operation on $\mathbb{Z}$, but is not associative or commutative. We have $(x + z) - (y + w) = (x - y) + (z - w)$ (and in particular $x - y = (x + z) - (y + z)$) and $y - x = -(x - y)$ for $x, y, z, w \in \mathbb{Z}$. We then have $|x - y| = |y - x|$ for all $x, y \in \mathbb{Z}$.

III.3.5. Multiplication and Division in $\mathbb{Z}$

III.3.5.1. We now discuss multiplication in $\mathbb{Z}$. We define multiplication on $\mathbb{N}_0$ as in III.2.8.6.. There is only one way to extend this multiplication to $\mathbb{Z}$ so that the distributive laws hold: since

$$0 = 0 \cdot 0 = 0(x + (-x)) = 0x + 0(-x) = 0 + 0(-x) = 0(-x)$$

for $x \in \mathbb{N}_0$, we must have $0y = 0$ for all $y \in \mathbb{Z}$. Similarly, $y0 = 0$ for all $y \in \mathbb{Z}$. Then, if $x, y \in \mathbb{Z}$,

$$0 = x0 = x(y + (-y)) = xy + x(-y)$$

so we must have $x(-y) = -(xy)$ for all x, y. Similarly, $(-x)y = -(xy)$ for all $x, y \in \mathbb{Z}$. Thus the following definition of multiplication is forced by the distributive laws:

III.3.5.2. DEFINITION. If $x, y \in \mathbb{N}$, define $x(-y) = (-x)y = -(xy)$, $(-x)(-y) = xy$, and $x0 = 0x = 0$.

It is obvious that this gives a well-defined binary operation on $\mathbb{Z}$. Despite the use of the distributive law in motivating the definition of multiplication, the associative, commutative, and distributive laws must still be checked. The commutative law is obvious, and the proof of the distributive law is essentially the reversal of the argument in III.3.5.1.. The associative law is a simple but tedious calculation checking case-by-case, using the associative law in $\mathbb{N}_0$. Thus we obtain:

III.3.5.3. THEOREM. Multiplication in $\mathbb{Z}$ is associative, commutative, and distributes over addition, and is uniquely determined by these properties and the requirement that $1 \cdot 1 = 1$.

The last statement follows from III.3.5.1. and III.2.5.6.(i).

The next results are easy consequences of the definitions, and the proofs are left to the reader.

III.3.5.4. PROPOSITION. If $x, y, z \in \mathbb{Z}$, $z \neq 0$, and $xz = yz$, then $x = y$. If $xy = 0$, then $x = 0$ or $y = 0$.

III.3.5.5. PROPOSITION. If $x, y \in \mathbb{Z}$, then $|xy| = |x||y|$. If $x, y \neq 0$, then $\operatorname{sgn}(xy) = \operatorname{sgn}(x)\operatorname{sgn}(y)$.

III.3.5.6. PROPOSITION. Let $x, y, z \in \mathbb{Z}$ with $x < y$. If $z > 0$, then $xz < yz$. If $z < 0$, then $xz > yz$.

III.3.5.7. We may define division in $\mathbb{Z}$ as in $\mathbb{N}$: if $x, y \in \mathbb{Z}$, $y \neq 0$, we say x is *divisible* by y in $\mathbb{Z}$ if there is a (necessarily unique) $z \in \mathbb{Z}$ with $x = yz$. We then say z is the *quotient* of x divided by y, written $z = \frac{x}{y}$. If $x, y \in \mathbb{N}$, this notion coincides with divisibility and quotient in $\mathbb{N}$. We have, for all $x \in \mathbb{Z}$, that x is divisible by 1 and $\frac{x}{1} = x$, and if $x \neq 0$, then x is divisible by x and $\frac{x}{x} = 1$. If x is divisible by y and z is divisible by w, then xz is divisible by yw and $\frac{xz}{yw} = \frac{x}{y}\frac{z}{w}$. In particular, if x is divisible by y, then xz is divisible by yz for any $z \in \mathbb{Z}$, $z \neq 0$, and $\frac{xz}{yz} = \frac{x}{y}$. If $x \neq 0$, then no meaning can be attached to $\frac{x}{0}$ since there is no z with $z0 = x$. We also can give no usable definition of $\frac{0}{0}$.

Quite apart from the impossibility of dividing by 0, which is insurmountable, division is still only partially defined on $\mathbb{Z}$. It is thus desirable to expand $\mathbb{Z}$ to a larger number system $\mathbb{Q}$ in which division (except division by 0) is everywhere defined.

III.3.6. Ring Axiomatization of $\mathbb{Z}$

We give here an alternate axiomatization of $\mathbb{Z}$. This axiomatization is not as "efficient" as the previous one, since it postulates existence of multiplication and its properties which can be deduced from the previous axioms, and also postulates commutativity of addition. However, this axiom scheme is commonly used, and natural subsets of these axioms define structures which are important in abstract algebra.

III.3.6.1. Ring Axioms for $\mathbb{Z}$:

(Op) There are two binary operations $+$ and $\cdot$ on $\mathbb{Z}$.

(A1) The binary operation $+$ is associative.

(A2) The binary operation $+$ is commutative.

(A3) There is an element $0 \in \mathbb{Z}$ such that $x + 0 = 0 + x = x$ for all $x \in \mathbb{Z}$.

(A4) For every $x \in \mathbb{Z}$ there is an element $(-x) \in \mathbb{Z}$ with $x + (-x) = (-x) + x = 0$.

(M1) The binary operation $\cdot$ is associative.

(M2) The binary operation $\cdot$ is commutative.

(M3) There is an element $1 \in \mathbb{Z}$ such that $x \cdot 1 = 1 \cdot x = x$ for all $x \in \mathbb{Z}$.

(D) The operation $\cdot$ distributes over addition.

(Cyc) If $A \subseteq \mathbb{Z}$, $1 \in A$, and $x \in A$ implies $x + 1 \in A$ and $(-x) \in A$, then $A = \mathbb{Z}$.

(Fr) If P is the smallest subset of $\mathbb{Z}$ containing 1, such that $x \in P$ implies $x + 1 \in P$, then $0 \notin P$.

We have seen that a set Z with a binary operation $+$ satisfying (A1), (A3), (A4), and with an element 1 satisfying (Cyc) and (Fr) is a model for $\mathbb{Z}$, hence also satisfies (A2), and there is a unique multiplication on Z for which the rest of the axioms are satisfied. Thus from III.3.3.5. we obtain:

III.3.6.2. THEOREM. Let $(Z, +, \cdot, 0, 1)$ be a model for the ring axioms of $\mathbb{Z}$. then there is a bijection $\phi : Z \to \mathbb{Z}$ with $\phi(0) = 0$, $\phi(1) = 1$, $\phi(x + y) = \phi(x) + \phi(y)$, and $\phi(x \cdot y) = \phi(x)\phi(y)$, where $\mathbb{Z}$ has its previously defined addition, multiplication, 0, and 1.

III.3.6.3. DEFINITION. A set R with two binary operations $+$ and $\cdot$ satisfying (A1)–(A4) [with $\mathbb{Z}$ replaced by R in the statements], (M1), and (D) is called a *ring*. A ring is *commutative* if it also satisfies (M2). A ring is *unital* if it also satisfies (M3) (an explicit assumption is usually made that $1 \neq 0$ in a unital ring; it is unnecessary to explicitly include this in our axiomatization of $\mathbb{Z}$ since it is an automatic consequence of (Fr)).

Several examples of rings are given in the Exercises, including commutative nonunital rings, noncommutative unital rings, and noncommutative nonunital rings. Note that the term *commutative*, when applied to a ring, refers to multiplication; addition is always commutative in a ring. Ring theory is an extensive and important part of abstract algebra, with many applications.

We will not develop the general theory of rings here, but we need one basic proposition:

III.3.6.4. PROPOSITION. Let R be a ring. Then

(i) R has cancellation for addition.

(ii) If $x \in R$, then $x0 = 0x = 0$.

(iii) If R is unital and contains more than one element, then $1 \neq 0$.

PROOF: The proof of (i) is the same as for $\mathbb{Z}$ (III.3.3.1.(i)), which uses only (G1)–(G3), which are the same as (A1), (A3), (A4). (Actually cancellation holds in any group by this proof.)

For (ii), for any $x \in R$ we have $0 + 0x = 0x = (0 + 0)x = 0x + 0x$, so $0x = 0$. Similarly, $x0 = 0$.

(iii) follows immediately from (ii): if $x \in R$, $x \neq 0$, then $0x = 0 \neq x = 1x$. ☝

III.3.6.5. The axiom (Cyc) is often replaced by the following axiom:

(Pos) There is a nonempty subset P' of $\mathbb{Z}$, closed under addition and multiplication, such that $\mathbb{Z}$ is the disjoint union of P', $\{0\}$, and $-P' = \{-x : x \in P'\}$.

It is obvious that (Pos) follows from the ring axioms for $\mathbb{Z}$ (taking $P' = P$). To show that (Pos) $\Rightarrow$ (Cyc) in the presence of the other axioms, we first need:

III.3.6.6. PROPOSITION. Let $(Z, +, \cdot, 0, 1)$ be a model for the ring axioms of $\mathbb{Z}$ with (Cyc) replaced by (Pos). Then $1 \in P'$.

PROOF: Since Z is a unital ring and $P' \neq \emptyset$, we have $0 \neq 1$ in Z by III.3.6.4.(iii). Thus either $1 \in P'$ or $(-1) \in P'$. If $(-1) \in P'$, then $1 = (-1)(-1) \in P'$. (In fact, $(-1) \notin P'$ since then $1 \in P' \cap (-P')$.) ☝

Once we know that $1 \in P'$, it follows from the proof of III.3.3.3. that $P = P'$ and that (Cyc) holds. So we get that the following set of axioms characterize $\mathbb{Z}$:

III.3.6.7. Ordered Ring Axioms for $\mathbb{Z}$:

(Op) There are two binary operations $+$ and $\cdot$ on $\mathbb{Z}$.

(A1) The binary operation $+$ is associative.

(A2) The binary operation $+$ is commutative.

(A3) There is an element $0 \in \mathbb{Z}$ such that $x + 0 = 0 + x = x$ for all $x \in \mathbb{Z}$.

(A4) For every $x \in \mathbb{Z}$ there is an element $(-x) \in \mathbb{Z}$ with $x + (-x) = (-x) + x = 0$.

(M1) The binary operation $\cdot$ is associative.

(M2) The binary operation $\cdot$ is commutative.

(M3) There is an element $1 \in \mathbb{Z}$ such that $x \cdot 1 = 1 \cdot x = x$ for all $x \in \mathbb{Z}$.

(D) The operation $\cdot$ distributes over addition.

(Pos) There is a nonempty subset P of $\mathbb{Z}$ such that

- (P1) If $x, y \in P$, then $x + y \in P$.
- (P2) If $x, y \in P$, then $xy \in P$.
- (P3) $0 \notin P$.
- (P4) If $x \in \mathbb{Z}$, $x \neq 0$, then either $x \in P$ or $(-x) \in P$, but not both.

($\text{Min}_{\mathbb{Z}}$) P is the smallest subset A of $\mathbb{Z}$ containing 1, such that $x \in A$ implies $x + 1 \in A$.

III.3.6.8. The ordering on $\mathbb{Z}$ can be cleanly expressed in terms of these axioms: $x < y$ if and only if $(y - x) \in P$. The specification of the ordering via (Pos) can be done alternately by the axiom

(Ord) There is a relation (strict total order) $<$ on $\mathbb{Z}$, such that

(O1) If $x, y \in \mathbb{Z}$, then exactly one of $x < y$, $y < x$, $x = y$ holds. *(Trichotomy)*

(O2) If $x, y, z \in \mathbb{Q}$, $x < y$, and $y < z$, then $x < z$. *(Transitivity)*

(O3) If $x, y, z \in \mathbb{Z}$ and $x < y$, then $x + z < y + z$. *(Translation Invariance)*

(O4) If $x, y, z \in \mathbb{Z}$, $x < y$, and $0 < z$, then $xz < yz$. *(Homogeneity or Scaling Invariance)*

It is easily seen that in the presence of the other axioms [($\text{Min}_{\mathbb{Z}}$) is not needed], (Pos) is equivalent to (Ord); if (Ord) is assumed, P is defined to be $\{x \in \mathbb{Z} : 0 < x\}$. It is a matter of taste whether the ordering is axiomatized using (Pos) or (Ord); both are used in the literature.

A commutative ring satisfying (Pos) (or, equivalently, (Ord)) is called an *ordered ring*.

III.3.6.9. REMARK. The axioms (P3) and (P4) can be replaced by

(P3′) If $x \in P$, then $(-x) \notin P$.

(P4′) If $x \in \mathbb{Z}$, $x \neq 0$, then either $x \in P$ or $(-x) \in P$.

It is obvious that (P3) and (P4) imply (P3′) and (P4′), and that (P3′) and (P4′) imply (P4). Note also that if (P3′) is assumed, then we cannot have $0 \in P$: if $0 \in P$, then $(-0) \notin P$ by (P3′), but $-0 = 0$, a contradiction. Thus (P3′) implies (P3).

III.3.7. Exercises

III.3.7.1. Show that the set of even integers, with its usual addition and multiplication, is a nonunital commutative ring.

III.3.7.2. Fix $n \in \mathbb{N}$, and let $M_n(\mathbb{R})$ be the set of all $n \times n$ matrices with real entries. Give $M_n(\mathbb{R})$ the usual matrix addition and multiplication. Show that $M_n(\mathbb{R})$, with this addition and multiplication, is a unital ring, which is not commutative if $n > 1$. Do the same for $M_n(\mathbb{Z})$, $M_n(\mathbb{Q})$, and $M_n(\mathbb{C})$.

III.3.7.3. Let R be the set of all sequences of 2×2 matrices (with real entries) which are eventually 0. Add and multiply sequences coordinatewise. Show that R with these operations is a noncommutative nonunital ring.

III.3.7.4. Let R be a ring. A subset J of R is an *ideal* of R if J is closed under addition, $0 \in J$, $x \in J$ implies $(-x) \in J$, and $x \in J$, $y \in R$ imply that $xy, yx \in J$ (cf. II.6.4.12.).

(a) Fix $n \in \mathbb{N}$, and let $n\mathbb{Z}$ be the set of all integers divisible by n. Show that $n\mathbb{Z}$ is an ideal in $\mathbb{Z}$. Show that every ideal of $\mathbb{Z}$ other than $\{0\}$ is of this form.

(b) Let J be an ideal in a ring R. If $x, y \in R$, set $x \sim y$ if $(y - x) \in J$. Show that $\sim$ is an equivalence relation.

(c) Let R/J be the set of $\sim$-equivalence classes $\{[x] : x \in R\}$ (cf. (b)). Show that $[x] + [y] = [x + y]$ and $[x][y] = [xy]$ give well-defined binary operations on R/J making it into a ring, which is commutative [resp. unital] if R is commutative [resp. unital]. R/J is called the *quotient ring* of R by J, pronounced "R mod J."

III.3.7.5. Use the notation and terminology of Exercise III.3.7.4..

(a) Show that the quotient ring $\mathbb{Z}/n\mathbb{Z}$ can be identified with the ring $\mathbb{Z}_n$ of "integers mod n," whose elements are $\{0, 1, 2, \ldots, n-1\}$ and where addition and multiplication are done modulo n.

(b) Show that $(\mathbb{Z}_n, 1, +)$ is a model for $\mathbb{N}$ with axiom ($\mathbb{N}$3) deleted.

(c) Show that $(\mathbb{Z}_n, +, 0, 1)$ is a model for $\mathbb{Z}$ with axiom (Fr) deleted.

III.3.7.6. This problem gives an extension of Theorem III.3.6.2.. Let R be a unital ring. Define a function $\phi : \mathbb{Z} \to R$ by $\phi(0) = 0$, $\phi(1) = 1$, and $\phi(x+1) = \phi(x) + 1$ and $\phi(-x) = -\phi(x)$ for $x \in \mathbb{N}$.

(a) Show that $\phi(x+y) = \phi(x) + \phi(y)$ and $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in \mathbb{Z}$ (ϕ is a *ring-homomorphism*).

(b) Show that $\phi(\mathbb{Z}) = \{\phi(x) : x \in \mathbb{Z}\}$ is a *subring* of R (a subset closed under addition and multiplication which is itself a ring), and is the smallest subring of R containing 1.

(c) Show that $J = \{x \in \mathbb{Z} : \phi(x) = 0\}$ is an ideal of $\mathbb{Z}$, and that ϕ gives a *ring-isomorphism* (bijective ring-homomorphism) from $\mathbb{Z}/J$ to $\phi(\mathbb{Z})$. Thus, either ϕ is injective and $\phi(\mathbb{Z})$ is isomorphic to $\mathbb{Z}$ as a ring, or $J = n\mathbb{Z}$ for some $n \in \mathbb{N}$, in which case $\phi(\mathbb{Z})$ is isomorphic to $\mathbb{Z}_n$ as a ring.

III.3.7.7. The Grothendieck Group. Let $(S, *)$ be an *abelian semigroup*, a set S with an associative, commutative binary operation $*$. There is a standard construction of an abelian group $G(S)$ which is virtually identical to the construction of $\mathbb{Z}$ from $\mathbb{N}$ or $\mathbb{N}_0$. The only change is that the equivalence relation must be modified to account for possible lack of cancellation in $(S, *)$.

(a) Define a relation $\sim$ on $S \times S$ by setting $(x_1, y_1) \sim (x_2, y_2)$ if there is a $z \in S$ with

$$x_1 * y_2 * z = x_2 * y_1 * z \ .$$

Show that $\sim$ is an equivalence relation on $S \times S$.

(b) Let $G(S)$ be the set of $\sim$-equivalence classes $\{[(x, y)] : x, y \in S\}$. Show that $[(x, y)]*[(z, w)] = [(x*z, y*w)]$ is a well-defined binary operation on $G(S)$, which is associative and commutative.

(c) Show that $[(x, x)] = [(y, y)]$ for any $x, y \in S$, and that this class gives an additive identity.

(d) If $-[(x, y)] = [(y, x)]$, show that $G(S)$ is an abelian (commutative) group, i.e. satisfies axioms (A1)–(A4).

(e) Show that $[(x * y, y)] = [(x * z, z)]$ for any $x, y, z \in S$. Show that the map $\iota : S \to G(S)$ defined by $\iota(x) = [(x * x, x)]$ is a (group-)homomorphism, i.e. $\iota(x * y) = \iota(x) * \iota(y)$ for all $x, y \in S$. Show that ι is injective if and only if S has cancellation.

(f) Show that this construction, applied to $(\mathbb{N}, \cdot)$, gives the positive rational numbers with multiplication.

(g) Show that the group $G(S)$ is *universal* in the sense that if H is any abelian group and $\phi : S \to H$ is a homomorphism, there is a unique homomorphism $\psi : G(S) \to H$ such that $\phi = \psi \circ \iota$.

(h) Show that $S \mapsto G(S)$ is a functor () from the category of abelian semigroups to the category of abelian groups. Part of showing this is to show that if S and S' are abelian semigroups, and $\phi : S \to S'$ is a homomorphism, there is a corresponding induced homomorphism from $G(S)$ to $G(S')$.

The term *abelian*, as a synonym for *commutative*, is used for groups and semigroups (and also for rings by some authors; abelian categories, functions, and varieties also use this name) in honor of the Norwegian mathematician Niels Henrik Abel, who made important contributions to both algebra and analysis before his death at age 26 from tuberculosis.

III.3.7.8. An element x in a commutative ring R is a *zero divisor* if $x \neq 0$ and there exists $y \in R$, $y \neq 0$, with $xy = 0$.

(a) Show that $xz = yz$ implies $x = y$ in a commutative ring R if and only if z is not a zero divisor. [Consider $(x - y)z$.]

(b) A unital commutative ring is an *integral domain* if it contains no zero divisors. Show that every unital subring of a field is an integral domain. Show that the polynomial ring $\mathbb{R}[X]$ (Exercise ()) is an integral domain.

III.4. The Rational Numbers

III.4.1. Introduction and Axiomatization of $\mathbb{Q}$

The next step in building the number system is the construction of the rational numbers, usually denoted $\mathbb{Q}$ (for "quotient"). We will follow the usual process of listing a set of axioms satisfied by $\mathbb{Q}$, and then showing that there is a model for the axioms and that the model is essentially unique.

III.4.1.1. Rational numbers are often described as "fractions," but this is incorrect. Rational numbers are not the same thing as fractions: many fractions such as $\frac{1}{2}$ and $\frac{2}{4}$ can represent the same rational number. It is more correct to say that a rational number is a number which can be represented as a fraction. (A similar distinction can be made for integers: an integer is a number that can be represented as a difference of natural numbers, but integers are not the same thing as differences of natural numbers.) Saying that rational numbers *are* fractions leads to other logical problems. For example, is $\cos\frac{\pi}{3}$ a rational number? What about $\sum_{k=0}^{\infty} \frac{1}{3^k}$?

III.4.1.2. The axioms satisfied by $\mathbb{Q}$ are very similar to the ones satisfied by $\mathbb{Z}$, along with an additional axiom stating that every nonzero element of $\mathbb{Q}$ has a multiplicative inverse. We state them in the ordered field form in analogy with III.3.6.7.. (These are not the most efficient axioms, but are the ones most commonly used.)

Ordered Field Axioms for $\mathbb{Q}$

(Op) There are two binary operations $+$ and $\cdot$ on $\mathbb{Q}$.

(A1) The binary operation $+$ is associative.

(A2) The binary operation $+$ is commutative.

(A3) There is an element $0 \in \mathbb{Q}$ such that $x + 0 = 0 + x = x$ for all $x \in \mathbb{Q}$.

(A4) For every $x \in \mathbb{Q}$ there is an element $(-x) \in \mathbb{Q}$ with $x + (-x) = (-x) + x = 0$.

(M1) The binary operation $\cdot$ is associative.

(M2) The binary operation $\cdot$ is commutative.

(M3) There is an element $1 \in \mathbb{Q}$ such that $x \cdot 1 = 1 \cdot x = x$ for all $x \in \mathbb{Q}$.

(M4) For every $x \in \mathbb{Q}$, $x \neq 0$, there is an element $x^{-1} \in \mathbb{Q}$ with $x^{-1}x = xx^{-1} = 1$.

(D) The operation $\cdot$ distributes over addition.

(Pos) There is a nonempty subset P of $\mathbb{Q}$ such that

(P1) If $x, y \in P$, then $x + y \in P$.

(P2) If $x, y \in P$, then $xy \in P$.

(P3) $0 \notin P$.

(P4) If $x \in \mathbb{Q}$, $x \neq 0$, then either $x \in P$ or $(-x) \in P$, but not both.

($\text{Min}_{\mathbb{Q}}$) P is the smallest subset A of $\mathbb{Q}$ containing 1, such that $x \in A$ implies $x + 1 \in A$ and $x \in A$ implies $x^{-1} \in A$.

As with $\mathbb{Z}$, the axiom (Ord) may be used in place of (Pos):

(Ord) There is a nonempty relation (strict total order) $<$ on $\mathbb{Q}$, such that

(O1) If $x, y \in \mathbb{Q}$, then exactly one of $x < y$, $y < x$, $x = y$ holds. *(Trichotomy)*

(O2) If $x, y, z \in \mathbb{Q}$, $x < y$, and $y < z$, then $x < z$. *(Transitivity)*

(O3) If $x, y, z \in \mathbb{Q}$ and $x < y$, then $x + z < y + z$. *(Translation Invariance)*

(O4) If $x, y, z \in \mathbb{Q}$, $x < y$, and $0 < z$, then $xz < yz$. *(Homogeneity or Scaling Invariance)*

III.4.1.3. A set with two binary operations satisfying (A1)–(A4), (M1)–(M4), and (D), and in which $0 \neq 1$, is called a *field*. A field satisfying (Pos) is an *ordered field* (note that an ordered field contains more than one element by (Pos), so $0 \neq 1$ automatically by III.3.6.4.(iii)).

A field is just a unital commutative ring in which every nonzero element has a multiplicative inverse. There are fields with finitely many elements (Exercise III.4.3.1.). An ordered field looks more like $\mathbb{Q}$ (Exercise III.4.3.7.), but there are many other ordered fields such as $\mathbb{R}$; see also Exercises III.5.3.1. and III.5.3.3.(d). See also Exercise III.6.3.2. for an interesting fieldlike ring which is not commutative.

III.4.2. The Standard Model of $\mathbb{Q}$

We can form a standard model for $\mathbb{Q}$ in much the same manner as for $\mathbb{Z}$, although the construction will be slightly more complicated since we must explicitly extend both addition and multiplication. The idea is that rational numbers can be represented by equivalence classes of fractions $\frac{x}{y}$, where $x \in \mathbb{Z}$, $y \in \mathbb{N}$, and $\frac{x}{y} = \frac{z}{w}$ if and only if $xw = yz$.

III.4.2.1. Definition. Define a relation $\sim$ on $\mathbb{Z} \times \mathbb{N}$ by $(x_1, y_1) \sim (x_2, y_2)$ if $x_1 y_2 = x_2 y_1$.

We think of (x, y) as representing the rational number $\frac{x}{y}$. We could just as well work on $\mathbb{Z} \times (\mathbb{Z} \setminus \{0\})$.

III.4.2.2. Proposition. (i) The relation $\sim$ is an equivalence relation on $\mathbb{Z} \times \mathbb{N}$.
(ii) If $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, then $(x_1 w_1 + y_1 z_1, y_1 w_1) \sim (x_2 w_2 + y_2 z_2, y_2 w_2)$ and $(x_1 z_1, y_1 w_1) \sim (x_2 z_2, y_2 w_2)$.

Proof: This is a straightforward calculation and is left to the reader. 👍

The operations in (ii) are motivated by the informal (at this point) relations

$$\frac{x}{y} + \frac{z}{w} = \frac{xw + yz}{yw} \quad \text{and} \quad \frac{x}{y}\frac{z}{w} = \frac{xz}{yw} .$$

We use these relations to define addition and multiplication:

III.4.2.3. DEFINITION. Let $\mathcal{Q}$ be the set of equivalence classes for $\sim$. If $[(x,y)], [(z,w)] \in \mathcal{Q}$, define

$$[(x,y)] + [(z,w)] = [(xw + yz, yw)] \quad \text{and} \quad [(x,y)][(z,w)] = [(xz, yw)] \ .$$

III.4.2.4. PROPOSITION. Addition and multiplication are well-defined binary operations on $\mathcal{Q}$ making $\mathcal{Q}$ into a field.

PROOF: Addition and multiplication are well defined by III.4.2.2.(ii). The associative law for multiplication, and the commutative laws for addition and multiplication, are obvious. The associative law for addition and the distributive law are straightforward calculations from the ring axioms of $\mathbb{Z}$, and are left to the reader.

The 0 element is $[(0,1)]$; it is obvious that $[(0,1)] + [(x,y)] = [(x,y)] + [(0,1)] = [(x,y)]$ for all x, y. Similarly, the negative of $[(x,y)]$ is $[(-x,y)]$, as is easily verified.

The 1 element is $[(1,1)]$. If $x \in \mathbb{N}$ (and of course $y \in \mathbb{N}$), then $[(x,y)]^{-1} = [(y,x)]$; if $(-x) \in \mathbb{N}$, then $[(x,y)]^{-1} = [(-y,-x)]$. These formulas can be easily verified by the reader. 👍

III.4.3. Exercises

III.4.3.1. Let $\mathbb{Z}_n$ be as in Exercise III.3.7.5.. Show that $\mathbb{Z}_n$ is a field if and only if n is prime. [If $n > 1$ is not prime, there are $x, y \in \mathbb{Z}_n$ with $x \neq 0$ and $y \neq 0$ but $xy = 0$. If n is prime, show that, for any $x \in \mathbb{Z}_n$, $x \neq 0$, the function $\phi_x : \mathbb{Z}_n \to \mathbb{Z}_n$ given by $\phi_x(y) = xy$ is injective. Use finiteness of $\mathbb{Z}_n$ to conclude that ϕ_x is also surjective.]

In particular, $\mathbb{Z}_2$ is a field with only two elements.

III.4.3.2. Let F be a field. Let ϕ be the ring-homomorphism from $\mathbb{Z}$ to F defined in Exercise III.3.7.6..

(a) If the ideal J of III.3.7.6.(c) is $\{0\}$, show that the injective ring-homomorphism ϕ extends to an injective ring-homomorphism from $\mathbb{Q}$ to F, also denoted ϕ. Show that $\phi(\mathbb{Q})$ is the smallest subfield of F.

(b) If the ideal J is $n\mathbb{Z}$, show that n is prime, and $\phi(\mathbb{Z}) \cong \mathbb{Z}_n$ is the smallest subfield of F. [Use Exercises III.4.3.1. and III.3.7.6.(c).]

The number n in case (b) is called the *characteristic* of F. In case (a), F is said to have *characteristic zero*. Thus, for example, $\mathbb{Z}_p$ has characteristic p and the fields of analysis, including $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, have characteristic zero.

III.4.3.3. (a) Show that $\mathbb{Z} + \mathbb{Z}\sqrt{2} = \{x + y\sqrt{2} : x, y \in \mathbb{Z}\}$ is a subring of $\mathbb{R}$.

(b) Show that $\mathbb{Q} + \mathbb{Q}\sqrt{2} = \{x + y\sqrt{2} : x, y \in \mathbb{Q}\}$ is a subfield of $\mathbb{R}$.

(c) Do the same for the subsets $\mathbb{Z} + \mathbb{Z}i$ and $\mathbb{Q} + \mathbb{Q}i$ of $\mathbb{C}$. Generalize. What if $\sqrt{2}$ is replaced by $\sqrt[3]{2}$? By π?

(d) If $\mathbb{Q} + \mathbb{Q}\sqrt{2}$ is regarded as a subset of $\mathbb{R}$, the ordering on $\mathbb{R}$ induces an ordering on $\mathbb{Q} + \mathbb{Q}\sqrt{2}$ making it into an ordered field. Can $\mathbb{Q} + \mathbb{Q}i$ be made into an ordered field? [If P is a positive cone in $\mathbb{Q} + \mathbb{Q}i$, then either $i \in P$ or $(-i) \in P$; in either case $(-1) \in P$.]

III.4.3.4. Rings of Fractions. Here is a construction which generalizes the construction of $\mathbb{Q}$ from $\mathbb{Z}$. Let R be a commutative ring, and D a nonempty subset of R which is closed under multiplication and which does not contain 0 or any zero divisor (the construction can be done more generally, but then becomes rather artificial and degenerate). We will embed R into a larger unital commutative ring $D^{-1}R$, in which the elements of D have multiplicative inverses.

(a) Put a relation $\sim$ on $R \times D$ by setting $(x_1, y_1) \sim (x_2, y_2)$ if $x_1 y_2 = x_2 y_1$. Show that $\sim$ is an equivalence relation. [Use Exercise 14.]

(b) Let $D^{-1}R$ be the set of $\sim$-equivalence classes. Show that $[(x, y)] + [(z, w)] = [(xw + yz, yw)]$ and $[(x, y)][(z, w)] = [(xz, yw)]$ give well-defined binary operations making $D^{-1}R$ into a unital commutative ring. [Think of $[(x, y)]$ as representing $\frac{x}{y}$.]

(c) Show that $\iota : R \to D^{-1}R$ given by $\iota(x) = [(xy, y)]$ for $y \in D$ is a well-defined injective ring-homomorphism.

(d) Show that every element of D has a multiplicative inverse in $D^{-1}R$.

The ring $D^{-1}R$ is called the *ring of fractions* of R with denominator D. In the special case where R is an integral domain (Exercise III.3.7.8.) and D is the set of nonzero elements of R, $D^{-1}R$ is a field, called the *quotient field*, or *field of fractions*, of R.

III.4.3.5. Use the notation and terminology of Exercise III.4.3.4..

(a) Describe $D^{-1}\mathbb{Z}$, where D is the set of powers of 2 in $\mathbb{N}$. This ring is called the *ring of dyadic rationals*.

(b) Describe $D^{-1}\mathbb{Z}$, where D is the set of odd numbers in $\mathbb{N}$. This ring is called the *ring of integers localized at 2*.

(c) Describe the quotient field of the ring $\mathbb{R}[X]$ of Exercise III.6.3.3. (cf. Exercise ()). This quotient field, usually denoted $\mathbb{R}(X)$, is, by slight abuse of language, called the *field of rational functions* in one variable with real coefficients. Explain why it is a slight abuse of language to call the elements of this field "functions." [Consider the "functions" $f(X) = X + 1$ and $g(X) = \frac{X^2-1}{X-1}$.]

III.4.3.6. Show that a field F can be made into an ordered field if and only if 0 is not a sum of nonzero squares in F (such an F is said to be *formally real*). [The set S of all quotients of sums of nonzero squares is closed under addition, multiplication, and reciprocals. Using Zorn's Lemma, embed S into a maximal set closed under addition, multiplication, and reciprocals, and not containing 0.] This is a version of the *Artin-Schreier Theorem*: every formally real field can be embedded in a real-closed field.

III.4.3.7. If F is an ordered field, show that the homomorphism $\phi : \mathbb{Z} \to F$ of Exercise III.3.7.6. is injective, and hence by Exercise III.4.3.2.(a) extends to an embedding of $\mathbb{Q}$ into F. Show that the induced ordering on $\mathbb{Q}$ is the ordinary one.

III.5. The Real Numbers

III.5.1. Construction of $\mathbb{R}$: the Dedekind Cut Approach

There are two standard approaches to the construction of the real numbers: via Dedekind cuts or via Cauchy sequences. (There is another nice approach using nonstandard analysis (III.10.8.12.).) Both are worthwhile, and each has uses and leads to insights about $\mathbb{R}$ not as easily obtainable from the other. The Dedekind cut approach (originally due, not surprisingly, to R. DEDEKIND in ()) is arguably slightly simpler, so we describe it first, in this section.

III.5.1.1. The idea is that if a is a real number, then a "cuts" the set of rational numbers into two sets $L_a = \{x \in \mathbb{Q} : x < a\}$ and $U_a = \{x \in \mathbb{Q} : x \geq a\}$ (if a is irrational, then $U_a = \{x \in \mathbb{Q} : x > a\}$; if a is rational, we arbitrarily put it in U_a instead of L_a). Every element of L_a is smaller than every element of U_a, and we have $\sup(L_a) = \inf(U_a) = a$ (). Conversely, if (L, U) is a division of $\mathbb{Q}$ into two nonempty subsets with each element of L less than every element of U, by completeness there will be a real number a, e.g. $\sup(L)$ (), such that $x \leq a$ for all $x \in L$ and $a \leq y$ for all $y \in U$; a must be unique by density of $\mathbb{Q}$ in $\mathbb{R}$, i.e. $a = \sup(L) = \inf(U)$, so that $L = L_a$ (provided L has no largest element) and $U = U_a$. Thus there is a one-one correspondence between real numbers and such pairs of subsets of $\mathbb{Q}$, which are called *Dedekind cuts*.

We may turn this argument around to *define* real numbers to be Dedekind cuts. (Technically, this will give us a model for $\mathbb{R}$.) In fact, the entire algebraic and order structure on $\mathbb{R}$ can be described directly in terms of the cuts in a fairly simple manner.

It turns out to be technically much easier to work only with positive numbers at first, i.e. to construct a model $\mathcal{P}$ of the set $\mathbb{P}$ of positive real numbers using Dedekind cuts in the set $\mathbb{Q}_+$ of positive rational numbers. The model for the whole set of real numbers is then constructed from $\mathcal{P}$ in a manner similar to the construction of $\mathbb{Z}$ from $\mathbb{N}$. (This approach is used in [Lan51].)

III.5.1.2. DEFINITION. (i) A subset S of $\mathbb{Q}_+ = \{x \in \mathbb{Q} : x > 0\}$ is *hereditary* if $x \in S$, $y \in \mathbb{Q}_+$, $y < x$ imply $y \in S$.

(ii) A *lower set* in $\mathbb{Q}_+$ is a nonempty hereditary subset of $\mathbb{Q}_+$ which is bounded above and has no largest element.

(iii) An *upper set* in $\mathbb{Q}_+$ is a subset U of $\mathbb{Q}_+$ such that $\mathbb{Q}_+ \setminus U$ is a lower set in $\mathbb{Q}_+$.

(iv) A *Dedekind cut* in $\mathbb{Q}_+$ is a pair (L, U), where L is a lower set in $\mathbb{Q}_+$ and $U = \mathbb{Q}_+ \setminus L$.

The next simple observation is obvious, but important.

III.5.1.3. PROPOSITION. If (L, U) is a Dedekind cut in $\mathbb{Q}_+$, $x \in L$, and $y \in U$, then $x < y$.

PROOF: We must have $x < y$ since L is hereditary. 👍

III.5.1.4. EXAMPLES. (i) If $a \in \mathbb{Q}_+$, set $L_a = \{x \in \mathbb{Q}_+ : x < a\}$ and $U_a = \{x \in \mathbb{Q}_+ : x \geq a\} = \mathbb{Q}_+ \setminus L_a$. L_a is nonempty since $\frac{1}{n} < a$ for some $n \in \mathbb{N}$ by (), and L_a is bounded above by a. L_a has no largest element since if $x \in L_a$, then $y = (a+x)/2 \in L_a$ and $x < y$. L_a is hereditary by transitivity of $<$. Thus L_a is a lower set, U_a is an upper set, and (L_a, U_a) is a Dedekind cut in $\mathbb{Q}_+$. The set L_a is often denoted a^*.

(ii) It is not *a priori* obvious that there are Dedekind cuts in $\mathbb{Q}_+$ not of the form (L_a, U_a) for $a \in \mathbb{Q}_+$, but it turns out that there are very many which are not. Here is an example which is at least not obviously of this form. Let

$$L = \{x \in \mathbb{Q}_+ : x^2 < 2\}$$

and $U = \mathbb{Q}_+ \setminus L$. L is hereditary since if $0 < y < x$, then $y^2 < x^2$. $1 \in L$, so L is nonempty, and L is bounded above since if $x \in L$, we have $x < 2$ since $x^2 < 2 < 2^2$ ().

We show that L has no largest element. If $x \in L$, set $z = \frac{4+3x}{3+2x}$. Then

$$z - x = \frac{4 - 2x^2}{3 + 2x} > 0 \quad \text{and} \quad 2 - z^2 = \frac{2 - x^2}{(3 + 2x)^2} > 0$$

by a simple computation; thus $x < z$ and $z \in L$.

Although there is no obvious $a \in \mathbb{Q}_+$ for which $L = L_a$, it takes a little work to show that in fact there is no such a (I.6.5.8.).

(iii) The only hereditary subset of $\mathbb{Q}_+$ which is not bounded above is all of $\mathbb{Q}_+$. For if S is such a set, and $y \in \mathbb{Q}_+$, then y is not an upper bound for S, so there is an $x \in S$ with $y < x$. But then $y \in S$ since S is hereditary, so $S = \mathbb{Q}_+$. Therefore a nonempty proper subset of $\mathbb{Q}_+$ which is hereditary and has no largest element is automatically a lower set.

(iv) An arbitrary union or intersection of hereditary subsets of $\mathbb{Q}_+$ is hereditary. An arbitrary union of lower sets is either $\mathbb{Q}_+$ or a lower set. A finite intersection of lower sets is a lower set. However, an arbitrary intersection of lower sets is not necessarily a lower set: it can be empty (e.g. $\cap_n L_{1/n} = \emptyset$) or it may have a largest element (e.g. $\cap_n L_{1+1/n}$).

(v) If S is a (necessarily proper) hereditary subset of $\mathbb{Q}_+$ which has a largest element a, let $L = S \setminus \{a\}$. Then L is nonempty ($\frac{1}{n} \in L$ for sufficiently large n), bounded above (by a), and hereditary. L also has no largest element since if $x \in L$, then $y = (x + a)/2$ is in L and $x < y$. Thus L is a lower set. In fact, L is exactly L_a of (i).

Here is a useful technical fact about Dedekind cuts:

III.5.1.5. THEOREM. Let (L, U) be a Dedekind cut in $\mathbb{Q}_+$, and $\epsilon \in \mathbb{Q}_+$. Then there exist $x \in L$ and $y \in U$, y not minimal in U, with $y - x = \epsilon$.

PROOF: Fix $a \in L$ and $b \in U$. The set $\{a + n\epsilon : n \in \mathbb{N}\}$ is unbounded in $\mathbb{Q}_+$ by (), hence there is an n with $a + n\epsilon > b$, and $a + n\epsilon \in U$ for this n. Thus there is a smallest n for which $a + n\epsilon \in U$. For this smallest n, set $x = a + (n-1)\epsilon$ and $y = n\epsilon$. To insure that y is not the smallest element of U, replace x and y by $x + \frac{1}{m}$ and $y + \frac{1}{m}$, where $m \in \mathbb{N}$ is large enough that $x + \frac{1}{m} \in L$. 👍

We now make our main definition:

III.5.1.6. DEFINITION. $\mathcal{P}$ denotes the set of all lower subsets of $\mathbb{Q}_+$.

Ordering on $\mathcal{P}$

III.5.1.7. $\mathcal{P}$ is naturally ordered by inclusion. If $L, L' \in \mathcal{P}$, write $L \leq L'$ if $L \subseteq L'$, and $L < L'$ if $L \leq L'$ and $L \neq L'$, i.e. if L is a proper subset of L'. This ordering is antisymmetric: if $L \leq L'$ and $L' \leq L$, then $L = L'$, or alternatively, at most one of $L = L'$, $L < L'$ and $L' < L$ holds.

Transitivity of this ordering is obvious.

As usual, we will also write $L' \geq L$ if $L \leq L'$, and $L' > L$ if $L < L'$.

III.5.1.8. PROPOSITION. The ordering $\leq$ is a total ordering on $\mathcal{P}$, i.e. if $L, L' \in \mathcal{P}$, then either $L \leq L'$ or $L' \leq L$. Alternatively, exactly one of $L < L'$, $L = L'$, $L' < L$ holds.

PROOF: If $L = L'$, there is nothing to prove. Suppose there is an $a \in L'$ with $a \notin L$. Thus $a \in \mathbb{Q}_+ \setminus L$, so $x < a$ for all $x \in L$ by III.5.1.3.. But any such x is in L' since L' is hereditary, so $L \subseteq L'$. 👍

Completeness of this ordering is deceptively simple:

III.5.1.9. PROPOSITION. If $\mathcal{L} = \{L_i : i \in I\}$ is a nonempty subset of $\mathcal{P}$ which is bounded above, then $\mathcal{L}$ has a supremum (least upper bound) in $\mathcal{P}$.

PROOF: $\mathcal{L}$ is bounded above if (and only if) there is an $L' \in \mathcal{P}$ with $L_i \subseteq L'$ for all $i \in I$. Fix such an L'. Let $L = \cup_{i \in I} L_i$. Since $L \subseteq L'$, L is a proper subset of $\mathbb{Q}_+$, hence is a lower set by III.5.1.4.(iii)–(iv). We have $L_i \leq L$ for all i, and if $L'' \in \mathcal{P}$ with $L_i \leq L''$ for all i, then $L \leq L''$, so L is the least upper bound of $\mathcal{L}$ in $\mathcal{P}$. 👍

Infima can be done similarly, but there is a slight technical glitch:

III.5.1.10. PROPOSITION. If $\mathcal{L} = \{L_i : i \in I\}$ is a nonempty subset of $\mathcal{P}$ which is bounded below, then $\mathcal{L}$ has an infimum (greatest lower bound) in $\mathcal{P}$.

PROOF: $\mathcal{L}$ is bounded below if (and only if) there is an $L' \in \mathcal{P}$ with $L' \subseteq L_i$ for all $i \in I$. Fix such an L'. Let $S = \cap_{i \in I} L_i$. Since $L' \subseteq S$, S is nonempty, and is a proper hereditary subset of $\mathbb{Q}_+$. But S could have a largest element. If S has no largest element, set $L = S$, and if S has a largest element a, set $L = S \setminus \{a\}$. In either case L is a lower set (cf. III.5.1.4.(v)). We have $L \leq L_i$ for all i, and if $L'' \in \mathcal{P}$ with $L'' \leq L_i$ for all i, then $L'' \leq L$, so L is the greatest lower bound of $\mathcal{L}$ in $\mathcal{P}$. 👍

III.5.1.11. So far, we could have equally well worked with Dedekind cuts on all of $\mathbb{Q}$, and defined $\mathbb{R}$ as an ordered set; restriction to positive numbers only has no advantages to this point. But defining the algebraic operations on $\mathbb{R}$ is a little messy, and it is much cleaner and easier to work with $\mathcal{P}$ instead.

Algebraic Operations on $\mathcal{P}$

We first make a general definition of sums and products of subsets of $\mathbb{Q}_+$:

III.5.1.12. DEFINITION. Let A and B be subsets of $\mathbb{Q}_+$. Define

$$A + B = \{x + y : x \in A, y \in B\}$$

$$AB = \{xy : x \in A, y \in B\} \ .$$

Note that $A + B$ and AB are subsets of $\mathbb{Q}_+$ since $\mathbb{Q}_+$ is closed under addition and multiplication.

III.5.1.13. **Caution:** Some authors, especially in older references or in the context of Boolean algebras, write $A + B$ for $A \cup B$ and AB for $A \cap B$. We will never use these meanings of $A + B$ and AB. Note that $A + B$ (as we have defined it) and $A \cup B$ are quite different in general, as are AB and $A \cap B$.

III.5.1.14. PROPOSITION. Let L and L' be lower sets in $\mathbb{Q}_+$. Then $L + L'$ and LL' are lower sets in $\mathbb{Q}_+$.

PROOF: $L + L'$ and LL' are clearly nonempty. If m and n are upper bounds for L and L' respectively, then $m + n$ and mn are upper bounds for $L + L'$ and LL' respectively by ().

We show $L + L'$ is hereditary. If $z \in L + L'$ and $w \in \mathbb{Q}_+$, $w < z$, write $z = x + y$ for $x \in L$, $y \in L'$. If $w \leq y$, there is an $n \in \mathbb{N}$ such that $\frac{1}{n} < w$ and $\frac{1}{n} < x$; then $\frac{1}{n} \in L$ and $0 < w - \frac{1}{n} < y$, so $y - \frac{1}{n} \in L'$ and $w = \frac{1}{n} + \left(w - \frac{1}{n}\right) \in L + L'$. On the other hand, if $y < w$, we have $0 < w - y < z - y = x$, so $(w - y) \in L$, and $w = (w - y) + y \in L + L'$ again.

The proof that LL' is hereditary is easier. If $z \in LL'$ and $w \in \mathbb{Q}_+$, $w < z$, write $z = xy$ for $x \in L$, $y \in L'$. Then $w < xy$, so $0 < \frac{w}{y} < x$ and hence $\frac{w}{y} \in L$ and $w = \frac{w}{y} y \in LL'$.

Finally, we show that $L + L'$ and LL' have no maximal elements. Let $z \in L + L'$, and write $z = x + y$ for $x \in L$, $y \in L'$. Since L has no largest element, there is an $r \in L$ with $x < r$. Then $r + y \in L + L'$, and $z < r + y$. The proof for LL' is similar.

It is now clear how to define addition and multiplication in $\mathcal{P}$:

III.5.1.15. DEFINITION. Define binary operations $+$ and $\cdot$ on $\mathcal{P}$ to be sum and product as defined in III.5.1.12..

As usual, we will write the product of L and L' either as $L \cdot L'$ or LL'. We will use the usual convention of grouping multiplication before addition, e.g. $LL' + L'' = (LL') + L''$.

The usual algebraic properties of addition and multiplication in $\mathcal{P}$ follow easily from those in $\mathbb{Q}$:

III.5.1.16. PROPOSITION. Let $L, L', L'' \in \mathcal{P}$. Then

(i) $(L + L') + L'' = L + (L' + L'')$ and $(LL')L'' = L(L'L'')$. *(Associative Law)*

(ii) $L + L' = L' + L$ and $LL' = L'L$. *(Commutative Law)*

(iii) $L(L' + L'') = LL' + LL''$. *(Distributive Law)*

PROOF: The proofs are quite easy, and all are essentially identical. We prove the associative law for addition, and leave the others to the reader. We have

$$(L + L') + L'' = \{(x + y) + z : x \in L,\ y \in L',\ z \in L''\}$$

$$L + (L' + L'') = \{x + (y + z) : x \in L,\ y \in L',\ z \in L''\}$$

and by the associative law in $\mathbb{Q}$ (A1), these two subsets of $\mathbb{Q}_+$ are identical. 👍

III.5.1.17. PROPOSITION. The element $1^* = L_1 = \{x \in \mathbb{Q}_+ : x < 1\}$ is a multiplicative identity for $\mathcal{P}$, i.e. $1^* \cdot L = L$ for any $L \in \mathcal{P}$.

PROOF: Let $L \in \mathcal{P}$. Since L is hereditary, it is clear that $1^* \cdot L \leq L$. On the other hand, if $x \in L$, choose $y \in L$ with $x < y$ (since L has no largest element); then $\frac{x}{y} < 1$, so $\frac{x}{y} \in 1^*$, and $x = \frac{x}{y}y \in 1^* \cdot L$. 👍

On the lower sets coming from elements of $\mathbb{Q}_+$ as in III.5.1.4.(i), addition and multiplication in $\mathcal{P}$ agree with addition and multiplication in $\mathbb{Q}_+$:

III.5.1.18. PROPOSITION. Let $a, b \in \mathbb{Q}_+$, and $L_a = \{x \in \mathbb{Q}_+ : x < a\}$ and $L_b = \{x \in \mathbb{Q}_+ : x < b\}$ the corresponding elements of $\mathcal{P}$. Then $L_a + L_b = L_{a+b}$ and $L_a L_b = L_{ab}$.

PROOF: It is obvious from () that $L_a + L_b \leq L_{a+b} = \{x \in \mathbb{Q}_+ : x < a + b\}$. On the other hand, if $z \in L_{a+b}$, i.e. $z < a + b$, set $d = a + b - z$. Then $x = a - \frac{d}{2} \in L_a$ and $y = b - \frac{d}{2} \in L_b$, and $z = x + y \in L_a + L_b$.

For the product, it follows from () that $L_a L_b \leq L_{ab}$. If $z \in L_{ab}$, we have $\frac{z}{b} < a$. Choose $x \in \mathbb{Q}_+$ with $\frac{z}{b} < x < a$, so $x \in L_a$. Then $\frac{z}{x} < b$, so $y = \frac{z}{x} \in L_b$, and $z = xy \in L_{ab}$. 👍

Multiplicative Inverses

III.5.1.19. Every element of $\mathcal{P}$ has a multiplicative inverse, but the argument has a slight subtlety. Suppose $L \in \mathcal{P}$, and $U = \mathbb{Q}_+ \setminus L$ is the corresponding upper set. Since inversion reverses order (), we should have

$$L^{-1} = \{x^{-1} : x \in U\}\ .$$

However, this may not be a lower set since it might have a largest element. In fact, this will happen if and only if $L = L_a$ for some $a \in \mathbb{Q}_+$ (we then obtain the set $\{x \in \mathbb{Q}_+ : x \leq a^{-1}\}$). In this case, we take $L^{-1} = \{x \in \mathbb{Q}_+ : x < a^{-1}\} = L_{a^{-1}}$. Thus we define L^{-1} differently depending on whether or not $L = L_a$ for some $a \in \mathbb{Q}_+$.

III.5.1.20. PROPOSITION. If $L \in \mathcal{P}$, and L^{-1} is defined as above, then L^{-1} is a lower set and $LL^{-1} = 1$.

PROOF: Clearly L^{-1} is nonempty. Suppose $x \in L^{-1}$, $y \in \mathbb{Q}_+$, and $y < x$. Then $x^{-1} < y^{-1}$ and since $x \notin L$, we have $y \notin L$ also since L is hereditary. Thus $y \in U$ and $y^{-1} \in L^{-1}$. So L^{-1} is hereditary. If $x \in L$, then $y^{-1} < x^{-1}$ for all $y \in U$ by () and III.5.1.3.; hence x^{-1} is an upper bound for L^{-1}. If $S = \{x^{-1} : x \in U\}$ has a largest element b, then $L^{-1} = S \setminus \{b\}$. If $x \in L^{-1}$, i.e. $x^{-1} \in U$ and $x < b$, and $y \in \mathbb{Q}_+$ with $x < y < b$ (e.g. $y = (x + b)/2$), then $y^{-1} > b^{-1}$, so $y \in S$, and $y \neq b$, so $y \in L^{-1}$; thus L^{-1} has no largest element in any case. So $L^{-1} \in \mathcal{P}$.

If $x \in L$ and $y \in L^{-1}$, then $y^{-1} \in U$, so $x < y^{-1}$ and therefore $y < x^{-1}$, so $xy < xx^{-1} = 1$. Thus $LL^{-1} \leq 1$.

The opposite inequality is trickier. We use III.5.1.5.. Fix $a \in L$. If $z < 1$, choose $x \in L$ and $y \in U$, y not minimal in U, with $y - x = a(1 - z)$. Then $\frac{x}{y} \in LL^{-1}$, and

$$1 - \frac{x}{y} = \frac{a}{y}(1 - z) < 1 - z$$

since $a < y$. Thus $z < \frac{x}{y}$ and $z \in LL^{-1}$. 👍

There are no additive inverses in $\mathcal{P}$. There is not even an additive identity:

III.5.1.21. PROPOSITION. Let $L, L' \in \mathcal{P}$. Then $L < L + L'$.

PROOF: If $x \in L$ and $y \in L'$, then $x < x + y \in L + L'$, so $x \in L + L'$ and $L \leq L + L'$. Set $U = \mathbb{Q}_+ \setminus L$. Fix $z \in L'$, and by III.5.1.5. choose $x \in L$, $y \in U$ with $y - x = z$. Then $y = x + z \in L + L'$, but $y \notin L$. 👍

Cancellation

Cancellation for multiplication follows immediately from the existence of inverses:

III.5.1.22. PROPOSITION. If $L, L', L'' \in \mathcal{P}$ and $LL' = LL''$, then $L' = L''$.

PROOF: We have $L^{-1}(LL') = L^{-1}(LL'')$. But we have

$$L^{-1}(LL') = (L^{-1}L)L' = 1^* \cdot L' = L'$$

and similarly $L^{-1}(LL'') = L''$, so $L' = L''$. 👍

Additive cancellation takes more work since there are no additive inverses. We first show that $<$ is translation-invariant:

III.5.1.23. PROPOSITION. Let $L, L', L'' \in \mathcal{P}$. If $L < L'$, then $L + L'' < L' + L''$.

PROOF: It is obvious that $L + L'' \leq L' + L''$. Fix $a \in L' \setminus L$, and $b \in L'$ with $a < b$. Let $U'' = \mathbb{Q}_+ \setminus L''$, and $x \in L''$, $y \in U''$ with $y - x = b - a$ (III.5.1.5.). Then $z = b + x \in L' + L''$, but if $z = s + t$ for $s \in L$, $t \in L''$, we have $s < a$, so $b - s > b - a = y - x$, so $t = z - s = (b - s) + x > (y - x) + x = y$, a contradiction. Thus $z \in (L' + L'') \setminus (L + L'')$. 👍

Additive cancellation then follows:

III.5.1.24. COROLLARY. Let $L, L', L'' \in \mathcal{P}$. If $L + L'' < L' + L''$, then $L < L'$. If $L + L'' = L' + L''$, then $L = L'$.

PROOF: Either $L < L'$, $L = L'$, or $L > L'$. In these cases $L + L'' < L' + L''$, $L + L'' = L' + L''$, and $L + L' > L' + L''$ respectively. 👍

Subtraction

We also get subtraction defined to the extent possible in light of III.5.1.21.:

III.5.1.25. PROPOSITION. Let $L, L' \in \mathcal{P}$. If $L < L'$, then there is a unique $L'' \in \mathcal{P}$ such that $L + L'' = L'$.

PROOF: Uniqueness follows immediately from III.5.1.24.. We show existence. Set $U = \mathbb{Q}_+ \setminus L$, and let

$$L'' = \{x - y : x \in L',\ y \in U,\ y < x\} .$$

We show L'' is a lower set. If $y \in L' \setminus L$ and $x \in L'$, $y < x$, then $x - y \in L''$, so L'' is nonempty. If $x \in L'$, $y \in U$, and $z \notin L'$, then $x - y < x < z$, so $z \notin L''$ and L'' is bounded. If $t \in L''$ and $0 < s < t$, write $t = x - y$ for $x \in L'$, $y \in U$, $y < x$; then $s + y < t + y \in L'$, and $s = (s + y) - y$, so $s \in L''$. Thus L'' is hereditary. Finally, if $t \in L''$ with $t = x - y$ for $x \in L''$, $y \in U$, $y < x$, let $w \in L'$, $x < w$; then $s = w - y \in L''$ and $t < s$, so L'' has no largest element. Thus $L'' \in \mathcal{P}$.

If $z \in L + L''$, write $z = w + (x - y)$ with $w \in L$, $x \in L'$, $y \in U$, $y < x$. Then $w < y$, so $z < w + (x - w) = x \in L'$, so $z \in L'$. Thus $L + L'' \leq L'$.

If $z \in L'$, let $b \in L'$, $z < b$, and by III.5.1.5. choose $x \in L$ and $y \in U$ with $y - x = b - z$. Then $z - x = b - y \in L''$, and $z = x + (z - x) \in L + L''$, $L' \leq L + L''$. 👍

Notation

We now streamline our notation for elements of $\mathcal{P}$, eliminating the clumsy L, L', etc.

III.5.1.26. By III.5.1.18., the map $\phi : \mathbb{Q}_+ \to \mathcal{P}$ given by $\phi(a) = L_a$ preserves addition and multiplication. It is obvious that ϕ is also order-preserving, i.e. if $a < b$, then $L_a < L_b$. We will identify $\mathbb{Q}_+$ with its image in $\mathcal{P}$, and just write a instead of $\phi(a)$, L_a, or a^*, unless there is possibility of confusion.

III.5.1.27. PROPOSITION. The image of $\mathbb{Q}_+$ is order-dense in $\mathcal{P}$, i.e. if $L, L' \in \mathcal{P}$ with $L < L'$, then there is an $a \in \mathbb{Q}_+$ with $L < a < L'$.

PROOF: Let $L < L'$. Then L is a proper subset of L'. Let $b \in L' \setminus L$, and let $a \in L'$ with $b < a$. Then $b \in L_a \setminus L$, so $L < L_a$, and $a \in L' \setminus L_a$, so $L_a < L'$. 👍

III.5.1.28. With this identification, we can say that if $L \in \mathcal{P}$, then $L = \{x \in \mathbb{Q}_+ : x < L\}$. (This statement seems strangely circular, but is not actually logically circular since it is not used as a *definition* of L.)

III.5.1.29. From now on, we will use lower-case letters to denote elements of $\mathcal{P}$ instead of L's. Thus we will write $a, b, x, y \in \mathcal{P}$, etc.; the a, b, x, y may denote elements of $\mathbb{Q}_+$, thought of as a subset of $\mathcal{P}$, or of $\mathbb{J}_+ = \mathcal{P} \setminus \mathbb{Q}_+$, called the set of *positive irrational numbers*. (We have not yet shown that this set is nonempty.) So, if $a \in \mathcal{P}$, we officially have

$$a = \{x \in \mathbb{Q}_+ : x < a\}$$

whether or not a itself is rational. (This statement is technically logically inconsistent if a is rational, since a has previously been defined to be something else, but we will gloss over this technicality since we are not regarding individual numbers as being any specific set-theoretic objects.)

Construction of the Model for $\mathbb{R}$ from $\mathcal{P}$

We now essentially copy the construction of $\mathbb{Z}$ from $\mathbb{N}$ to obtain a model for $\mathbb{R}$.

III.5.1.30. The construction is notationally easier if we adjoin the rational number 0 (which is not in $\mathcal{P}$) to $\mathcal{P}$. Write $\mathcal{P}_0$ for $\mathcal{P} \cup \{0\}$. We define $0 + x = x + 0 = x$ and $0 \cdot x = x \cdot 0 = 0$ for all $x \in \mathcal{P}_0$. This extends the addition and multiplication on $\mathbb{Q}_+ \cup \{0\}$, and the associative, commutative, and distributive laws continue to hold in $\mathcal{P}_0$. We also extend the ordering on $\mathcal{P}$ to $\mathcal{P}_0$ by setting $0 < x$ for all $x \in \mathcal{P}$.

III.5.1.31. DEFINITION. Define a relation on $\mathcal{P}_0 \times \mathcal{P}_0$ by setting $(x_1, y_1) \sim (x_2, y_2)$ if $x_1 + y_2 = x_2 + y_1$.

III.5.1.32. PROPOSITION. (i) The relation $\sim$ is an equivalence relation () on $\mathcal{P}_0 \times \mathcal{P}_0$.
(ii) If $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, then $(x_1 + z_1, y_1 + w_1) \sim (x_2 + z_2, y_2 + w_2)$.
(iii) If $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, then $(x_1z_1 + y_1w_1, x_1w_1 + y_1z_1) \sim (x_2z_2 + y_2w_2, x_2w_2 + y_2z_2)$.
(iv) $(x_1, y_1) \sim (x_2, y_2)$ and $(z_1, w_1) \sim (z_2, w_2)$, and $x_1 + w_1 < z_1 + y_1$, then $x_2 + w_2 < z_2 + y_2$.

PROOF: (i): If $(x_1, y_1) \sim (x_2, y_2)$, it is trivial that $(x_2, y_2) \sim (x_1, y_1)$. It is also trivial that $(x, y) \sim (x, y)$ for any x, y. If $(x_1, y_1) \sim (x_2, y_2)$ and $(x_2, y_2) \sim (x_3, y_3)$, we have

$$(x_1 + y_3) + (x_2 + y_2) = (x_1 + y_2) + (x_2 + y_3) = (x_2 + y_1) + (x_3 + y_2) = (x_3 + y_1) + (x_2 + y_2)$$

by the associative and commutative laws, and hence $x_1 + y_3 = x_3 + y_1$ by cancellation. Thus $(x_1, y_1) \sim (x_3, y_3)$.
(ii): By the associative and commutative laws, we have

$$(x_1 + z_1) + (y_2 + w_2) = (x_1 + y_2) + (z_1 + w_2) = (x_2 + y_1) + (z_2 + w_1) = (x_2 + z_2) + (y_1 + w_1)\ .$$

(iii): We have

$$(x_1z_1 + y_1w_1, x_1w_1 - y_1z_1) \sim (x_1z_1 + y_2z_1 + x_2w_1 + y_1w_1, x_1w_1 + x_2w_1 + y_2z_1 + y_1z_1)$$

$$= ((x_1+y_2)z_1+(x_2+y_1)w_1, x_1w_1+x_2w_1+y_2z_1+y_1z_1) = ((x_2+y_1)z_1+(x_1+y_2)w_1, x_1w_1+x_2w_1+y_2z_1+y_1z_1)$$

$$= (x_2z_1 + y_1z_1 + x_1w_1 + y_2w_1, x_1w_1 + x_2w_1 + y_2z_1 + y_1z_1) \sim (x_2z_1 + y_2w_1, x_2w_1 + y_2z_1) \ .$$

An almost identical argument shows $(x_2z_1 + y_2w_1, x_2w_1 + y_2z_1) \sim (x_2z_2 + y_2w_2, x_2w_2 + y_2z_2)$.

(iv): We have

$$(x_2 + w_2) + (y_1 + z_1) = (x_2 + y_1) + (z_1 + w_2) = (x_1 + y_2) + (z_2 + w_1)$$

$$= (x_1 + w_1) + (y_2 + z_2) < (z_1 + y_1) + (y_2 + z_2) = (z_2 + y_2) + (y_1 + z_1)$$

so $x_2 + w_2 < z_2 + y_2$ by cancellation. 👍

We will write $[(x, y)]$ for the equivalence class of the pair (x, y). We will think of $[(x, y)]$ as representing the real number $x - y$; the equivalence relation $\sim$ is designed so that every representative of the equivalence class represents the same real number.

III.5.1.33. Definition. Let $\mathcal{R}$ be the set of equivalence classes of ordered pairs. For $[(x, y)], [(z, w)] \in \mathcal{R}$, define

$$[(x, y)] + [(z, w)] = [(x + z, y + w)]$$

$$[(x, y)] \cdot [(z, w)] = [(xz + yw, xw + yz)]$$

$$[(x, y)] < [(z, w)] \text{ if } x + w < z + y \ .$$

Note that $+$, $\cdot$, and $<$ are well defined on $\mathcal{R}$ by III.5.1.32..

III.5.1.34. Theorem. $(\mathcal{R}, +, [(0, 0)], [(1, 0)])$ is a model for $\mathbb{R}$.

Proof: It is obvious that once we know that addition and multiplication of equivalence classes is well defined, $\mathcal{R}$ inherits associativity, commutativity, and the distributive law from $\mathcal{P}_0$.

We have $[(0, 0)]+[(x, y)] = [(0+x, 0+y)] = [(x, y)]$ for all $x, y \in \mathcal{P}_0$, and similarly $[(x, y)]+[(0, 0)] = [(x, y)]$ for all x, y; thus (A3) holds with 0-element $[(0, 0)]$. For $x, y \in \mathcal{P}_0$, we have $[(x, y)]+[(y, x)] = [(x+y, y+x)] = [(0, 0)]$, and similarly $[(y, x)] + [(x, y)] = [(0, 0)]$, so (A4) holds with $-[(x, y)] = [(y, x)]$.

Suppose $A \subseteq \mathcal{Z}$ contains $[(0, 0)]$, is closed under adding $[(1, 0)]$ and under negations. Then $A \cap \mathcal{N}_0 = \mathcal{N}_0$ since $\mathcal{N}_0$ is a model for $\mathbb{N}_0$, so $[(x, 0)] \in A$ for all $x \in \mathbb{N}_0$. Then also $-[(x, 0)] = [(0, x)] \in A$ for all $x \in \mathbb{N}_0$. If $x, y \in \mathbb{N}_0$, either $x = y$, or there is a $z \in \mathbb{N}_0$ with $x + z = y$, or there is a $w \in \mathbb{N}_0$ with $x = y + w$. In the first case $[(x, y)] = [(0, 0)] \in A$. If $x + z = y$, then $[(x, y)] = [(0, z)] \in A$. And if $x = y + w$, then $[(x, y)] = [(w, 0)] \in A$. Thus (Cyc) holds.

It is obvious that (Fr) holds, since we have shown that $[(x, 0)] \neq [(0, 0)]$ for any $x \in \mathbb{N}_0$, $x \neq 0$. 👍

III.5.1.35. Note that each equivalence class in $\mathcal{R}$ contains exactly one pair (x, y), where either x or y (or both) is 0. Thus we could have taken this set of ordered pairs as our model of $\mathbb{R}$. This model has the advantage that the elements are more concrete; they are specific ordered pairs rather than equivalence classes. But defining addition and proving associativity are somewhat messy in this approach.

III.5.1.36. We could have done the whole construction using $\mathcal{P}$ instead of $\mathcal{P}_0$. In this case, the 0 element is $[(1,1)]$, and the 1-element is $[(1+1,1)]$. The positive real number x is represented by $[(x+1,1)]$. The approach of III.5.1.35. can also be done using $\mathcal{P}$; details are left to the reader.

III.5.2. Construction of $\mathbb{R}$: the Cauchy Sequence Approach

The other standard construction of $\mathbb{R}$ is via equivalence classes of Cauchy sequences of rational numbers. This approach was pioneered by CANTOR.

III.5.2.1. The idea behind this construction is that every real number is the limit of a convergent sequence of rational numbers. CAUCHY found a criterion for convergence of a sequence which involves only the terms of the sequence, with no explicit reference to the limit; a sequence satisfying this criterion is called a *Cauchy sequence*. Many different sequences can converge to the same limit, and CAUCHY also found a criterion involving only the terms of the sequence for when this happens; Cauchy sequences satisfying this criterion are called *equivalent*. There is thus a one-one correspondence between real numbers and equivalence classes of Cauchy sequences of rational numbers.

The argument can be turned around to *define* real numbers to be equivalence classes of Cauchy sequences of rational numbers (this was CANTOR'S idea). The algebraic and order properties of real numbers can also be cleanly expressed in terms of these equivalence classes.

We first recall the definition of a convergent sequence. We will phrase it in terms of rational numbers, since we do not assume we have yet constructed anything beyond $\mathbb{Q}$. For a fuller discussion and motivation of the definition, see (). Recall that $\mathbb{Q}_+ = \{x \in \mathbb{Q} : x > 0\}$.

III.5.2.2. DEFINITION. Let (x_n) be a sequence in $\mathbb{Q}$, and $x \in \mathbb{Q}$. Then (x_n) *converges* to x, or x is the *limit* of the sequence (x_n), written $x_n \to x$ or $x = \lim\limits_{n\to\infty} x_n$, if for every $\epsilon \in \mathbb{Q}_+$ there is a $K \in \mathbb{Q}$ such that $|x_n - x| < \epsilon$ for all $n \in \mathbb{N}$, $n > K$.

A general sequence in $\mathbb{Q}$ need not converge; in fact, "most" sequences do not converge. It can be readily proved (IV.1.3.13.) that a sequence cannot have more than one limit, justifying reference to *the* limit of a convergent sequence.

CAUCHY'S observation is that convergent sequences have an intrinsic property expressible involving only the terms of the sequence:

III.5.2.3. DEFINITION. Let (x_n) be a sequence in $\mathbb{Q}$. Then (x_n) is a *Cauchy sequence* if, for every $\epsilon \in \mathbb{Q}_+$, there is a $K \in \mathbb{Q}$ such that $|x_n - x_m| < \epsilon$ for all $n, m > K$.

III.5.2.4. PROPOSITION. Let (x_n) be a sequence in $\mathbb{Q}$. If (x_n) converges, then (x_n) is a Cauchy sequence.

PROOF: Suppose $x_n \to x$. Let $\epsilon \in \mathbb{Q}_+$. Then there is a $K \in \mathbb{Q}$ such that $|x_n - x| < \frac{\epsilon}{2}$ for all $n > K$. If $n, m > K$, then by the triangle inequality we have

$$|x_n - x_m| \le |x_n - x| + |x - x_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$



III.5.2.5. CAUCHY'S idea was that, conversely, every Cauchy sequence should converge. This turns out to be true in $\mathbb{R}$ by completeness, but is false in $\mathbb{Q}$ (although this is not entirely obvious; cf. IV.1.12.7.). Thus the idea is to define real numbers to be limits of Cauchy sequences of rational numbers.

However, we also need to specify when two sequences (x_n) and (y_n) should converge to the same limit. This should happen if and only if x_n and y_n become arbitrarily close to each other as n becomes large. So we define:

III.5.2.6. DEFINITION. Let (x_n) and (y_n) be sequences in $\mathbb{Q}$ (not necessarily Cauchy sequences). Then (x_n) and (y_n) are *equivalent*, written $(x_n) \sim (y_n)$, if for every $\epsilon \in \mathbb{Q}_+$ there is a $K \in \mathbb{Q}$ such that $|x_n - y_n| < \epsilon$ for all $n > K$.

The term "equivalent" is justified:

III.5.2.7. PROPOSITION. The relation $\sim$ is an equivalence relation on the set of all sequences in $\mathbb{Q}$.

PROOF: It is trivial that $\sim$ is reflexive and symmetric. To prove transitivity, suppose (x_n), (y_n), and (z_n) are sequences in $\mathbb{Q}$ with $(x_n) \sim (y_n)$ and $(y_n) \sim (z_n)$. Let $\epsilon \in \mathbb{Q}_+$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - y_n| < \frac{\epsilon}{2}$ for all $n > K_1$, and a $K_2 \in \mathbb{Q}$ with $|y_n - z_n| < \frac{\epsilon}{2}$ for all $n > K_2$. Set $K = \max(K_1, K_2)$. If $n > K$, then by the triangle inequality

$$|x_n - z_n| \leq |x_n - y_n| + |y_n - z_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$



If (x_n) and (y_n) are sequences and $x_n = y_n$ for all but finitely many n, then $(x_n) \sim (y_n)$: take $K = \max\{n \in \mathbb{N} : x_n \neq y_n\}$.

Equivalent sequences converge to the same limit, and conversely:

III.5.2.8. PROPOSITION. Let (x_n) and (y_n) be sequences in $\mathbb{Q}$. If $(x_n) \sim (y_n)$ and $x_n \to x$, then $y_n \to x$. Conversely, if $x_n \to x$ and $y_n \to x$, then $(x_n) \sim (y_n)$.

PROOF: Let $\epsilon \in \mathbb{Q}_+$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x| < \frac{\epsilon}{2}$ for all $n > K_1$, and there is a $K_2 \in \mathbb{Q}$ such that $|x_n - y_n| < \frac{\epsilon}{2}$ for all $n > K_2$. Set $K = \max(K_1, K_2)$. If $n > K$, then

$$|y_n - x| \leq |y_n - x_n| + |x_n - x| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$

Conversely, suppose $y_n \to x$. Let $\epsilon \in \mathbb{Q}_+$, and let K_1 be as above. There is also a K_3 such that $|y_n - x| < \frac{\epsilon}{2}$ for all $n > K_3$. Set $K' = \max(K_1, K_3)$. If $n > K'$, then

$$|x_n - y_n| \leq |x_n - x| + |x - y_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$



Since a constant sequence clearly converges to its constant value (), we obtain:

III.5.2.9. COROLLARY. Let (x_n) be a sequence of rational numbers. Then (x_n) converges to the rational number r if and only if (x_n) is equivalent to the constant sequence (r).

A variation of III.5.2.8. for Cauchy sequences is useful:

III.5.2.10. PROPOSITION. Let (x_n) and (y_n) be sequences in $\mathbb{Q}$. If $(x_n) \sim (y_n)$ and (x_n) is a Cauchy sequence, then (y_n) is also a Cauchy sequence.

PROOF: Let $\epsilon \in \mathbb{Q}_+$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x_m| < \frac{\epsilon}{3}$ for all $n, m > K_1$, and there is a $K_2 \in \mathbb{Q}$ such that $|x_n - y_n| < \frac{\epsilon}{3}$ for all $n > K_2$. Set $K = \max(K_1, K_2)$. If $n, m > K$, then

$$|y_n - y_m| \leq |y_n - x_n| + |x_n - x_m| + |x_m - y_m| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon \ .$$



We now make our main definition:

III.5.2.11. DEFINITION. Let $\mathbf{R}$ be the set of equivalence classes of Cauchy sequences of rational numbers. Denote the class of a Cauchy sequence (x_n) by $[(x_n)]$. Denote by $\mathbf{Q}$ the set of equivalence classes of constant sequences.
There is an obvious one-one correspondence between $\mathbb{Q}$ and $\mathbf{Q}$.

Order Structure of R

The key to the order structure of $\mathbf{R}$ is the next result:

III.5.2.12. PROPOSITION. Let (x_n) and (y_n) be Cauchy sequences of rational numbers. If (x_n) and (y_n) are not equivalent, then there are $r, s \in \mathbb{Q}$ with $r < s$, and $K \in \mathbb{Q}$, such that either $x_n < r$ and $y_n > s$ for all $n > K$ or $y_n < r$ and $x_n > s$ for all $n > K$.

PROOF: If (x_n) and (y_n) are not equivalent, then there is an $\epsilon \in \mathbb{Q}_+$ such that, for every $K \in \mathbb{Q}$, there is an $n > K$ with $|x_n - y_n| \geq \epsilon$. For this ϵ, fix K such that $|x_n - x_m| < \frac{\epsilon}{3}$ and $|y_n - y_m| < \frac{\epsilon}{3}$ for all $n, m > K$. Fix $m > K$ with $|x_m - y_m| \geq \epsilon$. Then either $x_m - y_m \geq \epsilon$ or $y_m - x_m \geq \epsilon$. If $y_m - x_m \geq \epsilon$, Set $r = x_m + \frac{\epsilon}{3}$ and $s = y_m - \frac{\epsilon}{3}$. Then $r < s$, and for any $n > K$ we have $x_n < r$ and $y_n > s$. A similar argument shows that if $x_m - y_m \geq \epsilon$, and $r = y_m + \frac{\epsilon}{3}$, $s = x_m - \frac{\epsilon}{3}$, then $y_n < r < s < x_n$ for all $n > K$. 

III.5.2.13. PROPOSITION. Let (x_n) and (y_n) be sequences in $\mathbb{Q}$, and $r, s \in \mathbb{Q}$, $K \in \mathbb{Q}$ with $x_n < r < s < y_n$ for all $n > K$. If (z_n) and (w_n) are sequences in $\mathbb{Q}$ with $(z_n) \sim (x_n)$ and $(w_n) \sim (y_n)$, and $r', s' \in \mathbb{Q}$ with $r < r' < s' < s$, then there is a $K' \in \mathbb{Q}$ such that $z_n < r'$ and $w_n > s'$ for all $n > K'$.

PROOF: There is a $K_1 \in \mathbb{Q}$ such that $|x_n - z_n| < r' - r$ for all $n > K_1$, and $K_2 \in \mathbb{Q}$ such that $|y_n - w_n| < s - s'$ for all $n > K_2$. Set $K' = \max(K, K_1, K_2)$. If $n > K'$, we have $z_n = x_n + (z_n - x_n) < r + (r' - r) = r'$, and similarly $w_n = y_n - (y_n - w_n) > s - (s - s') = s'$. 

We can now define the order on $\mathbf{R}$:

III.5.2.14. DEFINITION. Let $[(x_n)], [(y_n)] \in \mathbf{R}$. Then $[(x_n)] < [(y_n)]$ if there are $r, s \in \mathbb{Q}$, $r < s$, and $K \in \mathbb{Q}$ such that $x_n < r$ and $y_n > s$ for all $n > K$.

By III.5.2.13., this order is well defined on equivalence classes. It is clearly transitive and antisymmetric. By III.5.2.12., it is a total order, i.e. trichotomy holds. As usual, write $[(x_n)] \leq [(y_n)]$ if $[(x_n)] < [(y_n)]$ or $[(x_n)] = [(y_n)]$, and define $>$ and $\geq$ accordingly. The induced ordering on $\mathbf{Q}$ is just the usual ordering on $\mathbb{Q}$, i.e. if $r, s \in \mathbb{Q}$, then $[(r)] < [(s)]$ if and only if $r < s$.

III.5.2.15. PROPOSITION. Let (x_n), (y_n) be Cauchy sequences in $\mathbb{Q}$. If there is a $K \in \mathbb{Q}$ with $x_n \leq y_n$ for all $n > K$, then $[(x_n)] \leq [(y_n)]$.

PROOF: Under the hypotheses, $[(x_n)] > [(y_n)]$ is impossible! 👍

III.5.2.16. The converse is also true by definition of the order. In fact, if $[(x_n)] < [(y_n)]$, then there are Cauchy sequences (z_n), (w_n) in $\mathbb{Q}$, with $(x_n) \sim (z_n)$ and $(y_n) \sim (w_n)$, such that $z_n < w_n$ for all n: just take $z_n = x_n$ and $w_n = y_n$ if $x_n < y_n$, and $z_n = 0$, $w_n = 1$ if $x_n \geq y_n$. Then $z_n = x_n$ and $w_n = y_n$ for all but finitely many n. The statement also clearly holds with $<$ replaced by $\leq$ throughout.

We then obtain density of $\mathbf{Q}$ in $\mathbf{R}$:

III.5.2.17. PROPOSITION. $\mathbf{Q}$ is order-dense in $\mathbf{R}$, i.e. if $[(x_n)] < [(y_n)]$, then there is a $t \in \mathbb{Q}$ with $[(x_n)] < [(t)] < [(y_n)]$.

PROOF: Let $r, s \in \mathbb{Q}$ with $r < s$, and $K \in \mathbb{Q}$, such that $x_n < r$ and $y_n > s$ for all $n > K$. Then $[(x_n)] \leq [(r)] < [(s)] \leq [(y_n)]$ by III.5.2.15.. If $t = (r + s)/2$, we have

$$[(x_n)] \leq [(r)] < [(t)] < [(s)] \leq [(y_n)] \ .$$

👍

Algebraic Structure of R

We would like to define addition and multiplication of equivalence classes of Cauchy sequences coordinatewise. The next proposition shows that addition is well defined in this way:

III.5.2.18. PROPOSITION. Let (x_n), (y_n), (z_n), (w_n) be sequences in $\mathbb{Q}$.

(i) If (x_n) and (y_n) are Cauchy sequences, so is $(x_n + y_n)$.

(ii) If $(x_n) \sim (z_n)$ and $(y_n) \sim (w_n)$, then $(x_n + y_n) \sim (z_n + w_n)$.

PROOF: The proof is similar to previous ones.

(i): Let $\epsilon > 0$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x_m| < \frac{\epsilon}{2}$ for all $n, m > K_1$ and $K_2 \in \mathbb{Q}$ such that $|y_n - y_m| < \frac{\epsilon}{2}$ for all $n, m > K_2$. Set $K = \max(K_1, K_2)$. If $n, m > K$, then

$$|(x_n + y_n) - (x_m + y_m)| = |(x_n - x_m) + (y_n - y_m)| \leq |x_n - x_m| + |y_n - y_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$

(ii): Let $\epsilon > 0$. There is a $K_3 \in \mathbb{Q}$ such that $|x_n - z_n| < \frac{\epsilon}{2}$ for all $n > K_3$ and $K_4 \in \mathbb{Q}$ such that $|y_n - w_n| < \frac{\epsilon}{2}$ for all $n > K_4$. Set $K' = \max(K_3, K_4)$. If $n > K'$, then

$$|(x_n + y_n) - (z_n + w_n)| = |(x_n - z_n) + (y_n - w_n)| \leq |x_n - z_n| + |y_n - w_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$

👍

Dealing with products is a little harder. We first need a result of general interest. A sequence (x_n) in $\mathbb{Q}$ is *bounded* if there is an $M \in \mathbb{Q}$ with $|x_n| \leq M$ for all $n \in \mathbb{N}$.

III.5.2.19. PROPOSITION. Every Cauchy sequence in $\mathbb{Q}$ is bounded.

PROOF: Let (x_n) be a Cauchy sequence. Fix $K \in \mathbb{Q}$ such that $|x_n - x_m| < 1$ for all $n, m > K$, and fix $n > K$. Then, if $m > n$, we have $|x_m| \leq |x_n| + |x_m - x_n| < |x_n| + 1$. The set $\{|x_k| : 1 \leq k < n\}$ is finite, hence has a maximum M'. Set $M = \max(M', |x_n| + 1)$. Then $|x_k| \leq M$ for all $k \in \mathbb{N}$. 👍

III.5.2.20. PROPOSITION. Let (x_n), (y_n), (z_n), (w_n) be sequences in $\mathbb{Q}$.

(i) If (x_n) and (y_n) are Cauchy sequences, so is $(x_n y_n)$.

(ii) If $(x_n) \sim (z_n)$ and $(y_n) \sim (w_n)$, then $(x_n y_n) \sim (z_n w_n)$.

PROOF: By III.5.2.19., there are $M, M', M'' \in \mathbb{Q}$ such that $|x_n| \leq M$, $|y_n| \leq M'$, $|z_n| \leq M''$ for all n. By increasing M, M', and M'' if necessary, we may assume $M, M', M'' > 0$.

(i): Let $\epsilon > 0$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x_m| < \frac{\epsilon}{2M'}$ for all $n, m > K_1$ and $K_2 \in \mathbb{Q}$ such that $|y_n - y_m| < \frac{\epsilon}{2M}$ for all $n, m > K_2$. Set $K = \max(K_1, K_2)$. If $n, m > K$, then

$$|x_n y_n - x_m y_m| = |x_n y_n - x_m y_n + x_m y_n - x_m y_m| \leq |x_n y_n - x_m y_n| + |x_m y_n - x_m y_m|$$

$$= |y_n||x_n - x_m| + |x_m||y_n - y_m| < M' \frac{\epsilon}{2M'} + M \frac{\epsilon}{2M} = \epsilon \ .$$

(ii): Let $\epsilon > 0$. There is a $K_3 \in \mathbb{Q}$ such that $|x_n - z_n| < \frac{\epsilon}{2M'}$ for all $n > K_3$ and $K_4 \in \mathbb{Q}$ such that $|y_n - w_n| < \frac{\epsilon}{2M''}$ for all $n > K_4$. Set $K' = \max(K_3, K_4)$. If $n > K'$, then

$$|x_n y_n - z_n w_n| = |x_n y_n - z_n y_n + z_n y_n - z_n w_n| \leq |x_n y_n - z_n y_n| + |z_n y_n - z_n w_n|$$

$$= |y_n||x_n - z_n| + |z_n||y_n - w_n| < M' \frac{\epsilon}{2M'} + M'' \frac{\epsilon}{2M''} = \epsilon \ .$$

👍

As a result, we may make the following definition:

III.5.2.21. DEFINITION. Define binary operations $+$ and $\cdot$ on $\mathbf{R}$ by $[(x_n)] + [(y_n)] = [(x_n + y_n)]$ and $[(x_n)] \cdot [(y_n)] = [(x_n y_n)]$.

III.5.2.22. Associativity of addition on $\mathbf{R}$ is immediate, since if (x_n), (y_n), and (z_n) are Cauchy sequences in $\mathbb{Q}$, the sequences $([x_n + y_n] + z_n)$ and $(x_n + [y_n + z_n])$ are not only equivalent but even identical. A similar observation shows that associativity of multiplication, commutativity of addition and multiplication, and the distributive law hold in $\mathbf{R}$.

We also have that $[(0)]$ is an additive identity and $[(1)]$ is a multiplicative identity. If (x_n) is a Cauchy sequence, then so is $(-x_n)$, and $[(-x_n)]$ is an additive inverse for $[(x_n)]$.

It is obvious that $\mathbf{Q}$ is closed under $+$ and $\cdot$, and the restrictions are the usual operations from $\mathbb{Q}$, i.e. we have $[(r)] + [(s)] = [(r + s)]$ and $[(r)] \cdot [(s)] = [(rs)]$ for $r, s \in \mathbb{Q}$.

There are also multiplicative inverses for nonzero elements of $\mathbf{R}$. First note the next immediate corollary of III.5.2.12.:

III.5.2.23. COROLLARY. Let (x_n) be a Cauchy sequence in $\mathbb{Q}$. If $[(x_n)] \neq [(0)]$, then there are $\alpha \in \mathbb{Q}_+$ and $K \in \mathbb{Q}$ with $|x_n| \geq \alpha$ for all $n > K$.

III.5.2.24. PROPOSITION. Every nonzero element of $\mathbf{R}$ has a multiplicative inverse.

PROOF: Let (x_n) be a Cauchy sequence in $\mathbb{Q}$ with $[(x_n)] \neq [(0)]$, and fix $\alpha \in \mathbb{Q}_+$ and $K \in \mathbb{Q}$ such that $|x_n| \geq \alpha$ for all $n \in \mathbb{K}$. Set

$$y_n = \begin{cases} x_n^{-1} & \text{if} \quad x_n \neq 0 \\ 1 & \text{if} \quad x_n = 0 \end{cases} .$$

We show (y_n) is a Cauchy sequence. Let $\epsilon \in \mathbb{Q}_+$. There is a $K_1 \in \mathbb{Q}$ such that $|x_n - x_m| < \alpha^2 \epsilon$ for all $n > K_1$. If $K' = \max(K, K_1)$, and $n, m > K'$, then $|x_n|, |x_m| \geq \alpha$, so $x_n, x_m \neq 0$ and $y_n = \frac{1}{x_n}$, $y_m = \frac{1}{x_m}$, and

$$|y_n - y_m| = \left| \frac{1}{x_n} - \frac{1}{x_m} \right| = \frac{|x_m - x_n|}{|x_m x_n|} \leq \frac{|x_m - x_n|}{\alpha^2} < \epsilon .$$

So $\mathbf{R}$ is a field.

III.5.2.25. We now check that $\mathbf{R}$ is an ordered field. It is easiest to check (Pos). Let $\mathbf{P} = \{[(x_n)] \in \mathbf{R} : [(x_n)] > [(0)]\}$. We have (P3) and (P4), i.e. that $\mathbf{R}$ is the disjoint union of $\mathbf{P}$, $\{[(0)]\}$, and $-\mathbf{P}$, so we need only check that $\mathbf{P}$ is closed under addition and multiplication. Let $[(x_n)], [(y_n)] \in \mathbf{P}$. then there are $\alpha, \beta \in \mathbb{Q}_+$ and $K_1, K_2 \in \mathbb{Q}$ such that $x_n \geq \alpha$ for all $n > K_1$ and $y_n \geq \beta$ for all $n > K_2$. Set $K = \max(K_1, K_2)$. If $n > K$, then $x_n + y_n \geq \alpha + \beta \in \mathbb{Q}_+$ and $x_n y_n \geq \alpha\beta \in \mathbb{Q}_+$. So $[(x_n + y_n)]$ and $[x_n y_n)]$ are in $\mathbf{P}$.

Summarizing, we have

III.5.2.26. THEOREM. $\mathbf{R}$ is an ordered field under $+$, $\cdot$, and $<$, with additive identity $[(0)]$ and multiplicative identity $[(1)]$. $\mathbf{Q}$ is an order-dense subfield order-isomorphic to $\mathbb{Q}$.

The Completeness Axiom

Showing that **R** satisfies the Completeness Axiom takes some work. Note that the order-theoretic completeness of the Completeness Axiom is not the same thing as metric completeness, which asserts that every Cauchy sequence converges. For metric completeness of $\mathbb{R}$, see ().

III.5.2.27. THEOREM. $(\mathbf{R}, <)$ satisfies the Completeness Axiom, i.e. every nonempty subset of **R** which is bounded above has a supremum.

PROOF: Let S be a nonempty bounded subset of **R**. We will construct two equivalent Cauchy sequences in $\mathbb{Q}$ whose equivalence class is $\sup(S)$ by a bisection process. First fix $a, b \in \mathbb{Q}$ such that b (actually $[(b)]$) is an upper bound for S, and a is not an upper bound. (For $r \in \mathbb{Q}$ we will write r for $[(r)]$ to streamline notation, except when there is possibility of confusion.)

Set $c_1 = (a+b)/2$. Either c_1 is an upper bound for S, or it is not. If c_1 is an upper bound for S, set $y_1 = a$ and $z_1 = c_1$; otherwise set $y_1 = c_1$ and $z_1 = b$. Then y_1 is not an upper bound for S, z_1 is, and $z_1 - y_1 = 2^{-1}(b-a)$. Repeat the process by setting $c_2 = (y_1 + z_1)/2$ and setting $y_2 = y_1$, $z_2 = c_2$ if c_2 is an upper bound for S and $y_2 = c_2$, $z_2 = z_1$ otherwise. Iterating the process, we obtain sequences (y_n), (z_n) in $\mathbb{Q}$ such that $y_n \le y_{n+1} < z_{n+1} \le z_n$ and $z_n - y_n = 2^{-n}(b-a)$.

To show that (y_n) and (z_n) are Cauchy sequences, and $(y_n) \sim (z_n)$, let $\epsilon > 0$, and fix $K \in \mathbb{N}$ such that $2^{-K}(b-a) \le \epsilon$. Then, if $n, m > K$, $n < m$, we have $y_K \le y_n \le y_m < z_m \le z_n \le z_K$, with at least one of the outside inequalities strict. Thus $|y_n - y_m|$, $|z_n - z_m|$, $|y_n - z_n|$ are all less than $z_K - y_K = 2^{-K}(b-a) \le \epsilon$.

Thus $[(y_n)] = [(z_n)] \in \mathbf{R}$. Call this element $\mathbf{y}$. We will show that $\mathbf{y} = \sup(S)$. First suppose $[(x_n)] \in S$. If $[(x_n)] > \mathbf{y} = [(z_n)]$, then by III.5.2.12. there are $r, s \in \mathbb{Q}$ with $r < s$, and $K_1 \in \mathbb{Q}$, such that $z_n < r$ and $x_n > s$ for all $n > K_1$. Fix $m > K_1$ and set $u = z_m$. Since the z_n are decreasing, we have $z_n \le u$ for all $n > m$, and thus $\mathbf{y} \le [(u)]$. We also have $x_n > s$ for all $n > m$ and $u < s$, so $[(u)] < [(x_n)]$. But by construction u (actually $[(u)]$) is an upper bound for S, a contradiction. Thus $[(x_n)] \le \mathbf{y}$, and $\mathbf{y}$ is an upper bound for S.

Now suppose $[(w_n)] \in \mathbb{R}$ with $[(w_n)] < \mathbf{y} = [(y_n)]$. Then there are $r, s \in \mathbb{Q}$, $r < s$, and $K_2 \in \mathbb{Q}$, such that $w_n < r$ and $y_n > s$ for all $n > K_2$. Fix $m > K_2$, and set $v = y_m$. By construction, v (actually $[(v)]$) is not an upper bound for S, so there is an $[(x_n)] \in S$ with $[(v)] < [(x_n)]$, so there is a $K_3 \in \mathbb{Q}$ with $x_n > v$ for all $n > K_3$. Since $v > s$, we have $w_n < r$ and $s < x_n$ for all $n > \max(K_2, K_3)$, so $[(w_n)] < [(x_n)]$, and $[(w_n)]$ is not an upper bound for S. Thus $\mathbf{y}$ is the least upper bound of S.

The bottom line is:

III.5.2.28. THEOREM. **R**, with its addition, multiplication, and order, is a model for $\mathbb{R}$.

III.5.3. Exercises

III.5.3.1. Let $\mathbb{R}(X)$ be the field of rational "functions" of Exercise III.4.3.5.(c). If $0 \ne f(X) = \frac{p(X)}{q(X)} \in \mathbb{R}(X)$, we say $f(X)$ is *positive* if the leading coefficients (coefficients of the highest powers of X) in $p(X)$ and $q(X)$ have the same sign. Let P be the set of positive elements.

(a) Show that the property of $f(X)$ being positive is independent of the way $f(X)$ is represented as a fraction $\frac{p(X)}{q(X)}$. If f is regarded as a function from $\mathbb{R} \setminus \{x : q(x) = 0\}$ to $\mathbb{R}$, show that f is positive if and only if

$f(x) > 0$ for all sufficiently large $x \in \mathbb{R}$. Show that a sum, product, or quotient of positive elements is positive.

(b) Show that $\mathbb{R}(X)$ is the disjoint union of P, $\{0\}$, and $-P$; thus $(\mathbb{R}(X), P)$ is an ordered field. (This ordering is called the *Fréchet ordering* on $\mathbb{R}(X)$.)

(c) If $\mathbb{R}$ is regarded as the subfield of $\mathbb{R}(X)$ of constant functions, the order on $\mathbb{R}(X)$ induces the usual order on $\mathbb{R}$.

(d) The element $f(X) = X$ is a positive element which is *infinitely large*, i.e. $f > r$ for every $r \in \mathbb{R}$. The element $g(X) = \frac{1}{X}$ is a positive element which is *infinitesimal*, i.e. $0 < g < \epsilon$ for all $\epsilon \in \mathbb{R}$, $\epsilon > 0$. Thus $\mathbb{R}(X)$ is a nonarchimedean ordered field.

III.5.3.2. Formal Power Series. Let F be a field, e.g. $F = \mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ (in fact the construction works over any ring).

(a) A *formal power series* over F is an expression of the form $\sum_{k=0}^{\infty} a_k X^k$, where the a_k are in F and X is an indeterminate symbol. Define addition and multiplication of power series by

$$\left(\sum_{k=0}^{\infty} a_k X^k\right) + \left(\sum_{k=0}^{\infty} b_k X^k\right) = \sum_{k=0}^{\infty} (a_k + b_k) X^k$$

$$\left(\sum_{k=0}^{\infty} a_k X^k\right) \cdot \left(\sum_{k=0}^{\infty} b_k X^k\right) = \sum_{k=0}^{\infty} c_k X^k \ , \text{ where } c_k = \sum_{j=0}^{k} a_j b_{k-j} \ .$$

Let $F[[X]]$ be the set of all formal power series over F. Show that $F[[X]]$ is a commutative ring under these operations, and that the polynomial ring $F[X]$ is a subring. In particular, F can be identified with the set of "constant" power series with $a_k = 0$ for $k > 0$.

(b) Show that a power series $\sum_{k=0}^{\infty} a_k X^k$ has a multiplicative inverse in $F[[X]]$ if and only if $a_0 \neq 0$. [Generate the coefficients of the multiplicative inverse series $\sum_{k=0}^{\infty} b_k X^k$ by successively solving a sequence of linear equations in F.]

The multiplication formula for formal power series can be motivated by writing the series out in informal notation

$$\sum_{k=0}^{\infty} a_k X^k = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + \cdots$$

and similarly for $\sum_{k=0}^{\infty} b_k X^k$, and then formally multiplying out

$$(a_0 + a_1 X + a_2 X^2 + a_3 X^3 + \cdots)(b_0 + b_1 X + b_2 X^2 + b_3 X^3 + \cdots)$$

and collecting together like powers of X.

The elements of $F[[X]]$ are called "formal power series" since there is no notion of convergence implied (in fact, over a general field, the idea of convergence of a power series does not even make sense). So formal power series are just algebraic expressions and should not be thought of as corresponding to "functions". For example, the series $\sum_{k=0}^{\infty} k! X^k$ is a perfectly good formal power series over $\mathbb{Q}$ (or $\mathbb{R}$, $\mathbb{C}$), even though it diverges when any nonzero complex number is substituted for X, and hence cannot be thought of as giving a function on $\mathbb{C}$ ($\mathbb{R}$, $\mathbb{Q}$). Later (cf. XII.9.9.17.), we will consider rings of convergent power series, where ideas from analysis come into the picture in an essential way.

III.5.3.3. Formal Laurent Series. This problem is somewhat of a continuation of Problem III.5.3.2.. Let F be a field. A *formal Laurent series* over F is a sum of the form $\sum_{k=m}^{\infty} a_k X^k$ with the a_k in F, where $m \in \mathbb{Z}$ depends on the series, and may be positive, negative, or zero. We will often implicitly assume when writing such an expression that $a_m \neq 0$ (unless all $a_k = 0$), although this is not necessary. We will informally write such a series as a formal sum of monomials

$$\sum_{k=m}^{\infty} a_k X^k = a_m X^m + a_{m+1} X^{m+1} + \cdots$$

where only finitely many negative powers of X (perhaps none) occur.

(a) Define addition and multiplication of formal Laurent series as in Problem III.5.3.2.. [The formula for multiplication is more complicated, but can be found by considering the informal term-by-term product

$$(a_m X^m + a_{m+1} X^{m+1} + \dots)(b_n X^n + b_{n+1} X^{n+1} + \cdots) \text{ .]}$$

Show that the set $F((X))$ of formal Laurent series is a commutative ring with this addition and multiplication, and the ring of formal power series $F[[X]]$ is a subring.

(b) Show that every nonzero element of $F((X))$ has a multiplicative inverse. [Use the method of problem III.5.3.2.(b). If the series is $\sum_{k=m}^{\infty} a_k X^k$ with $a_m \neq 0$, generate a series $\sum_{k=-m}^{\infty} b_k X^k$.] Thus $F((X))$ is a field.

(c) Show that $F((X))$ is naturally isomorphic to the quotient field () of $F[[X]]$. In particular, there is a canonical embedding of the field $F(X)$ of rational functions into $F((X))$, although the formula for the embedding is not obvious: the multiplicative inverse of a polynomial is usually not a polynomial or even a Laurent polynomial (Laurent series with only finitely many nonzero terms).

(d) Suppose F is an ordered field. Declare a nonzero power series $\sum_{k=m}^{\infty} a_k X^k$ with $a_m \neq 0$ to be *positive* if $a_m > 0$. Let P be the set of all positive series. Show that P satisfies (Pos) and hence makes $F((X))$ into an ordered field. Show that $F((X))$ is nonarchimedean: we have $X > 0$, but $nX < 1$ for all $n \in \mathbb{N}$ (identified with a subset of F as in ()), so X is an *infinitesimal* positive element of $F((X))$.

(e) The ordering on $F(X)$ induced by its embedding in $F((X))$ from (c) is not the same as the ordering from Problem III.5.3.1.. But show it can be obtained from this ordering by the substitution $X \mapsto \frac{1}{X}$ in $F(X)$.

III.5.3.4. Puiseux Series. This and the next two problems are a continuation of Problem III.5.3.3.; we will define three successively larger fields containing the field $\mathbb{R}((X))$. We will work over $\mathbb{R}$, although the construction can be done over any field.

A *Puiseux series* is a formal expression of the form $\sum_{k=m}^{\infty} a_{k/n} X^{k/n}$ for some $m \in \mathbb{Z}$, $n \in \mathbb{N}$, where $a_{k/n} \in \mathbb{R}$ for all k. In other words, a Puiseux series is a formal Laurent series in $X^{1/n}$ for some n (which may vary from series to series). Any formal Laurent series over $\mathbb{R}$ is a Puiseux series with $n = 1$. We identify $X^{k/n}$ and $X^{kr/nr}$ for any $r \in \mathbb{N}$; thus any two (or any finite number of) Puiseux series can be written as formal Laurent series in $X^{1/n}$ for a common n.

(a) Define addition and multiplication of Puiseux series in analogy with the definition for formal Laurent series. Show that the set $\mathcal{P}$ of Puiseux series is a field containing $\mathbb{R}((X))$ as a subfield. In particular, $\mathcal{P}$ contains $\mathbb{R}$ as the subfield of "constant" series, where $a_{k/n} = 0$ for all $k \neq 0$.

(b) Define an ordering on $\mathcal{P}$ by defining the positive cone as in Exercise III.5.3.3.(d): If $x = \sum_{k=m}^{\infty} a_{k/n} X^{k/n}$ is a nonzero element of $\mathcal{P}$, with m/n the smallest k/n for which $a_{k/n} \neq 0$ (m/n is called the *degree* of the

series), set $x > 0$ if $a_{m/n} > 0$. This ordering agrees with the previous one from Exercise III.5.3.3.(d) on $\mathbb{R}((X))$.

(c) Show by direct computation that every nonnegative element of $\mathcal{P}$ has a square root. [If $x \in \mathcal{P}$ with degree m/n with $m \geq 0$, write $x = a_{m/n}X^{m/n}(1+y)$ where $a_{m/n} > 0$ and y has degree > 0, and use the binomial series for $(1+y)^{1/2}$.]

In fact, $\mathcal{P}$ is a real-closed field, so is a first-order model for $\mathbb{R}$ (and is the "smallest" such model containing an infinitesimal). Puiseux series were first formally studied by PUISEUX in about 1850, although many of the basic ideas go back to NEWTON. See [BPR06, Chapter 2] for more information.

III.5.3.5. The Levi-Civita Numbers. This is a continuation of Problem III.5.3.4.. We will work over $\mathbb{R}$, although the construction can be done over any field. A *Levi-Civita number* is a formal expression of the form $\sum_{q\in\mathbb{Q}} a_q X^q$, where $a_q \in \mathbb{R}$ for all q, and for any such sum the set $S = \{q : a_q \neq 0\}$ is a *left-finite* set, i.e. for any $r \in \mathbb{Q}$ the set $\{q \in S : q < r\}$ is finite. In other words, either S is finite or the elements of S can be arranged in a strictly increasing sequence converging to $+\infty$. (The set S may vary from series to series.) Any Puiseux series over $\mathbb{R}$ is a Levi-Civita number.

(a) Define addition and multiplication of Levi-Civita numbers in analogy with the definition for Puiseux series. Show that the set $\mathcal{R}$ of Levi-Civita numbers is a field containing $\mathcal{P}$ as a subfield.

(b) Define an ordering on $\mathcal{R}$ by defining the positive cone as in Exercise III.5.3.3.(d): If $x = \sum_{q\in\mathbb{Q}} a_q X^q$ is a nonzero element of $\mathcal{R}$, with q_0 the smallest q for which $a_q \neq 0$ (q_0 is called the *degree* of the series), set $x > 0$ if $a_{q_0} > 0$. This ordering agrees with the previous one from Exercise III.5.3.4. on $\mathcal{P}$.

(c) Show by direct computation that every nonnegative element of $\mathcal{R}$ has a square root. [The argument is similar to the one in III.5.3.4.(c).]

(d) Show that $\mathcal{R}$ is complete in the following sense:

In fact, $\mathcal{R}$ is a real-closed field, so is a first-order model for $\mathbb{R}$ (and is the "smallest" such model which is complete in the sense of (d) and contains an infinitesimal). The Levi-Civita numbers are important in computational analysis. See [**?**], [SB10], [Rib92], or [DW96, Theorem 2.15] for more information.

III.5.3.6. Hahn Series. This is a further continuation of Problem III.5.3.5.. We will work over $\mathbb{R}$, although the construction can be done over any field. Let G be a totally ordered abelian group, e.g. $\mathbb{Q}$ or $\mathbb{R}$. A *Hahn series* over G is a formal expression of the form $\sum_{q\in G} a_q X^q$, where $a_q \in \mathbb{R}$ for all q, and for any such sum the set $S = \{q : a_q \neq 0\}$ is a well-ordered set, i.e. every subset of S has a smallest element (equivalently, S contains no strictly decreasing sequences). The set S may vary from series to series. Every Levi-Civita number is a Hahn series over $\mathbb{Q}$ (but there are Hahn series over $\mathbb{Q}$ which are not Levi-Civita numbers); a formal Laurent series is a Hahn series over $\mathbb{Z}$.

(a) Define addition and multiplication of Hahn series in analogy with the definition for Levi-Civita numbers. Show that the set $\mathcal{S}_G$ of Hahn series over G is a field.

(b) Define an ordering on $\mathcal{S}_G$ by defining the positive cone as in Exercise III.5.3.3.(d): If $x = \sum_{q\in\mathbb{Q}} a_q X^q$ is a nonzero element of $\mathcal{S}_G$, with q_0 the smallest q for which $a_q \neq 0$ (q_0 is called the *degree* of the series), set $x > 0$ if $a_{q_0} > 0$. This ordering agrees with the previous one from Exercise III.5.3.3.(d) on $\mathbb{R}((X))$.

(c) Show by direct computation that if G is 2-divisible, every nonnegative element of $\mathcal{S}_G$ has a square root. [The argument is similar to the one in III.5.3.4.(c).]

(d) Show that $\mathcal{S}$ is complete in the sense of III.5.3.5.(d).

In fact, $\mathcal{S}_G$ is also a real-closed field if G is divisible. See [DW96] for more information.

III.5.3.7. This problem is a continuation of III.5.3.2.. A formal power series $\sum_{k=0}^{\infty} a_k X^k$ has *order* m if $a_m \neq 0$ and $a_k = 0$ for $k < m$.

We will suggestively use letters $f, g, h, \ldots$ for elements of $F[[X]]$, or even more suggestively $f(X)$, etc., even though formal power series are not functions in any literal sense.

(a) Let (f_n) be a sequence of formal power series over F. Show that if for every m there are only finitely many f_n of order $\leq m$, then the infinite sum $\sum_{n=1}^{\infty} f_n$ is well defined in $F[[X]]$. [Add the coefficients of each power of X.]

(b) Give another proof of III.5.3.2.(b) by using the formal infinite series

$$(1-x)^{-1} = \sum_{n=0}^{\infty} x^n \ .$$

(c) If $f = \sum_{k=0}^{\infty} a_k X^k$ and $g = \sum_{k=0}^{\infty} b_k X^k$ are formal power series over F, and g has order ≥ 1 (i.e. $b_0 = 0$), then the infinite sum

$$\sum_{k=0}^{\infty} a_k g^k = \sum_{k=0}^{\infty} a_k [g(X)]^k$$

is a well-defined element of $F[[X]]$ which can be denoted $f(g(X))$ or $f \circ g$. [The order of g^n is at least n.] There is not a well-defined composition if $b_0 \neq 0$.

This can be motivated by regarding formal power series as defining "function germs" around $0 \in F$. It makes sense that composition can only be defined if the inside function sends 0 to 0.

(d) Prove that composition, when defined, is associative: if g and h have order ≥ 1, then $g \circ h$ has order ≥ 1, and for every f we have $(f \circ g) \circ h = f \circ (g \circ h)$.

(e) Prove the following "Formal Inverse Function Theorem" (where ι is the "identity function" X, i.e. $\iota = \sum_{k=0}^{\infty} a_k X^k$, where $a_1 = 1$ and $a_k = 0$ for $k \neq 1$) by successively solving a sequence of linear equations for the coefficients of g:

THEOREM. Let $f = \sum_{k=0}^{\infty} a_k X^k \in F[[X]]$. Then there exists $g \in F[[X]]$ with $f \circ g = \iota$ if and only if $a_0 = 0$ and $a_1 \neq 0$. If such g exists, it is unique, and also $g \circ f = \iota$.

If $f = \sum_{k=0}^{\infty} a_k X^k \in F[[X]]$, we may write $f(0) = a_0$ and $f'(0) = a_1$ (there is a difficulty carrying this over to higher derivatives unless F has characteristic 0 since a factorial should occur in the formula). Then, if the g exists in the theorem, we have $g(0) = 0$ and $g'(0) = [f'(0)]^{-1}$. In fact, there is a derivation δ (XII.12.1.1.) on $F[[X]]$ defined by

$$\delta\left(\sum_{k=0}^{\infty} a_k X^k\right) = \sum_{k=1}^{\infty} k a_k X^{k-1}$$

for which $f'(0) = [\delta(f)](0)$. We can write $f'(X) = \delta(f(X))$. Compare with V.3.10.9..

III.5.3.8. This problem is a continuation of III.5.3.2. and III.5.3.7.. A formal power series in two variables X and Y over a field F is a formal sum

$$\sum_{k,j=0}^{\infty} a_{k,j} X^k Y^j$$

where the $a_{k,j}$ are elements of F. There is an obvious way of adding and multiplying two such series, making the set $F[[X,Y]]$ of such formal power series into a ring.

(a) Show that there are natural isomorphisms

$$F[[X,Y]] \cong (F[[X]])[[Y]] \cong (F[[Y]])[[X]] \ .$$

In particular, $F[[X]]$ is a subring of $F[[X,Y]]$ consisting of all sums for which $a_{k,j} = 0$ for $j > 0$. Similarly, $F[[Y]]$ is a subring of $F[[X,Y]]$. Write f or $f(X,Y)$, etc., for a general element of $F[[X,Y]]$; elements of the subrings $F[[X]]$ and $F[[Y]]$ can be written $f(X)$, $g(Y)$, etc.

If $f(X,Y) = \sum_{k,j=0}^{\infty} a_{k,j} X^k Y^j$ is an element of $F[[X,Y]]$, we may write $f(0,0) = a_{0,0}$, $f_X(0,0) = \frac{\partial f}{\partial X}(0,0) = a_{1,0}$, and $f_Y(0,0) = \frac{\partial f}{\partial Y}(0,0) = a_{0,1}$. (As in III.5.3.7.(e), these come from derivations δ_X and δ_Y on $F[[X,Y]]$.)

(b) Describe the invertible elements of $F[[X,Y]]$.

(c) If $f \in F[[X]]$ and $g \in F[[X,Y]]$, with $g(0,0) = 0$, show that there is a well-defined element $f(g(X,Y)) \in F[[X,Y]]$. Similarly, if $f, g, h \in F[[X,Y]]$, with $g(0,0) = h(0,0) = 0$, show that there is a well-defined element $f(g(X,Y), h(X,Y)) \in F[[X,Y]]$. If $g, h \in F[[X]]$, then $f(g(X), h(X)) \in F[[X]]$ also.

(d) If $f(X)$ is a formal power series in one variable, define a formal power series $g(X,Y)$ in two variables by $g(X,Y) = f(X+Y)$ as in (c). Regard $g \in (F[[X]])[[Y]]$ to obtain

$$g(X,Y) = \sum_{k=0}^{\infty} f_k(X) Y^k$$

for power series $f_0(X), f_1(X), \ldots$. Show that $f_0 = f$ and $f_1 = f'$, and more generally (if F has characteristic 0) $f_k = \frac{1}{k!} f^{(k)}$, where the "derivative" is taken k times. This is an algebraic version of Taylor's Theorem (V.10.1.7.).

(e) Let $f(X) \in F[[X]]$. Define $g \in F[[X,Y]]$ by $g(X,Y) = f(Y) - f(X)$. Show that g is divisible by $Y - X$ in $F[[X,Y]]$ and if

$$h(X,Y) = \frac{g(X,Y)}{Y-X} = \frac{f(Y) - f(X)}{Y - X}$$

then $f'(X) = h(X,X)$.

(f) Prove the following "Formal Implicit Function Theorem" (cf. VIII.5.3.10.):

THEOREM. Let $f(X,Y) \in F[[X,Y]]$ with $f(0,0) = 0$ and $f_Y(0,0) \neq 0$. Then there is a unique $g \in F[[X]]$ such that $f(X, g(X)) = 0$ (where 0 denotes the zero element of $F[[X,Y]]$). We have $g(0) = 0$ and $g'(0) = -\frac{f_X(0,0)}{f_Y(0,0)}$.

(g) Discuss Laurent series in two variables and the quotient field of $F[[X,Y]]$.

(h) Generalize to formal power series rings $F[[X_1, \ldots, X_n]]$. Formulate and prove higher-dimensional versions of the Formal Implicit Function Theorem and Formal Inverse Function Theorem analogous to VIII.5.1.7., VIII.5.1.34., and VIII.5.2.1..

(i) To what extent can Puiseux series, Levi-Civita Numbers, and Hahn Series be generalized to several variables?

References: [**?**, 4.4.7], [Jz75], [Sok11], [KP02a].

III.5.3.9. The p-adic Numbers. The real numbers are not the only field formed by completing $\mathbb{Q}$. This problem describes the completion of $\mathbb{Q}$ with respect to a different metric. There is actually one of these for each prime number p. These fields are important in number theory.

Fix a prime number p. If $x \in \mathbb{Q}$, $x \neq 0$, then x can be written $p^n \cdot \frac{a}{b}$, where $n \in \mathbb{Z}$ and $a, b \in \mathbb{Z}$ are not divisible by p. The number n is uniquely determined by x; write $\nu_p(x) = n$ and $|x|_p = p^{-n} = p^{-\nu_p(x)}$. Set $\nu_p(0) = +\infty$ and $|0|_p = 0$.

(a) Show that if $x, y \in \mathbb{Q}$, $\nu_p(x+y) \geq \min(\nu_p(x), \nu_p(y))$ and $\nu_p(xy) = \nu_p(x) + \nu_p(y)$.

(b) Show that if $x, y \in \mathbb{Q}$, then $|x+y|_p \leq \max(|x|_p, |y|_p) \leq |x|_p + |y|_p$ and $|xy|_p = |x|_p|y|_p$.

(c) Show that d_p defined by $d_p(x,y) = |x-y|_p$ is a metric on $\mathbb{Q}$.

(d) Let $\mathbb{Q}_p$ be the set of equivalence classes of Cauchy sequences of rational numbers with respect to d_p, where equivalence is defined as in (). Mimic the proof of () to define addition and multiplication on $\mathbb{Q}_p$, and show that $\mathbb{Q}_p$ is a field. Extend $|\cdot|_p$ and d_p to $\mathbb{Q}_p$, and show that $\mathbb{Q}_p$ is complete with respect to d_p. (Note that the topology defined by d_p is quite different from the usual topology on $\mathbb{Q}$: $p^n \to 0$ in the d_p-topology, for example.)

(e) Show that if $x \in \mathbb{Z}$, then $|x|_p \leq 1$, and the closure of $\mathbb{Z}$ in $\mathbb{Q}_p$ is

$$\mathbb{Z}_p = \{x \in \mathbb{Q}_p : |x|_p \leq 1\} .$$

This is a subring of $\mathbb{Q}_p$, called the *ring of p-adic integers*. (The notation $\mathbb{Z}_p$ is inconsistent with (); this meaning of the notation $\mathbb{Z}_p$ will be used only in this problem.) Note that $\mathbb{Z}_p \cap \mathbb{Q}$ is larger than $\mathbb{Z}$; it includes all rational numbers $\frac{a}{b}$ where p does not divide b.

(f) Show that $\mathbb{Q}_p$ is locally compact with its metric topology, and $\mathbb{Z}_p$ is a compact open subset.

(g) Show that $\mathbb{Q}_p$ is totally disconnected.

(h) Show that the ring $\mathbb{Z}_p$ has only one maximal ideal J, the set of multiples of p, and $\mathbb{Z}_p/J \cong \mathbb{Z}/p\mathbb{Z}$ (). Every nonzero ideal in $\mathbb{Z}_p$ is a principal ideal generated by p^n (the set of all multiples of p^n) for some $n \in \mathbb{N}$.

(i) Show that the quotient field () of $\mathbb{Z}_p$ is $\mathbb{Q}_p$.

III.6. The Complex Numbers

Although they are not as familiar as the real numbers, the complex numbers $\mathbb{C}$ are the natural culmination of the expansion of the number system. Although we will make only very limited use of them in this book, they are also extremely important in analysis, and many parts of analysis, e.g. the theory of analytic functions () and Fourier series (), are naturally done on $\mathbb{C}$ instead of $\mathbb{R}$. Complex numbers and analysis done on them (complex analysis) are among the most beautiful parts of mathematics, and the most useful in applications.

From a purely algebraic standpoint, the real numbers have almost all the properties one might want in a number system. The only obvious shortcoming is that negative numbers do not have square roots in $\mathbb{R}$. The idea of making the complex numbers is to expand $\mathbb{R}$ to a larger system of numbers in which square roots of negative real numbers (and, as it turns out, square roots of arbitrary numbers in the expanded system) can be taken. The resulting system actually yields a lot more than just square roots.

III.6.1. Axiomatization of $\mathbb{C}$

III.6.1.1. Axioms for $\mathbb{C}$ are not standardized, but here is a set which can be used:

(ℂ1) $\mathbb{C}$ has two binary operations making it into a field.

(ℂ2) $\mathbb{C}$ contains $\mathbb{R}$ as a subfield (more precisely, $\mathbb{C}$ contains a specified subfield $\mathcal{R}$ with a specified isomorphism between $\mathbb{R}$ and $\mathcal{R}$).

(ℂ3) $\mathbb{C}$ contains an element i with $i^2 = -1$.

(ℂ4) Every element of $\mathbb{C}$ can be written as $a + bi$ for $a, b \in \mathbb{R}$ (or $\mathcal{R}$).

If $\mathcal{C}$ is any model for $\mathbb{C}$, then since $\mathcal{C}$ is a field, all the standard algebraic operations and identities hold in $\mathcal{C}$, and we are justified in using standard algebraic notation and manipulation in $\mathcal{C}$ without comment.

III.6.1.2. Proposition. Let $\mathcal{C}$ be a model for $\mathbb{C}$. If $z \in \mathcal{C}$, then z has a unique representation $z = x + yi$ where x and y are in the specified subfield $\mathcal{R}$ which is a model for $\mathbb{R}$.

Proof: There is such a representation for z by (ℂ4). To show uniqueness, if $z = x_1 + y_1 i = x_2 + y_2 i$, then $0 = (x_1 - x_2) + (y_1 - y_2)i$, so it suffices to show that if $0 = a + bi$ with $a, b \in \mathcal{R}$, then $a = b = 0$. But

$$0 = 0 \cdot (a - bi) = (a + bi)(a - bi) = a^2 - b^2 i^2 = a^2 + b^2$$

and thus $a = b = 0$. 👍

We can then quickly show that any two models for $\mathbb{C}$ are essentially identical (isomorphic):

III.6.1.3. Theorem. Let $(\mathcal{C}, \mathcal{R}, i)$ and $(\mathcal{C}', \mathcal{R}', i')$ be models for $\mathbb{C}$, and let $\psi : \mathcal{R} \to \mathcal{R}'$ be the isomorphism induced by the specified isomorphisms of $\mathcal{R}$ and $\mathcal{R}'$ with $\mathbb{R}$. Define $\phi : \mathcal{C} \to \mathcal{C}'$ by

$$\phi(a + bi) = \psi(a) + \psi(b)i'$$

(this is well defined by III.6.1.2.). Then ϕ is a bijection, and for all $z, w \in \mathcal{C}$,

$$\phi(z + w) = \phi(z) + \phi(w), \quad \phi(zw) = \phi(z)\phi(w) \ .$$

Thus the two models are isomorphic.

Proof: It is clear that ϕ is a bijection from III.6.1.2. and the fact that ψ is a bijection. We need check only the algebraic formulas. If $z = x + yi$ and $w = u + vi$, we have

$$zw = (x + yi)(u + vi) = xu + xvi + yui + yvi^2 = (xu - yv) + (xv + yu)i$$

$$\phi(z) + \phi(w) = [\psi(x) + \psi(y)i'] + [\psi(u) + \psi(v)i']$$
$$= [\psi(x) + \psi(u)] + [\psi(y) + \psi(v)]i' = \psi(x + u) + \psi(y + v)i' = \phi([x + u] + [y + v]i) = \phi(z + w)$$

$$\phi(z)\phi(w) = [\psi(x) + \psi(y)i'][\psi(u) + \psi(v)i']$$
$$= \psi(x)\psi(u) + [\psi(x)\psi(v) + \psi(y)\psi(u)]i' + \psi(y)\psi(v)i^2 = [\psi(x)\psi(u) - \psi(y)\psi(v)] + [\psi(x)\psi(v) + \psi(y)\psi(u)]i'$$
$$= \psi(xu - yv) + \psi(xv + yu)i' = \phi([xu - yv] + [xv + yu]i) = \phi(zw) \ .$$

👍

III.6.1.4. Thus, once we have shown that there is a model for $\mathbb{C}$, as usual we can talk about *the* complex numbers $\mathbb{C}$, and regard $\mathbb{R}$ as a subset of $\mathbb{C}$.

III.6.1.5. If $(\mathcal{C}, \mathcal{R}, i)$ is a model for $\mathbb{C}$, and $z = a + bi$ for $a, b \in \mathcal{R}$, the numbers a and b are uniquely determined by z and are called the *real* and *imaginary* parts of z, denoted $a = Re(z)$, $b = Im(z)$ (the notation $a = \Re(z)$ and $b = \Im(z)$ is also used). The element $a - bi$ of $\mathcal{C}$ is called the *complex conjugate* of z, denoted $\bar{z}$, and the real number $\sqrt{a^2 + b^2}$ is called the *absolute value* (or *modulus*, *magnitude*, *length*, or *norm*) of z, denoted $|z|$. We have $|i| = 1$, and $|\cdot|$ on $\mathcal{R}$ coincides with the usual absolute value on $\mathbb{R}$ under the specified identification.

The proofs in the next proposition are simple algebraic calculations and are left to the reader.

III.6.1.6. PROPOSITION. Let $(\mathcal{C}, \mathcal{R}, i)$ be a model for $\mathbb{C}$, and let $z = x + yi$ and $w = u + vi$ be elements of $\mathbb{C}$. Then

(i) $\overline{z + w} = \bar{z} + \bar{w}$ and $\overline{zw} = \bar{z}\bar{w}$.

(ii) $\bar{\bar{z}} = z$.

(iii) $z\bar{z} = |z|^2$.

(iv) $|zw| = |z||w|$. In particular, $|-z| = |z|$ and $|zi| = |z|$.

(v) $|Re(z)| \leq |z|$ and $|Im(z)| \leq |z|$.

It is also true that $|z + w| \leq |z| + |w|$, but we defer the proof to the next section.

III.6.1.7. It is impossible to make $\mathbb{C}$ (any model of $\mathbb{C}$) into an ordered field. For if P is a positive cone in $\mathbb{C}$ satisfying (Pos), then either i or $-i$ is in P; in either case $i^2 = (-i)^2 = -1$ would be in P, a contradiction.

III.6.2. The Standard Model for $\mathbb{C}$: the Complex Plane

There is a universally-used standard model for $\mathbb{C}$, the *complex plane*. In fact, this model for $\mathbb{C}$ is so standard that $\mathbb{C}$ is generally defined to be, or at least thought of as, this model. There is no real harm in this, and even considerable advantage, since the geometric properties of the complex plane are of essential use in complex analysis. But see III.6.2.24. and I.7.1.11. for some cautionary comments.

III.6.2.1. By considering real and imaginary parts, complex numbers can be naturally identified with ordered pairs of real numbers. So in our standard model, the elements will *be* ordered pairs of real numbers. Addition of complex numbers just corresponds to addition of the real parts and of the imaginary parts, i.e. coordinatewise addition, which is the usual addition of ordered pairs of numbers. The first calculation in the proof of III.6.1.3. gives a strong clue to how multiplication of ordered pairs should be defined. Thus we are led to the following definition:

III.6.2.2. DEFINITION. Let $\mathcal{C}$ be the set $\mathbb{R}^2$ of ordered pairs of real numbers. Define addition and multiplication in $\mathcal{C}$ by

$$(x,y)+(u,v)=(x+u,y+v)$$
$$(x,y)(u,v)=(xu-yv,xv+yu)\ .$$

III.6.2.3. PROPOSITION. The associative, commutative, and distributive laws hold for addition and multiplication in $\mathcal{C}$.

PROOF: The commutative laws obviously hold. The associative law for addition is a simple calculation:

$$[(x,y)+(u,v)]+(s,t)=([x+u]+s,[y+v]+t)=(x+[u+s],y+[v+t])=(x,y)+[(u,v)+(s,t)]\ .$$

The associative law for multiplication and the distributive law are straightforward but tedious calculations which should probably just be left to the interested, skeptical, or masochistic reader, but we give them for completeness:

$$[(x,y)(u,v)](s,t)=(xu-yv,xv+yu)(s,t)=(xus-yvs-[xvt+yut],xut-yvt+xvs+yus)$$
$$=(xus-xvt-[yvs+yut],xut+xvs+yus-yvt)=(x,y)(us-vt,ut+vs)=(x,y)[(u,v)(s,t)]\ .$$

$$(x,y)[(u,v)+(s,t)]=(x,y)(u+s,v+t)=(xu+xs-[yv+yt],xv+xt+yu+ys)$$
$$=([xu-yv]+[xs-yt],[xv+yu]+[xt+ys])=(xu-yv,xv+yu)+(xs-yt,xt+ys)=(x,y)(u,v)+(x,y)(s,t)\ .$$



III.6.2.4. The pair $(0,0)$ is an additive identity, and $(-x,-y)$ is an additive inverse for (x,y). We also have $(1,0)(x,y)=(x,y)$ for any (x,y), so $(1,0)$ is a multiplicative identity. There are also multiplicative inverses:

III.6.2.5. PROPOSITION. If $(x,y)\neq(0,0)$, then $x^2+y^2>0$ and $\left(\frac{x}{x^2+y^2},-\frac{y}{x^2+y^2}\right)$ is a multiplicative inverse for (x,y).

PROOF: The first statement follows from (). We have

$$(x,y)\left(\frac{x}{x^2+y^2},-\frac{y}{x^2+y^2}\right)=\left(x\frac{x}{x^2+y^2}-y\left[-\frac{y}{x^2+y^2}\right],x\left[-\frac{y}{x^2+y^2}\right]+y\frac{x}{x^2+y^2}\right)$$
$$=\left(\frac{x^2}{x^2+y^2}+\frac{y^2}{x^2+y^2},-\frac{xy}{x^2+y^2}+\frac{xy}{x^2+y^2}\right)=(1,0)\ .$$



We summarize these results:

III.6.2.6. THEOREM. $\mathcal{C}$ is a field with respect to the operations $+$ and $\cdot$ of III.6.2.2., with zero element $(0,0)$ and multiplicative identity $(1,0)$.

III.6.2.7. PROPOSITION. Let $a,b,u,v \in \mathbb{R}$. Then in $\mathcal{C}$, $(a,0)(u,v) = (au,av)$. In particular, $(a,0)(b,0) = (ab,0)$. So, since $(a,0)+(b,0) = (a+b,0)$, the map $\psi : \mathbb{R} \to \mathcal{C}$ with $\psi(x) = (x,0)$ is an isomorphism from $\mathbb{R}$ onto the subfield $\mathcal{R} = \{(x,0) : x \in \mathbb{R}\}$ of $\mathcal{C}$.

The proof is a simple calculation left to the reader.

Finally, we need i. The proof of the next proposition is also a simple calculation.

III.6.2.8. PROPOSITION. In $\mathcal{C}$, $(0,1)^2 = (-1,0)$. If $(x,y) \in \mathcal{C}$, then $(x,y) = (x,0) + (y,0)(0,1)$.

III.6.2.9. Thus, if we call $(0,1)$ i, we have that every complex number (x,y) can be written $a+bi$ for $a,b \in \mathcal{R}$, namely $a = (x,0)$, $b = (y,0)$. Under the identification of $\mathbb{R}$ with $\mathcal{R}$ via ψ, we have $i^2 = -1$, $x = Re((x,y))$, $y = Im((x,y))$, $\overline{(x,y)} = (x,-y)$. We usually write $x+yi$ instead of (x,y) for elements of $\mathcal{C}$.

The definition of the complex conjugate, along with III.6.1.6., motivates the calculation in III.6.2.5.: we have $z^{-1} = \frac{\bar{z}}{z\bar{z}} = \frac{\bar{z}}{|z|^2}$ if $z \neq 0$.

Summarizing again, we get our final result:

III.6.2.10. THEOREM. $(\mathcal{C}, \mathcal{R}, (0,1))$ is a model for $\mathbb{C}$. Every model for $\mathbb{C}$ is isomorphic to this one.

Geometry of $\mathbb{C}$

We thus normally represent $x+yi \in \mathbb{C}$ geometrically as $(x,y) \in \mathcal{C} = \mathbb{R}^2$. The real number x is represented as $(x,0)$, and i is represented as $(0,1)$. The x-axis in $\mathbb{R}^2$ is $\mathcal{R}$, called the *real axis*; the y-axis is the set of real multiples of i, called the *imaginary axis*. See Figure ().

III.6.2.11. We also have that $|x+yi| = \sqrt{x^2+y^2}$. This is the same as $\|(x,y)\|$, where $\|\cdot\|$ is the standard norm in $\mathbb{R}^2$ (). Thus $|x+yi|$ can be interpreted as the Euclidean distance between $x+yi = (x,y)$ and $0 = (0,0)$. The next result then follows immediately from ():

III.6.2.12. PROPOSITION. The *triangle inequality* holds in $\mathbb{C}$: if $z,w \in \mathbb{C}$, then $|z+w| \leq |z| + |w|$. In particular, $|z| \leq |Re(z)| + |Im(z)|$.

III.6.2.13. There are nice geometric interpretations of addition and multiplication in $\mathbb{C}$ (regarded as $\mathcal{C}$). Since addition in $\mathcal{C}$ coincides with vector addition in $\mathbb{R}^2$, it is given by the parallelogram law (); see Figure ().

Multiplication also has a simple interpretation using polar coordinates. We first describe the polar coordinates of an element of $\mathcal{C}$. Some of this discussion must be regarded as informal until the trigonometric functions are properly defined (cf. (); note that no complex numbers are used in the definition of the trigonometric functions or development of their properties).

III.6.2.14. Definition. Let $z = x + yi \in \mathbb{C}$, $z \neq 0$. If (r, θ) are the polar coordinates () of the point $(x, y) \in \mathbb{R}^2$, then $r = \sqrt{x^2 + y^2} = |x + yi|$ is called the *modulus*, *magnitude*, *length*, or *norm* of z, and θ is called the *argument* of z, denoted $Arg(z)$. We have $x = r\cos\theta$ and $y = r\sin\theta$; if $x \neq 0$, we have $\tan\theta = \frac{y}{x}$. We write $e^{i\theta}$ for the complex number $\cos\theta + i\sin\theta$.

Conversely, if (r, θ) is a pair of real numbers with $r \neq 0$, then there is a unique $z \in \mathcal{C}$ with $r = |z|$ and $\theta = Arg(z)$. We do not define $Arg(0)$. The argument of a positive real number is 0, and the argument of a negative real number is π.

Note that the argument of a complex number is only determined up to addition of a multiple of 2π. This ambiguity leads to considerable difficulties, but also a certain richness, in complex analysis; cf. III.6.2.21.. In particular, the argument cannot be defined continuously for all $z \in \mathbb{C} \setminus \{0\}$, although it can be continuously defined locally (except around 0). Note that even though θ is not uniquely determined by z, the complex number $e^{i\theta}$ is uniquely determined.

We will consider $e^{i\theta}$ to simply be a convenient notational abbreviation for $\cos\theta + i\sin\theta$, but it is actually a special case of the complex exponential function (XII.1.1.8.(iii)). This notation retains at least one important property of the usual exponential function:

III.6.2.15. Proposition. If θ and ϕ are real numbers, then $e^{i(\theta+\phi)} = e^{i\theta}e^{i\phi}$.

Proof: This is just an application of basic trigonometric identities. We have

$$e^{i(\theta+\phi)} = \cos(\theta+\phi) + i\sin(\theta+\phi) = [\cos\theta\cos\phi - \sin\theta\sin\phi] + i[\sin\theta\cos\phi + \cos\theta\sin\phi]$$
$$= (\cos\theta + i\sin\theta)(\cos\phi + i\sin\phi) = e^{i\theta}e^{i\phi} \ .$$



The next observation is simple, but fundamental:

III.6.2.16. Proposition. If $z \in \mathcal{C}$, $z \neq 0$, and $r = |z|$, $\theta = Arg(z)$, then $z = re^{i\theta}$. So $e^{i\theta} = \frac{z}{|z|}$.

As a corollary of the last two propositions, we get a geometric interpretation of multiplication:

III.6.2.17. Corollary. Let z and w be nonzero complex numbers. Then $|zw| = |z||w|$ and $Arg(zw) = Arg(z) + Arg(w)$ (where the last formula is interpreted modulo 2π). Alternatively, if $z = re^{i\theta}$ and $w = se^{i\phi}$ with $r, s, \theta, \phi \in \mathbb{R}$ and $r, s > 0$, then $zw = rse^{i(\theta+\phi)}$.

III.6.2.18. Multiplication by a fixed complex number z can thus be interpreted geometrically as scaling by a factor of $|z|$ and rotating counterclockwise around the origin by an angle $Arg(z)$. See Figure (). This interpretation is in line with the geometric interpretation of multiplication of vectors by real numbers (scalars): multiplication by a positive real number just scales length, and multiplication by a negative real number scales length and reverses direction (rotates by an angle π).

III.6.2.19. The polar coordinate representation provides an easy way to show that complex numbers have arbitrary roots. Suppose $z \in \mathbb{C}$ and $n \in \mathbb{N}$; we want to show there is an $x \in \mathbb{C}$ with $x^n = z$. If $z = 0$, this is obvious (and the only solution is $x = 0$). Suppose $z \neq 0$. Set $s = |z|^{1/n}$ and $\phi = Arg(z)/n$. Then $w = se^{i\phi}$ satisfies $w^n = z$. Any multiple of $\frac{2\pi}{n}$ can be added to ϕ with the same result; thus there are exactly n distinct solutions in $\mathbb{C}$ to $x^n = z$. Summarizing, we obtain:

III.6.2.20. Proposition. Let $z \in \mathbb{C}$ and $n \in \mathbb{N}$. Then the equation $x^n = z$ has a solution in $\mathbb{C}$. If $z = 0$, the unique solution is $x = 0$. If $z \neq 0$, there are exactly n solutions, distributed uniformly on the circle of radius $|z|^{1/n}$ around 0 at angles $\frac{Arg(z)}{n} + \frac{2k\pi}{n}$, $0 \leq k \leq n-1$.

III.6.2.21. In particular, every nonzero complex number z has exactly two square roots, which are negatives of each other on the line through the origin of angle $\frac{Arg(z)}{2}$ with the positive real axis. However, as z varies there is no way to choose one of the two square roots of z *continuously* on the whole complex plane, although a continuous choice can be made locally (except around 0). Thus there is no continuous square root function from all of $\mathbb{C}$ to $\mathbb{C}$.

There is a considerable generalization of III.6.2.20. which is one of the most important theorems of algebra, as its name suggests.

III.6.2.22. Theorem. [Fundamental Theorem of Algebra] Any nonconstant polynomial with complex coefficients has a root in $\mathbb{C}$. ($\mathbb{C}$ is *algebraically closed*.)

There are several quite different proofs of this theorem; see XII.1.4.2. for a nice one. III.6.2.20. is the special case of the polynomial $x^n - z$ for any $z \in \mathbb{C}$.

III.6.2.23. As a result, there is no algebraic reason to try to expand the number system beyond $\mathbb{C}$, so $\mathbb{C}$ is a natural culmination of the progression from $\mathbb{N}$ through $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$.

III.6.2.24. As mentioned earlier, although $\mathbb{C}$ is normally, and profitably, thought of as *being* the complex plane $\mathcal{C}$, it must be kept in the back of the mind that $\mathcal{C}$ is only a model for $\mathbb{C}$. For example, we want to regard $\mathbb{R}$ as a subfield of $\mathbb{C}$. But $\mathbb{R}$ is not a subset of $\mathcal{C}$: the subfield $\mathcal{R}$ of ordered pairs of real numbers with second coordinate 0 is not the same thing as $\mathbb{R}$, although there is an obvious and natural one-one correspondence between them. And there are other models of $\mathbb{C}$ which are useful in some contexts; see III.6.3.1., III.6.3.2.(e), III.6.3.3. for some examples.

III.6.3. Exercises

III.6.3.1. Consider the following subset of $M_2(\mathbb{R})$:

$$\mathfrak{C} = \left\{ \begin{bmatrix} x & -y \\ y & x \end{bmatrix} : x, y \in \mathbb{R} \right\} .$$

(a) Show that $\mathfrak{C}$ is a subring of $M_2(\mathbb{R})$.

(b) Show that $\mathfrak{C}$ is unital and commutative.

(c) Show that every element z of $\mathfrak{C}$ can be written uniquely as $z = x1 + yi$ for $x, y \in \mathbb{R}$, where

$$1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} .$$

Define $\bar{z} = x1 - yi$. Show that $z\bar{z} = (x^2 + y^2)1 = (det\ z)1$. If we define $|z| = \sqrt{x^2 + y^2}$, then $|zw| = |z||w|$ for all $z, w \in \mathfrak{C}$.

(d) Show that every nonzero element of $\mathfrak{C}$ has a multiplicative inverse. Thus $\mathfrak{C}$ is a field.

(e) Show that $\mathfrak{R} = \{x1 : x \in \mathbb{R}\}$ is a subfield of $\mathfrak{C}$ isomorphic to $\mathbb{R}$.

(f) Show that $(\mathfrak{C}, \mathfrak{R}, i)$ is a model for $\mathbb{C}$.

This model for $\mathbb{C}$ arises naturally from linear algebra. If $z = a + bi$ is a fixed element of $\mathcal{C}$, then multiplication by z is a real linear transformation M_z of $\mathcal{C} = \mathbb{R}^2$, and the matrix of this linear transformation with respect to the standard basis $\{1, i\}$ of $\mathcal{C} = \mathbb{R}^2$ is $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$.

III.6.3.2. Consider the following subset of $M_2(\mathbb{C})$:

$$\mathbb{H} = \left\{ \begin{bmatrix} z & -\bar{w} \\ w & \bar{z} \end{bmatrix} : z, w \in \mathbb{C} \right\} .$$

(a) Show that $\mathbb{H}$ is a subring of $M_2(\mathbb{C})$. (By considering $\mathbb{C}$ as a subring of $M_2(\mathbb{R})$ as in (), $\mathbb{H}$ can be rewritten as the subring

$$\left\{ \begin{bmatrix} a & -c & -b & -d \\ c & a & d & -b \\ b & -d & a & c \\ d & b & -c & a \end{bmatrix} : a, b, c, d \in \mathbb{R} \right\}$$

of $M_4(\mathbb{R})$.)

(b) Show that $\mathbb{H}$ is unital and noncommutative. [Consider the following elements:

$$1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad i = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad j = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \quad k = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} .$$

Compute ij, ji, jk, kj, ki, ik.]

(c) Show that every element x of $\mathbb{H}$ can be written uniquely as $x = a1 + bi + cj + dk$ for $a, b, c, d \in \mathbb{R}$. Define $\bar{x} = a1 - bi - cj - dk$. Show that $x\bar{x} = (a^2 + b^2 + c^2 + d^2)1$. If we define $|x| = \sqrt{a^2 + b^2 + c^2 + d^2}$, then $|xy| = |x||y|$ for all $x, y \in \mathbb{H}$.

(d) Show that every nonzero element of $\mathbb{H}$ has a multiplicative inverse. Thus $\mathbb{H}$ satisfies all the axioms for a field except (M2). Such a ring is called a *division ring* or *skew field*.

(e) Show that each of the sets $\mathbb{R}+\mathbb{R}i$, $\mathbb{R}+\mathbb{R}j$, and $\mathbb{R}+\mathbb{R}k$ is a subring of $\mathbb{H}$ isomorphic to $\mathbb{C}$.

(f) If $x = a1 + bi + cj + dk \in \mathbb{H}$, a is called the *real part* or *scalar part* of x, and $bi + cj + dk$ the *imaginary part* of x. If $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ are the standard basis vectors in $\mathbb{R}^3$, then $bi + cj + dk$ can be identified with the vector $\mathbf{v} = b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \in \mathbb{R}^3$, called the *vector part* of x. Thus an element of $\mathbb{H}$ may be regarded as a pair $(a, \mathbf{v})$, where $a \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^3$. Show that the addition is coordinatewise, and the multiplication is given by

$$(a, \mathbf{v})(b, \mathbf{w}) = (ab - \mathbf{v} \cdot \mathbf{w}, a\mathbf{w} + b\mathbf{v} + \mathbf{v} \times \mathbf{w})$$

where $\cdot$ and $\times$ are the dot and cross product of vectors in $\mathbb{R}^3$. (Although $\times$ is not associative, this multiplication *is* associative!)

The ring $\mathbb{H}$ is called the ring of *quaternions*. The name refers to the fact that it is a four-dimensional vector space (algebra) over $\mathbb{R}$. The letter $\mathbb{H}$ is used in honor of the Irish mathematician William Rowan Hamilton, who first defined the quaternions in 1843 and studied their properties. Hamilton became convinced that quaternions were *the* fundamental structure which could be used to describe and explain most of mathematics and physics. While quaternions have beautiful and useful algebraic and geometric structure, they have not quite lived up to Hamilton's expectations, although recently they have become an increasingly important computational tool in such areas as computer graphics and control theory. Quaternions are also of some importance in quantum mechanics; the matrices i, j, k are called the *Pauli spin matrices*. See () for details.

There is a way to make $\mathbb{R}^8$ into a somewhat similar algebraic structure called the *Cayley numbers* in which every nonzero element has a multiplicative inverse. Multiplication in the Cayley numbers is, however, not associative. It turns out that the only values of n for which $\mathbb{R}^n$ can be made into such a structure are $n = 1, 2, 4$, and 8, and the only structures possible are $\mathbb{R}$, $\mathbb{C}$, $\mathbb{H}$, and the Cayley numbers. See () for details.

III.6.3.3. Let $\mathbb{R}[X]$ denote the set of all polynomials with real coefficients, i.e.

$$\mathbb{R}[X] = \{a_0 + a_1X + a_2X^2 + \cdots + a_nX^n : n \in \mathbb{N}_0, a_0, \ldots, a_n \in \mathbb{R}\} .$$

(a) Show that $\mathbb{R}[X]$ is a unital commutative ring with usual addition and multiplication of polynomials.

(b) Let J be the set of all polynomials in $\mathbb{R}[X]$ which are divisible by $X^2 + 1$. Show that J is an ideal (Exercise III.3.7.4.) in $\mathbb{R}[X]$.

(c) Show that the quotient ring $\mathbb{R}[X]/J$ is a field which is a model for $\mathbb{C}$, with the classes of the constant functions playing the role of $\mathbb{R}$, and the class of X playing the role of i.

Levi-Civita numbers

III.7. Fundamental Theorems About $\mathbb{R}$

III.7.1. Partitions

In this section, we will discuss partitions and tagged partitions of intervals in $\mathbb{R}$. The primary use of partitions is in Riemann-type integration, but several important basic theorems in analysis can also be proved using tagged partitions.

DEFINITION. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$. A *partition* of $[a,b]$ is a division of $[a,b]$ into a finite number of subintervals $\{[x_{i-1},x_i] : 1 \le i \le n\}$ with only endpoints in common. A partition $\mathcal{P}$ is usually described by its set of endpoints $\{x_0,x_1,\dots,x_n\}$, where $a = x_0 < x_1 < \cdots < x_n = b$.

The *norm* or *mesh* of a partition $\mathcal{P} = \{x_0,x_1,\dots,x_n\}$ is

$$\|\mathcal{P}\| = \max\{x_i - x_{i-1} : 1 \le i \le n\}.$$

A *tag* in an interval $[c,d]$ is a specified number in $[c,d]$. A *tagged partition* is a partition $\{x_0,x_1,\dots,x_n\}$ with a specified tag t_i in $[x_{i-1},x_i]$ for each i. A tagged partition is usually denoted by $\dot{\mathcal{P}} = \{x_0,x_1,\dots,x_n;t_1,\dots,t_n\}$.

The norm or mesh of a partition is a measure of how fine the partition is. There is a more sophisticated and flexible notion of fineness of a partition based on gauges:

DEFINITION. A *gauge* on an interval $[a,b]$ is a function δ from $[a,b]$ to $(0,\infty)$.

If δ is a gauge on $[a,b]$, a tagged partition $\dot{\mathcal{P}} = \{x_0,x_1,\dots,x_n;t_1,\dots,t_n\}$ of $[a,b]$ is *δ-fine* if $x_i - x_{i-1} < \delta(t_i)$ for all $1 \le i \le n$.

The word "tagged" is frequently omitted in referring to δ-fine partitions, although it must be kept in mind that a δ-fine partition is by definition tagged.

Note that a gauge is just a strictly positive real-valued function on $[a,b]$; no continuity or other restrictions on δ are assumed. Allowing discontinuous gauges, and ones not bounded below by a positive number, give the theory of gauges and fine partitions its power.

The smaller a gauge δ is, the harder it is for a partition to be δ-fine. The most common gauges are constant functions, e.g. $\delta(x) = \epsilon$ for all x. For such a δ, a tagged partition $\dot{\mathcal{P}}$ is δ-fine if and only if $\|\mathcal{P}\| < \epsilon$.

If δ is a constant gauge, then it follows easily from the Archimedean property of $\mathbb{R}$ that there is a δ-fine partition of any interval $[a,b]$. However, it is not at all obvious that a nonconstant gauge δ on $[a,b]$ which is not bounded below by a positive number has a δ-fine partition. This turns out to be true, and is the main result of this section; this theorem is actually equivalent to the Completeness Property and easily gives other important equivalent properties of $\mathbb{R}$.

THEOREM.[FINE PARTITION THEOREM] Let δ be any gauge on an interval $[a,b]$ in $\mathbb{R}$. Then there is a δ-fine partition of $[a,b]$.

PROOF: Let $S = \{x \in [a,b] :$ there is a δ-fine partition of $[a,x]\}$. If $a < x < a + \delta(a)$, then $\dot{\mathcal{P}} = \{a,x;a\}$ is a δ-fine partition of $[a,x]$ (with one subinterval). So S is nonempty. Let $c = \sup S$; then $c \ge a + \delta(a) > a$. We will show that $c = b$. First note that there is a δ-fine partition of $[a,c]$: if $x \in S$ with $c - \delta(c) < x < c$, and $\{a,x_1,\dots,x_{n-1},x;t_1,\dots,t_n\}$ is a δ-fine partition of $[a,x]$, then $\{a,x_1,\dots,x_{n-1},x,c;t_1,\dots,t_n,c\}$ is a δ-fine partition of $[a,c]$. Now suppose that $c < b$, and let $y \in [a,b]$, $c < y < c + \delta(c)$; then, using the previous notation, $\{a,x_1,\dots,x_{n-1},x,c,y;t_1,\dots,t_n,c,c\}$ is a δ-fine partition of $[a,y]$, contradicting the definition of c. Thus $c = b$ and the theorem is proved. 👍

III.7.2. Compactness and Continuity

From the Fine Partition Theorem, we derive several fundamental results about the real numbers. The standard proofs of all these results resemble the proof of the Fine Partition Theorem; our derivation of the results from the Fine Partition Theorem follows R. A. Gordon [**?**].

A collection $\mathcal{U} = \{U_i : i \in \Omega\}$ of open sets in $\mathbb{R}$ (or in a general topological space) is an *open cover* of a set S if $S \subseteq \cup_i U_i$. A *subcover* of $\mathcal{U}$ for S is a subset of $\mathcal{U}$ which is also a cover of S. A subset F of $\mathbb{R}$ is *compact* if every open cover of F has a finite subcover.

THEOREM. A closed bounded interval $[a, b]$ in $\mathbb{R}$ is compact.

PROOF: Let $\mathcal{U} = \{U_i : i \in \Omega\}$ be an open cover of $[a, b]$. For each $x \in [a, b]$, choose an $i_x \in \Omega$ with $x \in U_{i_x}$, and choose $\delta(x) > 0$ such that $(x - \delta(x), x + \delta(x)) \subseteq U_{i_x}$. Then $\delta : [a, b] \to (0, \infty)$ is a gauge on $[a, b]$. Let $\{x_0, x_1, \ldots, x_n; t_1, \ldots, t_n\}$ be a δ-fine partition of $[a, b]$. Then $[x_{k-1}, x_k] \subseteq (t_k - \delta(t_k), t_k + \delta(t_k)) \subseteq U_{i_{t_k}}$ for each k, so $\{U_{i_{t_1}}, \ldots, U_{i_{t_n}}\}$ is a finite subcover of $\mathcal{U}$. 👍

COROLLARY. Every closed bounded subset of $\mathbb{R}$ is compact.

PROOF: If F is closed and contained in $[a, b]$, and $\mathcal{U}$ is an open cover of F, then $\mathcal{U} \cup \{F^c\}$ is an open cover of $[a, b]$, and this open cover has a finite subcover $\mathcal{V}$ for $[a, b]$. $\mathcal{V} \setminus \{F^c\} \subseteq \mathcal{U}$ is then a finite subcover of $\mathcal{U}$ for F. 👍

Virtually the same proof shows that any closed subset of a compact set is compact. The converse is also true:

PROPOSITION. A compact subset of $\mathbb{R}$ is closed and bounded.

PROOF: Let E be a compact subset of $\mathbb{R}$. If $U_m = (-m, m)$, then $\{U_m : m \in \mathbb{N}\}$ is an open cover of E, so there is a finite subcover $\{U_{m_1}, \ldots, U_{m_n}\}$. If $M = \max\{m_1, \ldots, m_n\}$, then $E \subseteq (-M, M)$. To see that E is closed, let $c \in E^c$, and for each $\epsilon > 0$ let $U_\epsilon = \{x \in \mathbb{R} : |x - c| > \epsilon\}$. Then $\{U_\epsilon : \epsilon > 0\}$ is an open cover of E, so there is a finite subcover $\{U_{\epsilon_1}, \ldots, U_{\epsilon_n}\}$. If $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_n\}$, then $(c - \epsilon, c + \epsilon) \subseteq E^c$, so E^c is open. 👍

Combining these two results, we obtain:

COROLLARY.[HEINE-BOREL THEOREM] A subset of $\mathbb{R}$ is compact if and only if it is closed and bounded.

The Heine-Borel Theorem gives a generalization of the Nested Interval Theorem:

THEOREM. Let (F_m) be a decreasing sequence of nonempty closed bounded subsets of $\mathbb{R}$. Then $\cap_m F_m$ is not empty.

PROOF: Suppose $F_1 \subseteq [a, b]$. If $\cap F_m = \emptyset$, then $\{F_m^c : m \in \mathbb{N}\}$ is an open cover of $[a, b]$, which has a finite subcover $\{F_{m_1}, \ldots, F_{m_n}\}$. If $M = \max\{m_1, \ldots, m_n\}$, then $F_M = F_{m_1} \cap \cdots \cap F_{m_n} = \emptyset$, a contradiction. 👍

From this, we get the important Bolzano-Weierstrass Theorem:

THEOREM.[BOLZANO-WEIERSTRASS THEOREM] Every bounded sequence in $\mathbb{R}$ has a convergent subsequence.

PROOF: Let (x_n) be a bounded sequence in $[a,b] \subseteq \mathbb{R}$. It suffices to show that (x_n) has a cluster point (). For each $m \in \mathbb{N}$, let F_m be the closure of $\{x_n : n \geq m\}$. Then F_m is a nonempty closed subset of $\mathbb{R}$, and $F_{m+1} \subseteq F_m$ and $F_m \subseteq [a,b]$ for all m. Thus $\cap_m F_m$ is nonempty, and any $c \in \cap_m F_m$ is a cluster point of (x_n). 👍

We then get some basic theorems about continuous functions as consequences:

THEOREM. Let E be a compact subset of $\mathbb{R}$, and let $f : E \to \mathbb{R}$ be continuous. Then $f(E)$ is compact.

PROOF: We first show that $f(E)$ is bounded. Suppose for each n there is an $x_n \in E$ with $|f(x_n)| \geq n$. Since E is bounded, (x_n) has a subsequence (x_{k_n}) converging to some $y \in \mathbb{R}$, and $y \in E$ since E is closed. Because f is continuous, the sequence $(f(x_{k_n}))$ converges to $f(y)$ by the Sequential Criterion. But this is impossible since $(f(x_{k_n}))$ is unbounded by construction, a contradiction.

Now suppose that c is in the closure of $f(E)$, and choose $x_n \in E$ such that $f(x_n) \to c$. Again by the Bolzano-Weierstrass Theorem, (x_n) has a subsequence (x_{k_n}) converging to some $y \in E$. But then $(f(x_{k_n}))$ converges to $f(y)$ by the Sequential Criterion; since $(f(x_{k_n}))$ converges to c, we have $c = f(y) \in f(E)$ and $f(E)$ is closed. 👍

COROLLARY.[MAX-MIN THEOREM] Let E be a closed bounded subset of $\mathbb{R}$, and $f : E \to \mathbb{R}$ a continuous function. Then f is bounded on E and attains its maximum and minimum on E, i.e. there are $c, d \in E$ such that $f(c) \leq f(x) \leq f(d)$ for all $x \in E$.

For the proof, just note that $f(E)$ is compact and that a compact subset of $\mathbb{R}$ has a maximum and a minimum element.

See () for a more complete discussion of compactness.

The next theorem is a variation of the Tietze Extension Theorem of topology (). It is certainly not true that a continuous function on a subset E of $\mathbb{R}$ can be extended to a continuous function on all of $\mathbb{R}$; there are many counterexamples, such as $f(x) = \frac{1}{x}$ on $(0,\infty)$. But if E is closed, there is always an extension:

THEOREM.[TIETZE EXTENSION THEOREM FOR $\mathbb{R}$] Let E be a closed subset of $\mathbb{R}$, and $f : E \to \mathbb{R}$ a continuous function. Then f can be extended to a continuous function on all of $\mathbb{R}$, i.e. there is a continuous function $g : \mathbb{R} \to \mathbb{R}$ which agrees with f on E.

PROOF: The definition of g is very simple. Let $U = E^c$; then U is open, so it consists of a union of countably many disjoint open intervals $I_n = (a_n, b_n)$ and possibly an unbounded interval of the form $(-\infty, a)$ (if E has a smallest element a) and/or an unbounded interval (b,∞) (if E has a largest element b.) Define g to be equal to f on E, linear on $[a_n, b_n]$ for each n, and if necessary constant on $(-\infty, a]$ and/or $[b,\infty)$. If $x \in I_n$, the formula for $g(x)$ is

$$g(x) = f(a_n) + \frac{f(b_n) - f(a_n)}{b_n - a_n}(x - a_n)\,.$$

It is geometrically clear that g is continuous everywhere, but takes a little work to write down a careful proof. Suppose $c \in \mathbb{R}$; we will show g is continuous at c. If $c \notin E$, this is obvious since g is linear in a neighborhood of c. Suppose $c \in E$. We will show that $\lim_{x \uparrow c} f(x) = f(c)$. Let $\epsilon > 0$ be given. If $c = b_n$ for some n, let $\delta = \min(b_n - a_n, \frac{b_n - a_n}{|f(b_n) - f(a_n)|}\epsilon)$ (if $f(a_n) = f(b_n)$, set $\delta = b_n - a_n$.) Then if $c - \delta < x < c$, g is linear on $[x, c]$, so $|f(x) - f(c)| < \epsilon$.

Now suppose that c is not one of the b_n. For $\epsilon > 0$, choose $\delta_1 > 0$ such that $c - \delta_1 < x < c$ and $x \in E$ imply $|f(x) - f(c)| < \epsilon$ (using the fact that f is continuous on E.) Since c is not one of the b_n, there is a

$d \in E$ with $c - \delta_1 < d < c$. Set $\delta = c - d$. If $c - \delta < x < c$, we will show that $|f(x) - f(c)| < \epsilon$. This is obvious if $x \in E$ since $\delta < \delta_1$; and if $x \in U$, then there is an n such that $d \le a_n < x < b_n < c$, so $f(x)$ is between $f(a_n)$ and $f(b_n)$, each of which is in $(f(c) - \epsilon, f(c) + \epsilon)$, so $|f(x) - f(c)| < \epsilon$. This proves that g is continuous from the left; the proof that g is continuous from the right is almost identical. 👍

THEOREM. Let $f : [a, b] \to \mathbb{R}$ be bounded, and continuous a.e. Then f is Riemann integrable on $[a, b]$.

PROOF: We suppose that $f : [a, b] \to \mathbb{R}$ is bounded, and continuous except on a set D of measure 0. Then we can find an open set U of arbitrarily small measure containing D. If $|f(x)| \le M$ for all $x \in [a, b]$, and $\epsilon > 0$ is given, we choose U to have measure $< \epsilon/4M$.

Suppose first that U can be chosen to be a *finite* disjoint union of open intervals $I_1, \ldots, I_n$. We will find step functions ϕ, ψ on $[a, b]$ with $\phi \le f \le \psi$ and $\int_{[a,b]} (\psi - \phi)\, d\lambda < \epsilon$. We need the following theorem from advanced calculus:

Theorem. Let F be a closed bounded set in $\mathbb{R}$. Then any continuous function from F to $\mathbb{R}$ is uniformly continuous.

Apply this theorem to $F = [a, b] \setminus U$: there is thus a $\delta > 0$ such that, whenever $x, t \in F$ and $|x - t| < \delta$, $|f(x) - f(t)| < \epsilon/2(b - a)$.

Now choose a partition $\mathcal{P}$ of $[a, b]$, containing the intervals $I_1, \ldots, I_n$, and such that the other intervals in the partition all have length $< \delta$. Let ϕ and ψ be the step functions with values $\inf_{x \in I} f(x)$ and $\sup_{x \in I} f(x)$ on I respectively, for each subinterval I in $\mathcal{P}$. (It doesn't matter how ϕ and ψ are defined on the endpoints of the subintervals so long as $\phi \le f \le \psi$ there.) Then $0 \le \psi - \phi \le 2M$, and $\psi - \phi \le \epsilon/2(b - a)$ on F. So

$$\int_{[a,b]} (\psi - \phi)\, d\lambda = \int_U (\psi - \phi)\, d\lambda + \int_F (\psi - \phi)\, d\lambda \le 2M\lambda(U) + \frac{\epsilon}{2(b-a)}\lambda(F)$$
$$< 2M\frac{\epsilon}{4M} + \frac{\epsilon}{2(b-a)}(b - a) = \epsilon.$$

Now consider the more difficult case where U is an infinite disjoint union of open intervals $I_1, I_2, \ldots$. We may define F and δ as before. We can choose a sequence $(\mathcal{P}_n)$ of partitions of $[a, b]$ such that $\mathcal{P}_n$ contains $I_1, \ldots, I_n$ and such that the rest of the subintervals have length $< \delta$, and define ϕ_n, ψ_n as above. We may with some care arrange for the $\mathcal{P}_n$ to become successively finer as n increases (i.e. each interval in $\mathcal{P}_n$ is a union of intervals in $\mathcal{P}_{n+1}$): it is only necessary to insure that every interval I_k in U is entirely contained in a single subinterval in $\mathcal{P}_1$, which is easily arranged since there are only finitely many of the I_k with length $\ge \delta/2$. If the $\mathcal{P}_n$ are successively finer, we have

$$\phi_1 \le \phi_2 \le \cdots \le f \le \cdots \le \psi_2 \le \psi_1.$$

Let $\phi = \sup \phi_n$, $\psi = \inf \psi_n$. Then the ϕ_n and ψ_n are step functions, but ϕ and ψ are not step functions in general. However, they are Lebesgue integrable. Also, we have $\phi \le f \le \psi$, $0 \le \psi - \phi \le 2M$, and $\psi - \phi \le \epsilon/2(b - a)$ on F, so as before we have $\int_{[a,b]} (\psi - \phi)\, d\lambda < \epsilon$. But by the Monotone Convergence Theorem (or the Dominated Convergence Theorem),

$$\int_{[a,b]} (\psi - \phi)\, d\lambda = \lim_{n \to \infty} \int_{[a,b]} (\psi_n - \phi_n)\, d\lambda$$

so $\int_{[a,b]} (\psi_n - \phi_n)\, d\lambda < \epsilon$ for sufficiently large n. 👍

III.7.3. The Creeping Lemma

An alternate approach to many fundamental theorems about $\mathbb{R}$ was presented in [MR68], based on the following fact which depends crucially on the Completeness Property of $\mathbb{R}$:

III.7.3.1. THEOREM. [CREEPING LEMMA] Let I be an interval in $\mathbb{R}$, and let R be a transitive relation on I with the following property: for every $x \in I$, there is a $\delta_x > 0$ (perhaps depending on x) such that, whenever $y \in I$, $x - \delta_x < y \leq x$, then yRx, and whenever $y \in I$, $x \leq y < x + \delta_x$, then xRy. Then R extends $\leq$ on I, i.e. for every $a, b \in I$, $a \leq b$, we have aRb.

PROOF: R is automatically reflexive by the assumed condition. Fix $a, b \in I$, $a \leq b$. We may suppose $a < b$. Let

$$S = \{x \in [a, b] : aRx\} \ .$$

Then S is a bounded subset of $\mathbb{R}$ which is nonempty since $a \in S$. Set $c = \sup(S)$. Then $c > a$ since $[a, a + \delta_a) \subseteq S$ for some $\delta_a > 0$, so $c \geq a + \delta_a$.

There is a $\delta_c > 0$ such that if $c - \delta_c < y \leq c$, then yRc. Since $c = \sup(S)$, there is an $x \in S$, $c - \delta_c < x \leq c$; then aRx and xRc. Thus aRc by transitivity.

If $c < b$, then if $c < x < \min(c + \delta_c, b)$, we have cRx; since aRc, we have aRx, so $x \in S$, contradicting that $c = \sup(S)$. Thus $c = b$. 👍

The next corollary is called the *Weak Creeping Lemma* in [MR68]. For a slick version of the proof, see ().

III.7.3.2. COROLLARY. Let I be an interval in $\mathbb{R}$, and let R be an equivalence relation on I with the following property: for every $x \in I$, there is a $\delta_x > 0$ (perhaps depending on x) such that, whenever $y \in I$, $|x - y| < \delta_x$, then xRy. Then R is the full relation on I, i.e. for every $a, b \in I$, we have aRb.

Here is a simple sample application. Only the weak version is needed.

III.7.3.3. PROPOSITION. Let f be a continuous function on a closed bounded interval $[a, b]$. Then f is bounded on $[a, b]$.

PROOF: Define an equivalence relation on $I = [a, b]$ by $x \sim y$ if f is bounded on the closed interval between x and y. Fix $x \in [a, b]$. There is a δ_x such that $|f(x) - f(y)| < 1$ if $y \in [a, b]$ and $|x - y| < \delta_x$. Thus $x \sim y$ if $|x - y| < \delta_x$. By III.7.3.2., $a \sim b$. 👍

III.7.3.4. COROLLARY. [EXTREME VALUE THEOREM] Let f be a continuous function on a closed bounded interval $[a, b]$. Then f is bounded on $[a, b]$, and attains a maximum and minimum on $[a, b]$.

PROOF: By III.7.3.3., f is bounded. Let $c = \sup(\{f(x) : x \in [a, b]\})$. If f never takes the value c on $[a, b]$, set $g(x) = \frac{1}{c - f(x)}$; then g is continuous on $[a, b]$, and positive everywhere. By III.7.3.3., g is bounded. If $d = \sup(\{g(x) : x \in [a, b]\})$, then

$$\frac{1}{c - f(x)} \leq d$$

$$\frac{1}{d} \leq c - f(x)$$

$$f(x) \leq c - \frac{1}{d}$$

for all $x \in [a, b]$, contradicting the definition of c. Thus f takes the value c on $[a, b]$. The proof that f attains a minimum is nearly identical. 👍

III.7.4. Exercises

III.7.4.1. Using the proof of III.7.1. as a model, give a direct proof of VI.1.3.2..

III.7.4.2. The Completeness Property of $\mathbb{R}$ implies the Fine Partition Theorem, as the proof of III.7.1. shows. Show that the conclusion of the Fine Partition Theorem implies the Completeness Property as follows.

(a) Suppose S is a nonempty bounded subset of $\mathbb{R}$ with no supremum. Let $a \in S$, and b an upper bound for S.

(b) Define a gauge δ on $[a, b]$ as follows. Let $x \in [a, b]$. If x is an upper bound for S, let y be a smaller upper bound for S, and define $\delta(x) = x - y$. And if x is not an upper bound for S, let $y \in S$ with $x < y$ and set $\delta(x) = y - x$. (To avoid using the Axiom of Choice, if x is an upper bound for S set $\delta(x) = \frac{1}{n}$, where n is the smallest natural number for which $x - \frac{1}{n}$ is an upper bound for S; if x is not an upper bound for S, let $\delta(x) = \frac{1}{n}$, where n is the smallest natural number for which there is a $y \in S$ with $x + \frac{1}{n} \leq y$.)

(c) Show that there is no δ-fine partition of $[a, b]$.

III.7.4.3. Explain why the name "Creeping Lemma" is appropriate. (See [Pét38].)

III.8. Representation of Real Numbers

III.8.1. Decimal Expansions

> "It is India that gave us the ingenious method of expressing all numbers by means of ten symbols, each symbol receiving a value of position as well as an absolute value; a profound and important idea which appears so simple to us now that we ignore its true merit. But its very simplicity and the great ease which it has lent to computations put our arithmetic in the first rank of useful inventions; and we shall appreciate the grandeur of the achievement the more when we remember that it escaped the genius of Archimedes and Apollonius, two of the greatest men produced by antiquity."
>
> *Pierre-Simon Laplace*[6]

The most familiar way to represent real numbers is by finite or infinite decimals. In fact, we are accustomed to thinking of real numbers as *being* decimals. For example, we are used to writing expressions like

$$\frac{1}{4} = 0.25$$

$$\frac{1}{3} = 0.333333333\cdots$$

$$\pi = 3.1415926\cdots$$

even though the precise meaning of the last expression is rather obscure (the first expression, and even the second, has a fairly evident meaning).

It is better from a logical standpoint to regard decimals as merely convenient *representations* of real numbers. Even natural numbers should not really be thought of as *being* decimal expressions, since it is sometimes convenient to represent them in other ways. For example, are

$$2\ , \quad 1.999999\cdots\ , \quad \sqrt{4}\ , \quad \int_0^\pi \sin x\, dx\ , \quad \sum_{k=0}^{\infty} \frac{1}{2^k}$$

all natural numbers? Are they the same number? And one runs into immediate logical difficulties in saying that rational numbers are the same thing as their decimal expansions.

In this section, we will carefully develop the theory of decimal expansions, and expansions to other bases, and show that there is a (near) one-to-one correspondence between real numbers and infinite decimals. The arguments are nice illustrations of the Completeness Axiom in various forms, particularly the Nested Intervals Property.

Bases

III.8.1.1. The first ingredient in decimal representations is to fix a *base* b, which is a natural number greater than 1. The idea is then to write all real numbers in terms of powers of b.

[6] H. Eves, *Return to Mathematical Circles*

III.8.1.2. It is pretty universal in modern human societies to use as base the number we call *ten*. In principle, this base is no better or worse than any other choice of base; it almost certainly came into general use simply because most of us have ten fingers (the basic meaning of the word *digit* is "finger or toe"). Indeed, the word *decimal* technically only correctly applies to base ten expansions, although we will, by abuse of language, use it more generally for expansions with different bases.

Other bases have occasionally been used. Besides ten, at least the numbers five, eight, twelve, twenty, twenty-seven, forty, and sixty (and reportedly also four, six, fifteen, and twenty-four) have been used at various times by different societies (see [**?**, p. 149] for a description of how counting in base eight, twenty-seven, or sixty can be done on the fingers, and has been by some societies). There are vestiges of these systems in the names of certain numbers in some modern-day languages (including the words *score* and *gross* in English, Russian *copok*, Danish *halvtreds*, Irish *fiche*, and French numbers like *quatre-vingt-onze*), and in our units of time and measurement of angles (although dividing a circle into 360 degrees apparently originated at least partly because there are approximately 360 days in a year). The notorious English system of measurement (and, until recently, money), can also be regarded as a number system with different base (actually more than one!)

I remember as a boy being fascinated by a book advocating use of a base twelve system (usually called *duodecimal* or *dozenal*, but J. NYSTROM called it *duodenal*, which was, according to [**?**], "perhaps the most unfortunate double entendre in the history of science.") The main argument was that arithmetic is simpler in this base since twelve has more divisors than ten; the multiplication table, although somewhat larger, is more systematic and easier to remember (and note that our school times tables typically go up to 12 anyway!), and many more rational numbers have terminating decimals. Although these claims are arguably true, the duodecimal movement has never gotten anywhere, though it continues to have its advocates.

With the advent of computers, it has become natural in certain settings to use binary (base 2) or hexadecimal (base 16) expansions since they match up nicely with computer architecture and hardware. Binary "decimal" digits are called *bits*, and groups of binary digits of fixed length (usually 8) are called *bytes*. Terms such as *megabyte* and *gigabyte* are now in common use in the language.

> "There are only 10 kinds of people in the world: those who understand binary, and those who don't."
>
> *Seen on a T-Shirt*

III.8.1.3. In describing decimal expansions, we will generally express everything in the usual base ten. The exposition can easily be adapted to any base b by just replacing 10 by b. We will use the usual base ten digit symbols 0 through 9; for another base b, if $b < 10$ the digits 0 through $b-1$ can be used, and if $b > 10$ additional symbols must be added. For example, in binary only the digits 0 and 1 are used, and in hexadecimal the letters A through F are customarily used to denote the numbers ten through fifteen, which are single digits in hexadecimal.

III.8.1.4. It is theoretically possible to vary the base from place to place (e.g. to use base ten for the first decimal digit, base 7 for the second, etc.); it seems almost inconceivable that this would ever be useful in practice, but the English essentially did this for centuries with money, and we still do it with time, and, in the U.S., other measurements. (The theory of continued fractions gives a much more useful version of varying bases; there the "base" varies not only from place to place, but from number to number.) The interested reader can figure out how to adapt the arguments to this case.

III.8.1.5. There is also a fascinating generalization of decimal expansions where the base b is not required to be a natural number, but can be an arbitrary real or complex number with $|b| > 1$, and the "digits" are a finite set of real or complex numbers containing 0. Such expansions are intimately connected with the theory of fractal sets in the plane. See [Edg08, §1.6] for details. Also see II.8.9.10. for a different kind of generalization.

Decimal Representations of Integers

All versions of decimal representations are based on the following "obvious" fact:

III.8.1.6. PROPOSITION. Let $b \in \mathbb{R}$, $b > 1$. Then $b^m < b^n$ if $m, n \in \mathbb{N}_0$, $m < n$ (setting $b^0 = 1$), and the set $\{b^n : n \in \mathbb{N}\}$ is not bounded above.

PROOF: It is easily proved by induction that $b^n > 1$ for all $n \in \mathbb{N}$, since

$$b^{n+1} = b \cdot b^n > 1 \cdot b^n = b^n$$

for any n. The first statement then follows since if $m < n$, we have

$$b^n = b^{n-m} \cdot b^m > 1 \cdot b^m = b^m \ .$$

For the second statement, suppose $S = \{b^n : n \in \mathbb{N}\}$ is bounded, and let $M = \sup(S)$. Choose $n \in \mathbb{N}$ such that $b^n > M - (b-1)$. Then

$$M \geq b^{n+1} = b^n + (b^{n+1} - b^n) = b^n + b^n(b-1) > [M - (b-1)] + (b-1) = M$$

since $b^n > 1$, a contradiction. 👍

III.8.1.7. We now fix a number $b \in \mathbb{N}$, $b > 1$, which we will use as base. We normally take b to be *ten*, but we don't want to write $b = 10$ now since we have not yet attached a meaning to this expression; after the fact we will obtain that the base b representation of b is 10 no matter what b is. Let D be the set of predecessors of b in $\mathbb{N}_0$; informally,

$$D = \{0, 1, \ldots, b-1\} \ .$$

Then D is a finite subset of $\mathbb{N}_0$ (III.1.4.11.) which we call the set of *digits* for base b. If b is ten, then

$$D = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} \ .$$

III.8.1.8. Now fix $x \in \mathbb{N}_0$. By III.8.1.6. and III.1.4.12., there is a smallest $n \in \mathbb{N}$ such that $b^n > x$. This n will be the number of digits in the base b expansion of x. If $n = 1$, then $d_1 := x \in D$, and the base b expansion of x is d_1.

Now suppose $n > 1$. Since

$$b^{n-1} \leq x < b^n = b \cdot b^{n-1}$$

there is a largest integer $d_n \in D$ such that $x \geq d_n b^{n-1}$. This d_n is the n'th digit in the base b expansion of x. Note that $d_n \geq 1$ since $x \geq b^{n-1}$.

III.8.1.9. Again under the assumption $n > 1$, let $x_{n-1} = x - d_n b^{n-1}$. Then $0 \leq x_{n-1} < b^{n-1}$. There is a largest integer $d_{n-1} \in D$ such that $d_{n-1} b^{n-2} \leq x_{n-1}$. If $n - 1 = 1$ (i.e. $n = 2$), then $x_{n-1} = d_{n-1}$, so

$$x = d_2 b + d_1$$

and the the base b expansion of x is $d_2 d_1$. If $n > 2$, set $x_{n-2} = x_{n-1} - d_{n-1} b^{n-2}$, i.e.

$$x_{n-2} = x - d_n b^{n-1} - d_{n-1} b^{n-2} = x - (d_n b + d_{n-1}) b^{n-1}$$

and repeat the process. After finitely many (n) steps we obtain

$$x = d_n b^{n-1} + d_{n-1} b^{n-2} + \cdots + d_2 b + d_1$$

and the base b expansion of x is $d_n d_{n-1} \dots d_2 d_1$. Note that each d_k is in D and $d_n \neq 0$ (unless $x = 0$, in which case $n = 1$). Conversely, if $c_1, \dots, c_m \in D$ and $c_m \neq 0$, and

$$x = c_m b^{m-1} + c_{m-1} b^{m-2} + \cdots + c_2 b + c_1$$

then by construction of the d_k we must have $m = n$ and $c_k = d_k$ for all k. Thus the base b expansion of x is unique.

If $x = b^m$, then $n = m + 1$ and $d_n = 1$, $d_k = 0$ for $1 \leq k < n$. In particular, the base b expansion of b is 10.

III.8.1.10. We have now defined a unique base b expansion for any nonnegative integer x. If x is a negative integer, we take its base b expansion to be $-d_n d_{n-1} \cdots d_2 d_1$, where $d_n d_{n-1} \cdots d_2 d_1$ is the decimal expansion of $-x$.

Finite Decimals

Now that we have base b expansions of integers, we continue to finite decimals. We can, and will, write $b = 10$, and we will only write out details for b equal to ten, although with slight notational changes the discussion will be valid for arbitrary b.

III.8.1.11. We first review the meaning of a finite decimal. This is an expression of the form

$$\pm m.d_1 d_2 \cdots d_n$$

where m is a nonnegative integer and $d_1, \dots, d_n$ are decimal digits between 0 and 9. This expression by definition represents the rational number

$$\pm \left[m + \frac{d_1 d_2 \cdots d_n}{10^n} \right]$$

where $d_1 d_2 \cdots d_n$ denotes a nonnegative integer in the usual way:

$$d_1 d_2 \cdots d_n = 10^{n-1} d_1 + 10^{n-2} d_2 + \cdots + 10 d_{n-1} + d_n \ .$$

Alternatively, we have

$$\pm m.d_1 d_2 \cdots d_n = \pm \left[m + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n}{10^n} \right] = \pm \left[m + \sum_{k=1}^{n} \frac{d_k}{10^k} \right] \ .$$

The m of course can also be written in decimal form in a similar manner.

III.8.1.12. Conversely, any rational number of the form

$$\frac{p}{10^q}$$

where $p, q \in \mathbb{Z}$ has a finite decimal representation: suppose $p > 0$ (the case $p \leq 0$ is an obvious variation), and $q > 0$, which can be arranged by multiplying top and bottom by a power of 10. Then we can uniquely write

$$p = 10^q m + r$$

where $0 \leq r < 10^q$, and we can then expand

$$r = 10^{q-1}d_1 + 10^{q-2}d_2 + \cdots + 10d_{q-1} + d_q$$

for digits $d_1, \ldots, d_q$. Thus the rational numbers whose denominator is a power of 10 (more precisely, the rational numbers which have such a fractional representation) are precisely the numbers with a finite decimal representation.

III.8.1.13. Definition. A rational number which can be written with a denominator which is a power of 10 is a *decadent rational*.

Any integer, or any rational number which can be written with a denominator divisible only by 2 and/or 5 is decadent. Every decadent rational is an integer times a power of 10 (and conversely). The set of decadent rational numbers is dense in $\mathbb{R}$ (in the sense that every interval in $\mathbb{R}$ contains such a number, in fact infinitely many).

III.8.1.14. There is one obvious sense in which the finite decimal expression for a decadent rational number is not unique, however: we have

$$m.d_1d_2\cdots d_n = m.d_1d_2\cdots d_n00\cdots 0$$

for any (finite) string of zeros. Thus any number of zeros can be appended or deleted at the end of the expression. Except for this, the finite decimal expansion of a decadent rational number is unique. (We can theoretically also add zeros in places to the left of the first digit in the integer part, but it is conventional not to write such zeros.)

III.8.1.15. The ordering on the decadent rational numbers is obtained easily from the decimal expansion. Working from left to right in the decimal expansions of (nonnegative) x and y, just find the first place where the expansions differ; the larger digit in this place denotes the larger number. (This is an effective procedure since there are only finitely many nonzero digits in each expansion; the expansions may have to be extended by adding zeros on the left and/or right ends to allow comparison.)

The Decimal Representation of a Real Number

III.8.1.16. We will show that every real number x has a decimal expansion of the form

$$\pm m.d_1d_2d_3\cdots$$

where m is a nonnegative integer and the d_k are digits between 0 and 9, that the decimal expansion is unique except for decadent rational numbers which have two such expansions, and that every such decimal expression is the decimal expansion of a unique real number.

We will require that the infinite decimal expansion of a decadent rational number be compatible with the finite expansion, i.e. if

$$x = \pm m.d_1 d_2 \cdots d_n$$

we should have that

$$\pm m.d_1 d_2 \cdots d_n 000 \cdots$$

is an infinite decimal expansion for x.

We also want the infinite decimal expansion to reflect the ordering on $\mathbb{R}$ in the same way as the finite expansions of III.8.1.15.. We must be careful, however: although the algorithm described in III.8.1.15. makes sense in this setting (i.e. it is a well-defined finite algorithm), it does not always give the correct answer due to the nonuniqueness of certain infinite decimal expansions (). We can, however, say the following: suppose

$$m.d_1 d_2 \cdots$$

is an infinite decimal expansion for a real number x. Then, for any n, we have

$$m + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n}{10^n} \leq x \leq m + \frac{d_1}{10} + \frac{d_2}{10^2} + \cdots + \frac{d_n + 1}{10^n} .$$

If $d_n \neq 9$, this inequality can be concisely written

$$m.d_1 d_2 \cdots d_n \leq x \leq m.d_1 d_2 \cdots d_{n-1} c_n$$

where $c_n = d_n + 1$. If $d_n = 9$, however, the right-hand inequality does not make sense, and it is notationally very complicated to write a valid version (one must "carry" back past any 9's at the end of the expression).

As we will see, these two provisions completely determine decimal expansions and give a hint how to find them.

III.8.1.17. The sign in an infinite decimal expression is self-explanatory. From now on, we will only deal with nonnegative real numbers; the decimal expansion of a negative number x will be the decimal expansion of $|x|$ preceded by a minus sign.

III.8.1.18. If x is nonnegative, the m is also clear: it is the unique nonnegative integer such that $m \leq x < m + 1$ (). This m is called the *integer part* of x, denoted $\lfloor x \rfloor$.

III.8.1.19. By subtracting off the integer part, we may reduce the problem of defining (or computing) the decimal expansion of a number x to the case where $0 \leq x \leq 1$. We will make this restriction from now on.

Generating the Decimal Expansion of a Real Number

Finding a Real Number with a Given Decimal Expansion

Decimal Expansions of Rational and Irrational Numbers

Defining Real Numbers as Decimals

III.8.1.20. Since there is essentially a one-one correspondence between real numbers and decimal expressions, it is possible to define the real numbers via infinite decimals, and this approach is taken in some references. But there are some serious difficulties with this approach.

III.8.1.21. Besides the general problem of defining the real numbers as *being* one particular model, be it decimal expansions, Dedekind cuts, equivalence classes of Cauchy sequences, or whatever, there are two additional problems in defining the real numbers via decimal expressions:

(i) The correspondence is not quite one-to-one.

(ii) It is not easy or obvious how to define or describe the algebraic operations on $\mathbb{R}$ using the decimal expressions.

III.8.1.22. The first difficulty is easily overcome. One way is just to exclude decimals ending in an infinite string of 9's.

III.8.1.23. The second difficulty is not easily overcome. In school we learn algorithms for adding, subtracting, multiplying, and dividing integers expressed in decimal notation, including memorization of the multiplication table for one-digit natural numbers; later, we also learn the simple modifications needed to use the same algorithms for arithmetic of *finite* (i.e. terminating) decimals. But there are no finite algorithms for adding or multiplying infinite decimals, since there is no place to "start". For example, can you even determine the first decimal digit of

$$x = (0.2531486073\cdots) + (0.3468513926\cdots)$$

without further information? Could you even be sure this digit can be determined in a finite length of time with a finite amount of information about the decimal expansions? (The problem is even worse, and more obvious, for multiplication.)

III.8.1.24. The only reasonable way to proceed is to appeal to one of the other approaches, most obviously the Cauchy sequence approach. In the example of the last paragraph, although even the first decimal digit of x is not clearly determined by the information given, x itself is determined to a high degree of accuracy (within 2×10^{-10}) by the ten decimal places of the given expressions. If one has the complete decimal expansions of the summands, or an algorithm for generating them, the number x can be determined to arbitrary accuracy, i.e. a sequence of rational numbers (terminating decimals) can be generated which converges to x.

III.8.1.25. The problem is that real numbers which are very close together can have different expansions starting early in the expression if they happen to be on opposite sides of a number with a small terminating expansion (i.e. an integer times a small negative power of 10).

Another way of expressing the difficulty is that if we know that real numbers x and y satisfy $|x-y| < 10^{-n}$, we do not necessarily know that x and y agree to n, or even $n-1$, decimal places, although they "usually" will; in unusual cases they may not even agree to one decimal place if one expansion has a long string of 0's after the first place and the other a corresponding long string of 9's. They will not even necessarily agree when *rounded* to one decimal place, e.g. if $x = .24999999999999999$ and $y = .25000000000000001$. In fact, for any n, the functions which truncate or round a real number to n decimal places are step functions with jump discontinuities of height 10^{-n} ().

III.8.2. Continued Fractions

A somewhat more complicated, but in some ways more efficient, way to represent real numbers by sequences of natural numbers is by continued fractions. The theory of continued fractions is usually, and rightly, considered to be a part of number theory; however, the applications in analysis and probability, along with applications in closely related fields such as dynamical systems, are important enough that the basic theory of continued fractions should be a more integral part of the standard analysis curriculum than is currently the case.

While continued fractions give a much better way of representing real numbers than decimal representations for some purposes (such as finding good rational approximations), they are not nearly as good as decimal representations for the purposes of doing arithmetic, and this shortcoming is the principal reason their use has not become so widespread or popular. But mathematicians need to be familiar with both ways of representing numbers since they are useful for different purposes.

We will cover only a few important basic properties of the theory of continued fractions here which are most relevant in analysis and probability, and omit most of the many fascinating connections with number theory (some of these are described at the end of the section and in the Exercises). Books on number theory such as the classic [HW08] (there are also many others; perhaps my favorite is the little gem [Khi97]) cover continued fractions in more detail; [Sta78] has a thorough and geometrically appealing treatment of the theory.

The fundamental difference between decimal representations and continued fractions is that decimal representations require an initial choice of a base, and all representations are tied to that choice of base; some numbers have nice representations with respect to the chosen base, while others do not. But the continued fraction representation of a number is not tied to any initial choice of base; indeed, one way of thinking of the theory is that at each step in the expansion of a number, the perfect "base" is used for that number at that step.

As a result, the theory of continued fractions concerns the most efficient approximation of real numbers by rational numbers. Of course, the finite parts of the decimal expansion of a real number give rational approximations, but one can do much better in general by a different approach. Any real number can be approximated arbitrarily closely by rational numbers, but close rational approximations to irrational numbers will necessarily have large denominators. By an "efficient" approximation we will mean an approximation which is close by some measure depending on the size of the denominator of the rational number when expressed in lowest terms. For example, we could ask:

III.8.2.1. QUESTION. For an irrational number x, and a natural number n, what is the rational number

with denominator $\leq n$ which most closely approximates x?

The theory of continued fractions gives a complete answer to this question.

III.8.2.2. A *continued fraction* is an expression of the form

$$x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_n + \cdots}}}}} .$$

This is just a formal expression, which may be finite or infinite. In interpreting the meaning of this expression, the x_k may stand for numbers, functions, indeterminates, ... For example, there has been a large mathematical industry in expanding functions in continued fractions; cf. [Wal48], [JT80], [LW08]. For us, the x_k will always stand for positive real numbers, usually natural numbers.

More general continued fraction expressions of this type are sometimes considered, where the numerators are not necessarily 1's; in this case, continued fractions with 1's in the numerators (the only kind we will consider) are called *simple continued fractions*.

III.8.2.3. DEFINITION. A *finite continued fraction* of depth $n \in \mathbb{N}$ is an expression of the form

$$x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_{n-1} + \cfrac{1}{x_n}}}}}} .$$

The theory of continued fraction expansions of real numbers is based on the following "obvious" fact, which can be easily proved by induction (Exercise ()).

III.8.2.4. PROPOSITION. (i) Let $x_1, \ldots, x_n$ be positive real numbers, and x_0 a real number. Then the expression

$$x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_{n-1} + \cfrac{1}{x_n}}}}}}$$

unambiguously defines a real number by the usual rules of algebra. If $x_0, \ldots, x_n$ are rational, then the number defined by the expression is rational.
(ii) If $x_1, \ldots, x_n, x'_n$ are positive real numbers, and x_0 is a real number, and

$$x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_{n-1} + \cfrac{1}{x_n}}}}}} = x_0 + \cfrac{1}{x_1 + \cfrac{1}{x_2 + \cfrac{1}{x_3 + \cfrac{1}{\ddots + \cfrac{1}{x_{n-1} + \cfrac{1}{x'_n}}}}}} ,$$

then $x_n = x'_n$.

Statement (i) can fail if the x_k ($k \geq 1$) are not positive, since then a denominator of zero can arise during simplification.

It is notationally clumsy to write these complex fractions; we will use the common notation $[x_0; x_1, \ldots, x_n]$ for the fraction in (i). Variations of this notation, and many other notations for continued fractions, are used in the literature.

We now define a function which will be used repeatedly:

III.8.2.5. DEFINITION. Let x be a real number. Then the *integer part* of x, denoted $\lfloor x \rfloor$, is the largest integer which is $\leq x$.

We can now inductively define the continued fraction expansion of a real number:

III.8.2.6. DEFINITION. Let x be a real number. Define numbers as follows:

Set $\alpha_0 = x$ and $r_0 = a_0 = \lfloor \alpha_0 \rfloor$. Let $p_0 = r_0$ and $q_0 = 1$.

Set $\beta_1 = \alpha_0 - a_0 = x - r_0$. Then $0 \leq \beta_1 < 1$.

If $\beta_1 = 0$ (i.e. if $x \in \mathbb{Z}$), terminate the expansion. Otherwise set $\alpha_1 = \frac{1}{\beta_1}$ and $a_1 = \lfloor \alpha_1 \rfloor$. Then $a_1 \in \mathbb{N}$.

Set $r_1 = a_0 + \frac{1}{a_1} = \frac{a_0 a_1 + 1}{a_1}$. Then $r_1 \in \mathbb{Q}$, and $r_0 < x \leq r_1$. Set $p_1 = a_0 a_1 + 1$, $q_1 = a_1$, so $r_1 = \frac{p_1}{q_1}$, and define β_2 by $\beta_2 = \alpha_1 - a_1$, i.e. $x = a_0 + \frac{1}{a_1 + \beta_2}$. Then $0 \leq \beta_2 < 1$.

If $\beta_2 = 0$ (i.e. if $x = r_1$), terminate the expansion. Otherwise set $\alpha_2 = \frac{1}{\beta_2}$ and $a_2 = \lfloor \alpha_2 \rfloor \in \mathbb{N}$.

Set $r_2 = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = \frac{p_2}{q_2}$. Then $r_2 \leq x < r_1$. Define $\beta_3 = \alpha_2 - a_2$ (i.e. $x = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \beta_3}}$). Then $0 \leq \beta_3 < 1$.

Continue inductively defining β_n, α_n, a_n, $r_n = \frac{p_n}{q_n}$ for all n (or until the process terminates).

The finite or infinite sequence $[a_0; a_1, a_2, \ldots]$ is called the *continued fraction* (expression, expansion, decomposition) of x. The a_n are called the *terms* (or *elements*) of the continued fraction.

The numbers a_n, r_n, p_n, q_n, α_n, β_n depend on x; we will write them $a_n(x)$, etc., when it is necessary to specify the dependence on x.

The name "continued fraction" for this construction is appropriate, since the idea is to write

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\ddots}}}} \ .$$

We will discuss later how this expression can be precisely interpreted as the limit of the finite continued fractions r_n.

It is easy to verify the next result (note that we have a recursive formula $\alpha_n = a_n + \frac{1}{\alpha_{n+1}}$ for all n):

III.8.2.7. PROPOSITION. Let $x \in \mathbb{R}$, and $n \in \mathbb{N}$. Then

$$x = [a_0(x); a_1(x), \ldots, a_{n-1}(x), \alpha_n(x)]$$
$$r_n(x) = [a_0(x); a_1(x), \ldots, a_n(x)] \ .$$

We thus get the following "contraction" principle:

III.8.2.8. PROPOSITION. Let $x \in \mathbb{R}$, and $n \in \mathbb{N} \cup \{0\}$. If the continued fraction construction for x lasts at least $n+1$ steps (i.e. $a_n(x)$ is defined), then $a_n(x) = a_k(\alpha_{n-k}(x))$ for $0 \leq k \leq n$.

We then get a recursive formula for r_n:

III.8.2.9. COROLLARY. Let $x \in \mathbb{R}$, and $n \in \mathbb{N} \cup \{0\}$. If the continued fraction construction for x lasts at least $n+1$ steps, then

$$r_n(x) = a_0 + \frac{1}{r_{n-1}(\alpha_1(x))} = a_0 + \frac{1}{a_1 + \frac{1}{r_{n-2}(\alpha_2(x))}} = \cdots = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots \; + \frac{1}{r_0(\alpha_n(x))}}}}.$$

We also have the following uniqueness statement:

III.8.2.10. PROPOSITION. Let $n \in \mathbb{N}$, $b_0 \in \mathbb{Z}$, $b_1, \ldots, b_n \in \mathbb{N}$. If $x \in \mathbb{R}$, and the continued fraction expansion of x lasts at least $n+2$ steps (i.e. $a_{n+1}(x)$ is defined), then $a_k(x) = b_k$ for $0 \leq k \leq n$ if and only if $x = [b_0; b_1, \ldots, b_n, \alpha]$ for some $\alpha > 1$, i.e. if and only if x is in the open interval between $[b_0; b_1, \ldots, b_n]$ and $[b_0; b_1, \ldots, b_{n-1}, b_n + 1]$.

III.8.2.11. There is one slight ambiguity in continued fraction expansions, however: for any $b_0 \in \mathbb{Z}$, $b_1, \ldots, b_n \in \mathbb{N}$,

$$[b_0; b_1, \ldots, b_n, 1] = [b_0; b_1, \ldots, b_n + 1]$$

and the second expression, not the first, is the correct continued fraction expansion of the corresponding rational number. Thus a finite continued fraction expression ending in 1 can be "contracted", and a terminating continued fraction expansion (of length at least 2) cannot end in 1.

Since the r_n are rational, it is obvious that if x is irrational, then the construction never terminates, so the continued fraction expression for x is an infinite sequence. Conversely, if x is rational, then the construction necessarily terminates, although this is not immediately obvious:

III.8.2.12. PROPOSITION. If r is rational, then the continued fraction expansion of x terminates.

PROOF: Write $x = \frac{p}{q}$. We may suppose $p, q \in \mathbb{N}$ for simplicity (this will happen after one step if $x < 0$, and the result is obvious if $x = 0$). If $p < q$, then $a_0 = 0$ and $\alpha_1 = \frac{q}{p}$; otherwise $x = a_0 + \frac{s}{q}$ with $0 \leq s < q \leq p$. If $s = 0$, the process terminates. Otherwise, $\alpha_1 = \frac{q}{s} = a_1 + \frac{t}{s}$ with $0 \leq t < s < q$. If $t = 0$, the process terminates. Otherwise, $\alpha_2 = \frac{s}{t}$, and we have $s, t \in \mathbb{N}$, $s < p$, $t < q$, so after 2 steps the remaining procedure computes the continued fraction expansion of a fraction with strictly smaller numerator and denominator. Thus the process must terminate in no more than $2q + 1$ steps. 👍

In fact, for $x = \frac{p}{q}$ the continued fraction construction is really the same as the Euclidean Algorithm for determining the greatest common divisor of p and q. The termination estimate of $1 + 2q$ steps is rather

crude: it is known (LAMÉ'S Theorem) that if $p, q > 0$, then the number of steps to termination is bounded by

$$2 + 5\min(\log_{10}(p), \log_{10}(q))$$

which is roughly 5 times the number of decimal digits in the smaller of p and q.

III.8.2.13. If $[b_0; b_1, b_2, \dots]$ is any sequence with $b_0 \in \mathbb{Z}$ and $b_n \in \mathbb{N}$ for $n \geq 1$, we can define r_n in the analogous manner, i.e.

$$r_n = [b_0; b_1, \dots, b_n] = b_0 + \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{\ddots + \frac{1}{b_n}}}} \ .$$

By III.8.2.10., the b_n can be recovered from the continued fraction expansions of the r_n.

III.8.2.14. In III.8.2.6., we were not completely precise about the definition of p_n and q_n. These are defined by simply converting the compound fraction to a simple fraction and taking numerator and denominator. For example, we have

$$r_2 = a_0 + \frac{1}{a_1 + \frac{1}{a_2}} = a_0 + \frac{a_2}{a_1 a_2 + 1} = \frac{a_0 a_1 a_2 + a_0 + a_2}{a_1 a_2 + 1}$$

so

$$p_2 = a_0 a_1 a_2 + a_0 + a_2 = p_0 + a_2 p_1, \qquad q_2 = a_1 a_2 + 1 = q_0 + a_2 q_1 \ .$$

In fact, if $[a_0; a_1, a_2, \dots]$ is any sequence with $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for $n \geq 1$, and

$$r_n = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \frac{1}{a_n}}}}$$

and p_n, q_n are defined from r_n by clearing complex fractions, the same formulas hold for $n = 2$.

III.8.2.15. PROPOSITION. For any $n \geq 2$,

(i) $p_n = p_{n-2} + a_n p_{n-1}$ and $q_n = q_{n-2} + a_n q_{n-1}$.

(ii) $p_{n-1} q_{n-2} - p_{n-2} q_{n-1} = (-1)^n$.

In particular, p_n and q_n are relatively prime, and $q_n > q_{n-1}$.

In particular, these formulas hold for a sequence arising as the continued fraction of a real number. (We will see later (III.8.2.19.) that every sequence occurs as a continued fraction.)
PROOF: We prove (i) by induction, using III.8.2.8. and III.8.2.9.. The formula holds for $n = 2$ by the previous calculation. Suppose it holds for n for all x. Fix x, and let $\alpha_1 = \alpha_1(x)$. Then, from III.8.2.9., $p_{k-1}(\alpha_1) = q_k(x)$ and $q_{k-1}(\alpha_1) = p_k(x) - a_0(x) q_k(x)$ for all k. Then

$$p_{n+1}(x) = a_0(x) p_n(\alpha_1) + q_n(\alpha_1) \ .$$

By the inductive hypothesis applied to α_1, this equals

$$\begin{aligned} a_0(x)[p_{n-2}(\alpha_1) + a_n(\alpha_1)p_{n-1}(\alpha_1)] + [q_{n-2}(\alpha_1) + a_n(\alpha_1)q_{n-1}(\alpha_1)] \\ = a_0(x)[q_{n-1}(x) + a_{n+1}(x)q_n(x)] \\ +([p_{n-1}(x) - a_0(x)q_{n-1}(x)] + a_{n+1}(x)[p_n(x) - a_0(x)q_n(x)]) \\ = p_{n-1}(x) + a_{n+1}(x)p_n(x) \ . \end{aligned}$$

A similar calculation shows that

$$q_{n+1}(x) = q_{n-1}(x) + a_{n+1}(x)q_n(x) \ .$$

We now prove (ii) by induction. It holds for $n = 2$ by direct calculation. If it holds for n, then by (i)

$$\begin{aligned} p_nq_{n-1} - p_{n-1}q_n = (p_{n-2} + a_np_{n-1})q_{n-1} - p_{n-1}(q_{n-2} + a_nq_{n-1}) \\ = p_{n-2}q_{n-1} - p_{n-2}q_{n-1} + a_np_{n-1}q_{n-1} - a_np_{n-1}q_{n-1} \\ = -(p_{n-1}q_{n-2} - p_{n-2}q_{n-1}) = -(-1)^n = (-1)^{n+1} \end{aligned}$$

so the formula holds for $n + 1$.
It follows from (ii) that p_n and q_n are relatively prime, and it is obvious from (i) that $q_n > q_{n-1}$ since $q_{n-2} \geq 1$ and $a_n \geq 1$. 👍

III.8.2.16. Corollary. (i) Let $[a_0; a_1, \dots]$ be a sequence with $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for $n \geq 1$, and let (r_n) be the corresponding sequence of finite continued fractions. Then

$$r_0 < r_2 < r_4 < \cdots < r_5 < r_3 < r_1$$

and $|r_n - r_{n-1}| = \frac{1}{q_nq_{n-1}}$ for $n \geq 1$.
(ii) Let $x \in \mathbb{R}$ with nonterminating continued fraction expansion $[a_0; a_1, \dots]$, and (r_n) the corresponding sequence of finite continued fractions. Then

$$r_0 < r_2 < r_4 < \cdots < x < \cdots < r_5 < r_3 < r_1$$

and $r_n \to x$. In fact, for $n \geq 1$,

$$|x - r_n| < |r_{n+1} - r_n| = \frac{1}{q_nq_{n+1}} < \frac{1}{a_{n+1}q_n^2} \leq \frac{1}{q_n^2}$$

(the extreme inequality also obviously holds for $n = 0$).

Because of (ii), r_n is often called the n'th *continued fraction convergent* to x, and we may justifiably write

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \frac{1}{\cdots}}}} \ .$$

An alternate way of phrasing the convergence is that the alternating series

$$a_0 + \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{q_nq_{n-1}}$$

converges to x. (Note that this series is absolutely convergent since $q_n \geq n$ for all n.)

III.8.2.17. Corollary. If x and y are distinct real numbers, their continued fraction expansions are different.

III.8.2.16. almost shows that every sequence as in (i) is the continued fraction expansion of a (necessarily irrational) real number. For under the conditions of (i), by the Nested Intervals Property there is a unique $x \in \mathbb{R}$ satisfying

$$r_0 < r_2 < r_4 < \cdots < x < \cdots < r_5 < r_3 < r_1$$

and we need only show that the continued fraction expansion of this x is $[a_0; a_1, \dots]$. This follows from the next proposition, which is essentially a rephrasing of III.8.2.10.:

III.8.2.18. Proposition. Let $[b_0; b_1, \dots]$ be a sequence with $b_0 \in \mathbb{Z}$ and $b_n \in \mathbb{N}$ for $n \geq 1$, and let (r_n) be the corresponding sequence of finite continued fractions. If x is between r_n and r_{n+1}, then $a_k(x) = b_k$ for $0 \leq k \leq n$.

III.8.2.19. Corollary. Let $[a_0; a_1, \dots]$ be a sequence with $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for $n \geq 1$, and let (r_n) be the corresponding sequence of finite continued fractions. Let x be the unique real number satisfying

$$r_0 < r_2 < r_4 < \cdots < x < \cdots < r_5 < r_3 < r_1 \ .$$

Then the continued fraction expansion of x is $[a_0; a_1, \dots]$.

We also have the following useful corollary about the continuity of continued fraction decompositions:

III.8.2.20. Corollary. Let $x \in \mathbb{R}$ have continued fraction expansion of length at least $n+1$, and (r_k) the corresponding convergents. If y is between r_n and r_{n+1}, then $a_k(y) = a_k(x)$ for $0 \leq k \leq n$.

We have the following relations between x and the $\alpha_k(x)$, which can be easily proved by induction using III.8.2.15.(i):

III.8.2.21. Proposition. Let $x \in \mathbb{R}$, and $n \in \mathbb{N}$. If the continued fraction construction for x lasts at least $n+1$ steps, we have

$$x = \frac{\alpha_1 a_0 + 1}{\alpha_1}$$

and, if $n \geq 2$,

$$x = \frac{\alpha_n p_{n-1} + p_{n-2}}{\alpha_n q_{n-1} + q_{n-2}} \ .$$

This gives a refinement of III.8.2.16.(ii) with a lower bound:

III.8.2.22. Corollary. Let x be an irrational number, and $r_n = \frac{p_n}{q_n}$ its n'th continued fraction convergent. Then

$$\frac{1}{q_n(q_n + q_{n+1})} < |x - r_n| < \frac{1}{q_n q_{n+1}} .$$

We combine and summarize our main results in a theorem:

III.8.2.23. Theorem. The continued fraction expansion establishes a one-one correspondence between the set $\mathbb{J}$ of irrational numbers and the set of sequences $[a_0; a_1, a_2, \dots]$ with $a_0 \in \mathbb{Z}$ and $a_n \in \mathbb{N}$ for $n \geq 1$. If (x_m) is a sequence of irrational numbers and x an irrational number, then $x_m \to x$ if and only if, for each n, there is an m_0 such that $a_n(x_m) = a_n(x)$ for all $m \geq m_0$.

Approximation by Continued Fractions

We conclude with a survey of some number-theoretic results about continued fractions, mostly stated without proof.

The continued fraction convergents to an irrational number x are the "best" rational approximants to x in the sense of Question III.8.2.1.. We first make a couple of definitions:

III.8.2.24. Definition. Let x be an irrational number. A rational number $\frac{p}{q}$ with $p \in \mathbb{Z}$, $q \in \mathbb{N}$ is a *good approximation* to x if

$$\left| x - \frac{p'}{q'} \right| > \left| x - \frac{p}{q} \right|$$

for all $p', q' \in \mathbb{Z}$, $1 \leq q' \leq q$, $(p', q') \neq (p, q)$.
A rational number $\frac{p}{q}$ with $p \in \mathbb{Z}$, $q \in \mathbb{N}$ is a *best approximation* to x if

$$|q'x - p'| > |qx - p|$$

for all $p', q' \in \mathbb{Z}$, $1 \leq q' \leq q$, $(p', q') \neq (p, q)$.

A best approximation to x is obviously a good approximation to x. Good approximations and best approximations are sometimes respectively called "best approximations of the first kind" and "best approximations of the second kind."

III.8.2.25. Theorem. The continued fraction convergents $\frac{p_n}{q_n}$ to an irrational number x for $n \geq 1$ are precisely the best approximations to x, i.e. every continued fraction convergent is a best approximation, and every best approximation is a continued fraction convergent.

Not every good approximation is a continued fraction convergent, but all good approximations to x are *secondary convergents*, numbers of the form $[a_0; a_1, \dots, a_{n-1}, c_n]$, where $[a_0; a_1, \dots]$ is the continued fraction expansion of x and $1 \leq c_n \leq a_n$. If $a_n > 1$ and $a_n/2 < c_n \leq a_n$, then the secondary convergent is a good approximation; if a_n is even and $c_n = a_n/2$, the secondary convergent is sometimes, but not always, a good approximation, cf. [].

These solve Question III.8.2.1.: since the solution to III.8.2.1. for a given n is obviously a good approximation, one only needs to find the good approximation with the largest denominator $\leq n$. One only needs to know the continued fraction expansion to the point where $q_k \geq n$.

The next theorem is an important result about continued fractions for number theory purposes, and further justifies the statement that the continued fraction convergents to a real number are the "best" rational approximants.

III.8.2.26. THEOREM. Let x be an irrational number, and, for $n \in \mathbb{N}$, let $r_n = \frac{p_n}{q_n}$ be the n'th convergent to x.
(i) If $n \geq 2$, then at least one of the following inequalities holds:

$$|x - r_{n-1}| < \frac{1}{2q_{n-1}^2} \text{ or } |x - r_n| < \frac{1}{2q_n^2} \ .$$

(ii) Conversely, if $r = \frac{p}{q}$ $(p, q \in \mathbb{Z})$ satisfies

$$|x - r| < \frac{1}{2q^2}$$

then $r = r_n$ for some n.

PROOF: We prove only (i). Since x is between r_{n-1} and r_n, we have

$$|x - r_{n-1}| + |x - r_n| = |r_n - r_{n-1}| = \frac{1}{q_{n-1}q_n} \ .$$

If the statement is false, i.e.

$$|x - r_{n-1}| \geq \frac{1}{2q_{n-1}^2} \text{ and } |x - r_n| \geq \frac{1}{2q_n^2}$$

we would have

$$\frac{1}{2q_{n-1}^2} + \frac{1}{q_n^2} \leq \frac{1}{q_{n-1}q_n}$$

and multiplying both sides by $2q_{n-1}^2 q_n^2$, we have

$$q_n^2 + q_{n-1}^2 \leq 2q_n q_{n-1}$$
$$(q_n - q_{n-1})^2 \leq 0$$

which is a contradiction since $q_n \neq q_{n-1}$. 👍

Actually, one can do a little better:

III.8.2.27. THEOREM. Let x be an irrational number, and, for $n \in \mathbb{N}$, let $r_n = \frac{p_n}{q_n}$ be the n'th convergent to x. Then for infinitely many n (at least one of every three consecutive n),

$$|x - r_n| < \frac{1}{\sqrt{5}q_n^2} \ .$$

The example of the golden ratio (Exercise ()) shows that $\sqrt{5}$ is the largest possible constant M for which $|x - r_n| < \frac{1}{Mq_n^2}$ for infinitely many n in general.

III.8.2.28. If the $a_n(x)$ are unbounded, which is the case for almost all x (III.8.2.29.), we can do much better, since if $a_{n+1}(x)$ is large, $r_n(x)$ is a very good approximation to x by III.8.2.16.. For example, since $a_4(\pi) = 292$, we have that $r_3(\pi) = \frac{355}{113}$ satisfies

$$\left|\pi - \frac{355}{113}\right| < \frac{1}{292 \cdot 113^2} \approx 2.7 \times 10^{-7}$$

so $\frac{355}{113}$ agrees with π to (at least, in fact exactly) 6 decimal places. (The approximation $\frac{355}{113}$ to π was found by the Chinese mathematician Zu Chongzhi in the fifth century.) A theorem of Lochs () says that for almost all $x \in \mathbb{J}$ (in the sense of Lebesgue measure), $r_n(x)$ asymptotically agrees with x to about n decimal places.

Asymptotic Behavior of Continued Fractions

Here are some asymptotic results about the growth rate of the coefficients $a_n(x)$ for "typical" irrational numbers:

III.8.2.29. Theorem. Let $\phi : \mathbb{N} \to (1, \infty)$ be a function. Then
(i) If $\sum_{n=1}^{\infty} \frac{1}{\phi(n)}$ diverges, then the set of $x \in \mathbb{J}$ for which $a_n(x) = O(\phi(n))$ as $n \to \infty$ has Lebesgue measure 0. In particular, the set of $x \in \mathbb{J}$ for which $(a_n(x))$ is bounded has Lebesgue measure 0.
(ii) If $\sum_{n=1}^{\infty} \frac{1}{\phi(n)}$ converges, then $a_n(x) = o(\phi(n))$ as $n \to \infty$ for almost all $x \in \mathbb{J}$, i.e. the set of x for which this fails has Lebesgue measure 0.

Thus, almost all irrational numbers have many a_n which are "large" but which do not grow too rapidly. Theorem III.8.2.29. is roughly a version of the Borel-Cantelli Lemma () (it is not exactly the Borel-Cantelli Lemma since the random variables (a_n) are not quite independent; cf. ().) In contrast, Khinchin [Khi97] proved the following remarkable theorem, which is essentially a version of the Strong Law of Large Numbers ():

III.8.2.30. Theorem. For almost all $x \in \mathbb{J}$, the geometric mean of the a_n approaches a constant

$$\kappa = \prod_{k=1}^{\infty} \left[1 + \frac{1}{k(k+2)}\right]^{\log_2 k} = \prod_{k=1}^{\infty} k^{\log_2\left[1 + \frac{1}{k(k+2)}\right]} = 2.685452\cdots$$

i.e. $\lim_{n\to\infty} [a_1(x)a_2(x)\cdots a_n(x)]^{1/n} = \kappa$.

So, for almost all irrational numbers x, "most" of the $a_n(x)$ are small. Note, however, that the *arithmetic* mean of the a_n's almost always diverges: since $\sum \frac{1}{n \log n}$ diverges, for almost all x we have $a_n(x) > n \log n$ for infinitely many n by III.8.2.29.(i), and thus, for almost all x,

$$\frac{1}{n} \sum_{k=1}^{n} a_k(x) > \log n$$

for infinitely many n.

Khinchin's result is more general:

III.8.2.31. THEOREM. Let $f : \mathbb{N} \to [0, \infty)$ be a function for which there exist positive constants C and δ such that

$$f(k) < Ck^{1/2-\delta}$$

for all k. Then, for almost all $x \in \mathbb{J}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(a_k(x)) = \sum_{k=1}^{\infty} f(k) \log_2 \left[1 + \frac{1}{k(k+2)}\right] .$$

III.8.2.30. is the case $f(k) = \log k$. Another interesting case is to let $f(k_0) = 1$ for some k_0 and $f(k) = 0$ for $k \neq k_0$. This case shows that the number k_0 appears with the same asymptotic relative frequency in the continued fraction expansion of almost all numbers, and the frequency is

$$\log_2(k_0^2 + 2k_0 + 1) - \log_2(k_0^2 + 2k_0) .$$

For example, for almost all x, the relative frequency of 1 in the continued fraction expansion of x is

$$\log_2 4 - \log_2 3 = .4150 \cdots$$

i.e. asymptotically about 41.5% of the terms in the continued fraction expansion of x are 1's. Similarly, about 17% are 2's, 9.3% are 3's, etc. This result is reminiscent of, but not the same as, Benford's Law ().

Approximation of Algebraic and Transcendental Numbers

III.8.2.32. DEFINITION. A real (or complex) number x is *algebraic* if x is a root of a polynomial with integer coefficients. A real (or complex) number which is not algebraic is *transcendental*.

A rational number $\frac{p}{q}$ is a root of the polynomial $qx - p$, hence is algebraic. It is a standard theorem of abstract algebra that every algebraic number is a root of a unique (up to constant multiple) polynomial with integer coefficients which is irreducible over $\mathbb{Z}$ (or equivalently over $\mathbb{Q}$). The degree of this polynomial is the *degree* of the algebraic number. Degree 1 algebraic numbers are precisely the rational numbers; degree 2 algebraic numbers are called *quadratic numbers*.

Our first result is a simple argument showing that irrational algebraic numbers cannot be approximated too closely by rational numbers. This result is due to LIOUVILLE in 1844.

III.8.2.33. THEOREM. Let x be an algebraic number of degree $n > 1$. Then there is a constant C, $0 < C < 1/2$, such that

$$|x - r| > \frac{C}{q^n}$$

for all rational numbers $r = \frac{p}{q}$.

PROOF: Let x be a root of an irreducible polynomial f of degree n with integer coefficients, and let M be the maximum of $|f'|$ on the interval $\left[x - \frac{1}{2}, x + \frac{1}{2}\right]$. Then $M > 0$ since f is not constant. By the Mean Value Theorem, if $r = \frac{p}{q}$ satisfies $|x - r| < \frac{1}{2}$, there is a c between x and r such that

$$|q^n f(r)| = |q^n(f(r) - f(x))| = q^n |f'(c)||r - x| \leq Mq^n |x - r| .$$

Note that $q^n f(r) \in \mathbb{Z}$, and $f(r) \neq 0$ since f has no rational roots. Thus, if $C < \frac{1}{M}$, then $1 \leq Mq^n|x-r|$, $|x-r| > \frac{C}{q^n}$ if $|x-r| < \frac{1}{2}$. Also, $M > 2$ since there is an $s \in \mathbb{Z}$ with $|x-s| < \frac{1}{2}$, and by the above argument

$$1 \leq |f(s)| \leq M|x-s|$$

so $C < \frac{1}{2}$ and $\left|x - \frac{p}{q}\right| > \frac{C}{q^n}$ also if $\left|x - \frac{p}{q}\right| > \frac{1}{2}$. 👍

This result shows that irrational numbers which can be approximated very closely by rational numbers must be transcendental:

III.8.2.34. DEFINITION. An irrational number x is a *Liouville number* if, for every n, there are $p, q \in \mathbb{Z}$ with $q \geq 2$ and

$$\left|x - \frac{p}{q}\right| < \frac{1}{q^n} \ .$$

By III.8.2.33., every Liouville number is transcendental.

III.8.2.35. EXAMPLE. Let $x = \sum_{k=1}^{\infty} 10^{-k!}$. For each n, we can take $\frac{p}{q} = \sum_{k=1}^{n} 10^{-k!}$. This was LIOUVILLE'S original example, and the first explicit number to be proved transcendental.

The set L of Liouville numbers is quite an interesting subset of $\mathbb{R}$. It is very "small" in a measure-theoretic sense, but "large" in a topological (category) sense:

III.8.2.36. THEOREM. L is a dense G_δ; hence the complement of L is a meager set.

PROOF: For each n, let

$$U_n = \bigcup_{\frac{p}{q} \in \mathbb{Q}} \left(\frac{p}{q} - \frac{1}{q^n}, \frac{p}{q} + \frac{1}{q^n}\right) \ .$$

Then U_n is an open set containing $\mathbb{Q}$, hence dense; and $L = \mathbb{J} \cap \cap_n U_n$. Since $\mathbb{J}$ and $\cap_n U_n$ are dense G_δ's, so is L. 👍

III.8.2.37. THEOREM. The set L has s-dimensional Hausdorff measure 0, for every $s > 0$. So L has Hausdorff dimension 0 and Lebesgue measure 0.

PROOF: It suffices to show that if $0 < s \leq 1$ is fixed, the s-dimensional Hausdorff measure of $L \cap (-m, m)$ is 0 for every $m \in \mathbb{N}$, i.e. that for every $\epsilon > 0$ there is a sequence (I_k) of open intervals with $\sum_k [\ell(I_k)]^s < \epsilon$ and $L \cap (-m, m) \subseteq \cup I_k$. Fix $m \in \mathbb{N}$ and $\epsilon > 0$. Then, for any n,

$$L \cap (-m, m) \subseteq U_n \cap (-m, m) = \bigcup_{q=2}^{\infty} \bigcup_{p=-mq}^{mq} \left(\frac{p}{q} - \frac{1}{q^n}, \frac{p}{q} + \frac{1}{q^n}\right) \ .$$

We have that

$$\ell\left(\left(\frac{p}{q}-\frac{1}{q^n},\frac{p}{q}+\frac{1}{q^n}\right)\right)=\frac{2}{q^n}$$

so, for sufficiently large n, we have the sum of the s powers of the lengths of the intervals is

$$\sum_{q=2}^{\infty}\sum_{p=-mq}^{mq}\left(\frac{2}{q^n}\right)^s=\sum_{q=2}^{\infty}2^s(2mq+1)\frac{1}{q^{ns}}\le\sum_{q=2}^{\infty}2^s(2mq+q)\frac{1}{q^{ns}}$$

$$=2^s(2m+1)\sum_{q=2}^{\infty}\frac{1}{q^{ns-1}}\le 2^s(2m+1)\int_1^{\infty}\frac{1}{x^{ns-1}}\,dx=\frac{2^s(2m+1)}{ns-2}<\epsilon\ .$$



It can be shown (Exercise ()) that every Liouville number is in the complement of the set of III.8.2.29.(ii) for some ϕ, giving another proof that L has Lebesgue measure 0.

III.8.2.38. However, it is far from true that all transcendental numbers have rapidly converging continued fraction convergents (or, equivalently, rapidly growing continued fraction coefficients), as III.8.2.29.(ii) shows. In fact, there are transcendental numbers whose continued fraction convergents converge nearly as slowly as those of φ (Exercise III.8.4.2.), namely numbers whose continued fraction expansion is all 1's except for some (infinitely many!) sparsely distributed 2's (to see that there are transcendental numbers with such continued fraction expansions, it suffices to note that there are uncountably many such sequences of 1's and 2's no matter what your definition of "sparsely distributed" may be, as long as a subset of a sparsely distributed set is sparsely distributed. Actually, it is conjectured that every algebraic number of degree > 2 has unbounded continued fraction expansion; if so, all numbers with bounded continued fraction expansion which is not eventually periodic (cf. Exercise III.8.4.4.) would be transcendental.) In fact, there appears to be no foolproof way to distinguish between algebraic and transcendental numbers by a statistical analysis of the pattern of the terms in the continued fraction expansion.

Liouville's theorem has been improved by various authors, beginning with Thue in 1909. The best result known is due to Roth in 1955:

III.8.2.39. Theorem. Let x be an irrational algebraic number. Then for every $\epsilon > 0$ there is a $C > 0$ such that

$$|x-r|>\frac{C}{q^{2+\epsilon}}$$

for all $r=\frac{p}{q}\in\mathbb{Q}$.

This can be used to construct other transcendental numbers (cf. Exercise ()).

III.8.3. Measure Theory of Continued Fractions

In this section, we develop some measure-theoretic properties of continued fraction expansions, and the asymptotic behavior of the terms and convergents.

First, given $n, k \in \mathbb{N}$, what is the probability that, for a randomly chosen x, we have $a_n(x) = k$? To make this precise, let

$$E_n^{(k)} = \{x \in (0,1) : a_n(x) = k\}$$

and let $P_n^{(k)}$ be the Lebesgue measure of $E_n^{(k)}$. Then $P_n^{(k)}$ is a reasonable interpretation of the "probability" that $a_n(x) = k$ for a randomly chosen x.

III.8.3.40. The case $n = 1$ is the easiest to analyze. We have

$$E_1^{(k)} = \left(\frac{1}{k+1}, \frac{1}{k}\right]$$

and so

$$P_1^{(k)} = \frac{1}{k(k+1)} = \frac{1}{k^2+k} \ .$$

Thus with high probability $a_1(x)$ is small for a randomly chosen x: in fact, the probability that $a_1(x) \geq k$ is the Lebesgue measure of $(0, 1/k] = 1/k$.

It is not easy to give such a precise description of $P_n^{(k)}$ for $n > 1$. But in general, we have the following estimate:

III.8.3.41. THEOREM. For any n and k,

$$\frac{1}{3k^2} < P_n^{(k)} < \frac{2}{k^2} \ .$$

Note that the bounds are independent of n.

PROOF: The case $n = 1$ follows from the previous calculation. Let $n > 1$. Fix $k_1, \ldots, k_{n-1} \in \mathbb{N}$, and let

$$J(k_1, \ldots, k_{n-1}) = \{x \in (0,1] : a_j(x) = k_j \text{ for } 1 \leq j \leq n-1\}.$$

Then we have () that $J(k_1, \ldots, k_{n-1})$ is the interval with endpoints $\frac{p_{n-1}}{q_{n-1}}$ and $\frac{p_{n-1}+p_{n-2}}{q_{n-1}+q_{n-2}}$, where p_j and q_j are the numerator and denominator of the j'th convergent to the finite continued fraction $[0; k_1, \ldots, k_{n-1}]$. Thus we have that the length of $J(k_1, \ldots, k_{n-1})$ is

$$\left|\frac{p_{n-1}}{q_{n-1}} - \frac{p_{n-1}+p_{n-2}}{q_{n-1}+q_{n-2}}\right| = \frac{1}{q_{n-1}(q_{n-1}-q_{n-2})} = \frac{1}{q_{n-1}^2\left(1+\frac{q_{n-2}}{q_{n-1}}\right)} \ .$$

Similarly, if $k \in \mathbb{N}$ is fixed, the endpoints of $J(k_1, \ldots, k_{n-1}, k)$ are $\frac{p_{n-1}k+p_{n-2}}{q_{n-1}k+q_{n-2}}$ and $\frac{p_{n-1}(k+1)+p_{n-2}}{q_{n-1}(k+1)+q_{n-2}}$, and its length is

$$\left|\frac{p_{n-1}k+p_{n-2}}{q_{n-1}k+q_{n-2}} - \frac{p_{n-1}(k+1)+p_{n-2}}{q_{n-1}(k+1)+q_{n-2}}\right| = \frac{1}{(q_{n-1}k+q_{n-2})(q_{n-1}(k+1)+q_{n-2})}$$

$$= \frac{1}{q_{n-1}^2 k^2 \left(1 + \frac{q_{n-2}}{kq_{n-1}}\right)\left(1 + \frac{1}{k} + \frac{q_{n-2}}{kq_{n-1}}\right)}$$

and so

$$\frac{\lambda(J(k_1, \ldots, k_n, k))}{\lambda(J(k_1, \ldots, k_n))} = \frac{1}{k^2} \frac{1 + \frac{q_{n-2}}{q_{n-1}}}{1 + \frac{q_{n-2}}{kq_{n-1}}} \frac{1}{1 + \frac{1}{k} + \frac{q_{n-2}}{kq_{n-1}}}$$

and we have

$$1 \leq \frac{1 + \frac{q_{n-2}}{q_{n-1}}}{1 + \frac{q_{n-2}}{kq_{n-1}}} < 2$$

$$1 < 1 + \frac{1}{k} + \frac{q_{n-2}}{kq_{n-1}} < 3$$

so we conclude that

$$\frac{1}{3k^2} < \frac{\lambda(J(k_1, \ldots, k_{n-1}, k))}{\lambda(J(k_1, \ldots, k_{n-1}))} < \frac{2}{k^2}$$

$$\frac{1}{3k^2}\lambda(J(k_1, \ldots, k_{n-1})) < \lambda(J(k_1, \ldots, k_{n-1}, k)) < \frac{2}{k^2}\lambda(J(k_1, \ldots, k_{n-1}))\ .$$

We have that, as $(k_1, \ldots, k_{n-1})$ varies over $\mathbb{N}^{n-1}$, $E_n^{(k)}$ is the disjoint union of the $J(k_1, \ldots, k_{n-1}, k)$ and $(0, 1]$ is the disjoint union of the $J(k_1, \ldots, k_{n-1})$; adding the inequalities over all $(k_1, \ldots, k_{n-1})$, we obtain the desired result. 👍

It is interesting to note the following example. Let D be the set of real numbers x such that there exists a strictly increasing sequence (r_n) in $\mathbb{N}$ for which $a_{r_n}(x)$ divides $a_{r_{n+1}}(x)$ for all n. Then D is a large subset of $\mathbb{R}$ (it contains all x which have infinitely many 1's in its continued fraction expansion, and hence $\mathbb{R} \setminus D$ has Lebesgue measure 0).

III.8.3.42. THEOREM. *D is not a Borel set in $\mathbb{R}$.*

In fact, D is an analytic subset of $\mathbb{R}$. $\mathbb{R} \setminus D$ is not analytic. These are perhaps the simplest such subsets of $\mathbb{R}$ to explicitly describe.

III.8.4. Exercises

III.8.4.1. Prove III.8.2.4. by induction on n. [Hint: apply the inductive hypothesis to the denominator of the fraction.]

III.8.4.2. Let φ be the "golden ratio" $\frac{1+\sqrt{5}}{2}$ (often called τ in the mathematical literature). This φ is the unique positive number such that $\varphi^{-1} = \varphi - 1$ (or, equivalently, $\varphi^2 - \varphi = 1$, the positive root of $p(x) = x^2 - x - 1$).

(a) Prove that the continued fraction expansion of φ is $[1; 1, 1, \dots]$, and relate it to the Fibonacci sequence $(1, 1, 2, 3, 5, 8, \dots)$, where each term is the sum of the two previous terms. This is the continued fraction expansion which converges most slowly.

(b) Show that the area of a regular pentagon with side 1 is $\dfrac{5\varphi}{4\sqrt{3-\varphi}}$.

(c) Show that the volume of a regular dodecahedron with edge 1 is $\dfrac{5\varphi^3}{2(3-\varphi)}$.

(d) Show that the volume of a regular icosahedron with edge 1 is $\dfrac{5\varphi^5}{6}$.

The golden ratio is a number which has been of considerable mathematical and cultural significance. Parts (b)–(d) barely scratch the surface of the appearances throughout mathematics. See [Liv02] for a cultural history of φ.

III.8.4.3. (a) Let $d \in \mathbb{N}$, not a perfect square. Find the continued fraction expansion of $\sqrt{d}$. Relate the expansion to the integral solutions of *Pell's equation* [the name is a misnomer, since Pell made little if any contribution to the theory of this equation; the basic work is due to Fermat]

$$x^2 - dy^2 = \pm 1 \ .$$

(b) Let r_n be the n'th continued fraction approximation to $\sqrt{2}$. If Newton's approximation to the positive root of $f(x) = x^2 - 2$ is used with initial approximation r_n, i.e.

$$x_{m+1} = \frac{1}{2}\left(x_m + \frac{2}{x_m}\right)$$

with $x_0 = r_n$, show that the next approximation x_1 is r_{2n+1}. Find the relation between continued fractions and Newton's method for $\sqrt{d}$, $d > 2$ [Duj01], [Mik54], [CMR95], [Ele97].

III.8.4.4. (a) Let α be a quadratic number, i.e. an irrational root of a quadratic polynomial $ax^2 + bx + c$ with $a, b, c \in \mathbb{Z}$. Prove that the continued fraction expansion of α is eventually periodic.
(b) Conversely, suppose β is an irrational number and the continued fraction expansion of β is eventually periodic. Prove that β is a quadratic number.

(Part (a) is an involved problem, and it is unreasonable for the reader to solve it without referring to number theory texts. Part (b) is fairly easy. The two parts were respectively theorems of Lagrange and Euler, two of the most outstanding mathematicians of the 18th century.)

III.8.4.5. Show that the continued fraction expansion of e is

$$[2; 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, \dots, 1, 2k, 1, \dots] .$$

[First find the continued fraction expansion of $\frac{e+1}{e-1}$.] This expansion was found by EULER.

In contrast, the first few terms of the continued fraction expansion of π are

$$[3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, 2, 1, 84, 2, 1, 1, 15, 3, 13, 1,$$

$$4, 2, 6, 6, 99, 1, 2, 2, 6, 3, 5, 1, 1, 6, 8, 1, 7, 1, 2, 3, 7, 1, 2, 1, 1, 12, 1, 1, 1, 3, 1, 1,$$

$$8, 1, 1, 2, 1, 6, 1, 1, 5, 2, 2, 3, 1, 2, 4, 4, 16, 1, 161, 45, 1, 22, 1, 2, 2, 1, 4, 1, 2, \dots]$$

and no pattern is known (although there are other regular nonsimple continued fraction representations of π). Millions of terms of this expansion can be found on the internet.

III.8.4.6. (a) Write out the details of the proof that every Liouville number is transcendental.
(b) Prove that the number $\sum_{k=1}^{\infty} 10^{-k!}$ is a Liouville number.
(c) Prove that the number with continued fraction expansion $[0; 10, 100, \dots, 10^{n!}, \dots]$ is a Liouville number.

III.8.4.7. Let φ be the golden ratio, and F_n the n'th Fibonacci number (Exercise ()). Since $\lim_{n\to\infty} \frac{F_{n+1}}{F_n} = \varphi$, the F_n grow exponentially: if b is any number with $0 < b < \varphi$, then $F_n > b^n$ for all sufficiently large n (and if $c > \varphi$, $F_n < c^n$ for sufficiently large n).
(a) If x is any irrational number, show that $q_n(x) \geq F_n$ for all n. [Prove by induction, using III.8.2.15.(i).] Thus $q_n(x)$ grows at least exponentially. (The inequality $q_n(x) \geq F_n$ also applies for x rational whenever $q_n(x)$ is defined.)
(b) Show by induction on k and III.8.2.15.(i) that if x is irrational and $n \geq 2$, then, for all k,

$$\frac{q_{n+k}(x)}{q_n(x)} > F_{k+1}$$

and that F_{k+1} is the largest possible universal constant for the right-hand side. [Note that $q_n(x)$ can be arbitrarily large compared to $q_{n-1}(x)$ if $a_n(x)$ is large.]

III.8.4.8. (a) Prove Proposition III.8.2.21. by induction, using III.8.2.15.(ii) and the relation $\alpha_n(x) = a_n(x) + \frac{1}{\alpha_{n+1}(x)}$.
(b) Prove the left-hand inequality in III.8.2.22., using $a_n \leq \alpha_n < a_n + 1$.

III.8.4.9. (a) Using III.8.2.22., show that for an irrational number x, the following are equivalent:

(i) x is a Liouville number.

(ii) For each $k \in \mathbb{N}$, $q_{n+1}(x) \geq q_n(x)^k$ for infinitely many n.

(iii) For each $k \in \mathbb{N}$, $a_{n+1}(x) \geq q_n(x)^k$ for infinitely many n.

(b) A Liouville number x can have many small $a_n(x)$'s. In fact, we can have the lim inf of the geometric means of the $a_n(x)$ equal to 1 if many 1's are interspersed between the a_{n+1}'s of (iii). But (iii) implies that the lim sup of the geometric means of the $a_n(x)$ is infinite. Thus no Liouville numbers are in the set for which the conclusion of III.8.2.30. holds.

III.8.4.10. *Champernowne's constant* is obtained by concatenating all positive integers into a single decimal:

$$c = 0.12345678910111213 14\cdots$$

The continued fraction expansion is

$$[0; 8, 9, 1, 149083, 1, 1, 1, 4, 1, 1, 1, 3, 4, 1, 1, 1, 15, a_{18}, 6, 1, 1, 21, 1, \dots]$$

where a_{18} is a 166 digit number. This is only the beginning: $a_{40}(c)$ has 2504 digits, and $a_{34062}(c)$ (the largest one calculated so far) has 7,311,111,092 digits! See http://mathworld.wolfram.com/ChampernowneConstant.html for more information.

Despite the apparent rapid growth of the $a_n(c)$, c is not a Liouville number, although it is transcendental [**?**].

III.8.4.11. This problem gives an analog of () for the growth of the q_n. Let $\psi : \mathbb{N} \to [1, \infty)$ be a function.
(a) Show that if $\sum_{q=1}^{\infty} \frac{1}{\psi(q)}$ converges, then for almost all x we have that the inequality

$$\left| x - \frac{p}{q} \right| < \frac{1}{q\psi(q)}$$

has only finitely many solutions. [We may assume $0 < x < 1$. Fix N. For each p and q, $q \geq N$, $1 \leq p \leq q$, let $I_{p,q}$ be an interval of length $\frac{2}{q\psi(q)}$ centered at $\frac{p}{q}$. Let $S_N = \cup_{p,q} I_{p,q}$. Then $\lambda(S_N) \leq \sum_{q=N}^{\infty} \frac{1}{\psi(q)}$. If the inequality has infinitely many solutions for x, x is in $\cap_N S_N$.]
(b) [] If $\frac{\psi(q)}{q}$ increases and $\sum_{q=1}^{\infty} \psi(q)$ diverges, then for almost all x the inequality

$$\left| x - \frac{p}{q} \right| < \frac{1}{q\psi(q)}$$

has infinitely many solutions.

III.8.4.12. Consider a_n as a random variable on the interval $(0, 1)$.
(a) Find the conditional probability $P(a_2 = k | a_1 = j)$ [The set

$$\{x \in (0, 1) : a_1(x) = j, a_2(x) = k\}$$

is an interval.]
(b) Compute $P_2^{(k)} = P(a_2 = k)$.
(c) Show that a_m and a_n are not independent for any m and n.
(d) Using III.8.3.41., show that there are independent random variables (b_n) and positive constants α and β such that

$$\alpha < \frac{a_n(x)}{b_n(x)} < \beta \text{ a.e.}$$

III.8.4.13. ([?], [Wag93, p. 75-76]; cf. [?]) For each $t > 0$, define

$$\nu_t = \sum_{n=1}^{\infty} 2^{\left(2^{\lfloor nt \rfloor} - 2^{n^2}\right)} .$$

The ν_t are called the *von Neumann numbers.*

(a) Show that the series converges for all $t > 0$.

(b) Show that each ν_t is a Liouville number, and in particular transcendental.

(c) Show that ν_t is not algebraic over the subfield F_t of $\mathbb{R}$ generated by $\{\nu_s : s < t\}$, i.e. ν_t is not a root of a nonzero polynomial with coefficients in F_t. In other words, $\{\nu_t : t > 0\}$ is an algebraically independent set of real numbers of cardinality $2^{\aleph_0}$. [Show that ν_t can be approximated much more closely by rationals than ν_s for $s < t$.] Note that the Axiom of Choice is not needed to construct this set of algebraically independent numbers or to prove its properties.

III.8.4.14. Let f be a strictly increasing function from $\mathbb{N}$ to $\mathbb{N}$, and let $\mu = \limsup_{n\to\infty} \dfrac{f(n+1)}{f(n)}$. Let $d \in \mathbb{N}$, $d > 1$, and set $x = \displaystyle\sum_{n=1}^{\infty} \frac{1}{d^{f(n)}}$.

(a) Show that the equation

$$|x - r| < \frac{C}{q^{\mu-\epsilon}}$$

has infinitely many solutions in $r = \frac{p}{q} \in \mathbb{Q}$ for any fixed $C > 0$ and $\epsilon > 0$.

(b) Use Roth's Theorem (III.8.2.39.) to show that x is transcendental if $\mu > 2$.

III.8.4.15. (a) If $x = \frac{p}{q}$ is a rational number, show that the decimal expansion of x is eventually periodic, with the length of the period not more than $q - 1$. [Divide q into p by long division.]

(b) Conversely, if $x \in \mathbb{R}$ has an eventually periodic decimal expansion, then x is rational. [If the length of the period of the expansion is k, show that $10^k x - x$ is a decadent rational.]

If $0 \le x \le 1$ and the decimal expansion of x is exactly periodic, a fractional representation of x is easy to obtain. For example,

$$x = .250425042504\cdots = \frac{2504}{9999}$$

since $10000x - x = 2504$, and the same pattern works in general, with k 9's in the denominator if the period is k. (This representation will not be in lowest terms in general.)

(c) What is the exact relationship between the length of the period of the decimal expansion of $\frac{p}{q}$ and the denominator q?

(d) Show that analogous results hold for "decimal" expansions with any base.

III.8.4.16. Show that if a and b are positive numbers (e.g. natural numbers), then

$$\sqrt{a^2 + b} = a + \cfrac{b}{2a + \cfrac{b}{2a + \cfrac{b}{2a + \cfrac{b}{\cdots}}}} .$$

III.8.4.17. For $x \in \mathbb{J}$, let $a_n(x)$ be the n'th term in the continued fraction expansion of x. Let

$$C = \left\{ x \in \mathbb{J} \cap [0,1] : \sum_{n=1}^{\infty} \frac{1}{a_n(x)} < \infty \right\} .$$

Is C Lebesgue measurable? Is it a Borel set? What is the Lebesgue (outer) measure of C? (cf. III.8.2.29., IV.2.16.4..)

III.8.4.18. Fix $n \in \mathbb{N}$. Find the integers p, q, r, s with $0 \leq p, q, r, s \leq n$ such that

$$\pm\frac{p}{q} \pm \sqrt{\frac{r}{s}}$$

most closely approximates π. Is there a better way to solve this than brute-force calculation, by a method analogous to the continued-fraction expansion? (The method should work not just for π, but equally well for any irrational number which is not quadratic.) For $n = 100$, the answer (found by brute force) is

$$\frac{71}{28} + \sqrt{\frac{29}{79}}$$

which is accurate within 10^{-8} (https://mathoverflow.net/questions/351400/approximations-to-pi). The number

$$4 - \sqrt{\frac{14}{19}}$$

found by hand is accurate to about 1.2×10^{-5}. And there is always $\sqrt{10}$, which is off by about .02 (but not quite as good as $\left(\frac{4}{3}\right)^4$, cf. Figure I.13, and not nearly as good as $\frac{22}{7}$).

III.9. Irrational and Transcendental Numbers

> "It can be of no practical use to know that π is irrational, but if we can know, it surely would be intolerable not to know."
>
> *E. C. Titchmarsh*[7]

III.9.1. Introduction

III.9.1.1. Analytic methods can be used to prove that specific numbers such as e and π are irrational and even transcendental. In this section, we give a few proofs not requiring anything beyond basic analysis techniques. More sophisticated results can be obtained by combining such methods with ideas from algebraic number theory; we survey some of these results, mostly without proof.

III.9.1.2. It was known as a result of Euler's work on continued fractions (III.8.4.5.) that e (and also e^2) is irrational, and in fact, using Lagrange's Theorem (III.8.4.4.), not algebraic of degree 2. Another proof based on the power series representation was given by Fourier in 1815 (cf. V.5.5.6.; this proof can be easily modified to also give the quadratic result.)

III.9.1.3. The first proof that π is irrational was given by Lambert in 1766, using continued fractions (his proof was not rigorous by modern standards). Legendre gave a rigorous proof along the same lines in 1794 for π and many other numbers. Hermite developed another method in 1873 giving the transcendence of e, which was adapted by Lindemann to give broad transcendence results including transcendence of π. Significant advances beginning with the same general method were obtained throughout the twentieth century. See, for example, [Niv56] or the introduction to [Shi89] for a more complete survey.

III.9.2. Irrationality of π

III.9.2.1. We first give a simple proof of the irrationality of π (and π^2) along the lines of the arguments of Hermite and Lindemann, due to I. Niven [Niv47] and Y. Iwamoto [Iwa49] (cf. [Wac49], [Niv56, 2.6]), which will introduce the techniques used in later proofs of transcendence, even though the result will be subsumed in III.9.3.2.. The idea in all these proofs is to assume that a certain rational or polynomial relation exists, and obtain a contradiction by constructing an integral whose value is an integer strictly between 0 and 1. The functions used in the arguments can seem to materialize out of thin air, but actually arise from considerations related to continued fraction expansions; see [Sie49, I, §3], for example, for motivation.

We first make the following simple observation:

III.9.2.2. Lemma. Let $g(x)$ be a polynomial with integer coefficients, and $n \in \mathbb{N}$. Set $f(x) = \frac{x^n g(x)}{n!}$. Then the k'th derivative $f^{(k)}(0)$ is an integer for every k.

Proof: Let $x^n g(x) = \sum_{k=0}^{n+d} c_k x^k$. Then $c_k \in \mathbb{Z}$ for all k, and $c_k = 0$ for $k < n$. We have, for all k,

$$f^{(k)}(0) = \frac{k! c_k}{n!} \in \mathbb{Z} .$$

[7] In N. Rose, *Mathematical Maxims and Minims*, Rome Press Inc., Raleigh, NC, 1988.

III.9.2.3. THEOREM. The number π^2 is irrational. (Hence π is also irrational.)

PROOF: Let n be a fixed positive integer (which we will choose later), and set

$$f(x) = \frac{x^n(1-x)^n}{n!} \ .$$

Note that if $0 < x < 1$, $0 < f(x) < \frac{1}{n!}$. By III.9.2.2., $f^{(k)}(0) \in \mathbb{Z}$ for all k; and $f^{(k)} \equiv 0$ for $k > 2n$. Since $f(1-x) = f(x)$ for all x, we also have $f^{(k)}(1) \in \mathbb{Z}$ for all k.

Suppose $\pi^2 = \frac{a}{b}$ for relatively prime $a, b \in \mathbb{N}$. Set

$$F(x) = b^n[\pi^{2n}f(x) - \pi^{2n-2}f^{(2)}(x) + \pi^{2n-4}f^{(4)}(x) - \cdots + (-1)^n f^{(2n)}(x)]$$

and note that $F(0), F(1) \in \mathbb{Z}$. Next,

$$\frac{d}{dx}\left(F'(x)\sin\pi x - \pi F(x)\cos\pi x\right) = \left(F''(x) + \pi^2 F(x)\right)\sin\pi x$$

$$= b^n\pi^{2n+2}f(x)\sin\pi x = \pi^2 a^n f(x)\sin\pi x$$

so we have

$$\pi a^n \int_0^1 f(x)\sin\pi x\,dx = \left[\frac{F'(x)\sin\pi x}{\pi} - F(x)\cos\pi x\right]_0^1 = F(1) + F(0) \in \mathbb{Z} \ .$$

On the other hand, since $0 < f(x)\sin\pi x < \frac{1}{n!}$ for $0 < x < 1$,

$$0 < \pi a^n \int_0^1 f(x)\sin\pi x\,dx < \frac{\pi a^n}{n!} < 1$$

for sufficiently large n, a contradiction.

A nearly identical proof ([Kok49], [Wac49]) shows that e^n is irrational for all n (see III.9.3.1. for a better result). A modified version of this argument, somewhat more complicated but still elementary (cf. [Niv56, 2.5]), gives:

III.9.2.4. THEOREM. If α is any nonzero rational number, then $\cos\alpha$ is irrational.

III.9.2.5. This theorem gives another proof that π is irrational, since otherwise $\cos\pi = -1$ would be irrational. It also implies that $\sin^2\alpha = \frac{1}{2}(1-\cos 2\alpha)$ and hence $\sin\alpha$ and $\cos^2\alpha = 1-\sin^2\alpha$ are irrational, for nonzero rational α. In addition, for α nonzero rational, $\tan^2\alpha$ is irrational, since otherwise $\cos^2\alpha = \frac{1}{1+\tan^2\alpha}$ would be rational, and hence $\tan\alpha$ is irrational. Thus, if α is nonzero rational, $\arcsin\alpha$ and $\arctan\alpha$ are irrational. These results will be subsumed by the much more general result III.9.3.6.(iii).

III.9.3. Transcendence of e and Related Numbers

A proof along the same general lines can be given of the transcendence of e, due to A. Hurwitz (1893):

III.9.3.1. Theorem. The number e is transcendental.

Proof: Suppose e is algebraic. Then there are integers $a_0, a_1, \ldots, a_m$ with $a_0 \neq 0$ and

$$a_0 + a_1 e + a_2 e^2 + \cdots + a_m e^m = 0 \ .$$

If $n \in \mathbb{N}$ and f is any polynomial of degree n with real coefficients, and we set

$$F(x) = f(x) + f'(x) + f''(x) + \cdots + f^{(n)}(x) \ ,$$

then by repeated integration by parts we obtain, for any $b > 0$,

$$\int_0^b f(t)e^{-t}\,dt = f(0) - f(b)e^{-b} + \int_0^b f'(t)e^{-t}\,dt = \cdots = F(0) - F(b)e^{-b}$$

called *Hermite's Identity*.

We will apply Hermite's Identity in the case where

$$f(x) = \frac{1}{(p-1)!}x^{p-1}(x-1)^p(x-2)^p\cdots(x-m)^p$$

where m is the degree of the polynomial relation for e given above and p is a large prime number which will be chosen later (recall that there are infinitely many prime numbers, so p can be chosen as large as desired).

Since f is a polynomial of degree $n = p(m+1)-1$, we may apply Hermite's Identity for $b = 1, \ldots, k, \ldots, m$, we obtain

$$F(0)e^k - F(k) = e^k \int_0^k f(t)e^{-t}\,dt \ .$$

The equation also holds when $k = 0$ (both sides are 0). Multiplying by a_k and adding, we get the fundamental relation

$$\sum_{k=0}^m a_k e^k F(0) - \sum_{k=0}^m a_k F(k) = \sum_{k=0}^m a_k e^k \int_0^k f(t)e^{-t}\,dt \ . \tag{III.1}$$

The first sum is 0 by the assumed polynomial relation for e, so the left side becomes

$$-\sum_{k=0}^m a_k F(k) = -\sum_{k=0}^m \sum_{r=0}^n a_k f^{(r)}(k) \ .$$

If g is a polynomial with integer coefficients, then for any r we have that all coefficients of $g^{(r)}$ are divisible by $r!$ (V.12.5.6.). So, if $r \geq p$, all the coefficients of $f^{(r)}$ are integers divisible by p, and hence for $r \geq p$ we have that $f^{(r)}(k)$ is an integer divisible by p for any $k \in \mathbb{Z}$ (and, in particular, for $k = 0, 1, \ldots, m$). Also, for $k = 1, \ldots, m$, f has a zero of order p at k, so $f^{(r)}(k) = 0$ for $0 \leq r < p$. Thus $f^{(r)}(k)$ is an integer divisible by p for all r.

Since f has a zero of order $p-1$ at 0, we also have $f^{(r)}(0) = 0$ for $0 \le r < p-1$. We can calculate directly from the definition that

$$f^{(p-1)}(0) = (-1)^p(-2)^p \cdots (-m)^p = (-1)^{mp}(m!)^p$$

which is an integer not divisible by p if $p > m$.

So every term on the left is an integer divisible by p except the term $a_0 f^{(p-1)}(0)$, which is an integer not divisible by p if $p > m$ and $p > |a_0|$ (recall that $a_0 \neq 0$). So if p is sufficiently large, the left side of III.1 is a nonzero integer.

Now consider the right side of III.1. For $0 \le x \le m$, we have

$$|f(x)| \le \frac{m^{p-1}m^p m^p \cdots m^p}{(p-1)!} = \frac{m^{mp+p-1}}{(p-1)!} = \frac{(m^{m+1})^p}{m(p-1)!} \ .$$

So, if we set

$$c = e^m \sum_{k=0}^{m} |a_k| \text{ and } d = m^{m+1}$$

(note that c and d do not depend on p), we get for the right side

$$\left| \sum_{k=0}^{m} a_k e^k \int_0^k f(t)e^{-t}\,dt \right| \le \sum_{k=0}^{m} \left| a_k e^k m \frac{(m^{m+1})^p}{m(p-1)!} \right| \le c\frac{d^p}{(p-1)!}$$

which is strictly less than 1 for p sufficiently large, contradicting that the left side is a nonzero integer. ☝

III.9.3.2. A proof that π is transcendental can be given along the same lines, but it is slightly more complicated and uses some elementary results from number theory; see, for example, [Bak90, 1.3] or [Shi89, p. 47-50].

Using a refined version of the same argument, plus some facts about algebraic numbers, Lindemann in his 1882 paper outlined a proof of a much more general theorem, for which Weierstrass later provided a complete and simplified proof (Hermite gave a proof of III.9.3.5.):

III.9.3.3. Theorem. Let $\alpha_1, \ldots, \alpha_n$ be distinct algebraic numbers, and $\beta_1, \ldots, \beta_n$ nonzero algebraic numbers. Then

$$\beta_1 e^{\alpha_1} + \cdots + \beta_n e^{\alpha_n} \neq 0 \ .$$

This result has the following equivalent formulation, which is often called Lindemann's Theorem:

III.9.3.4. THEOREM. Let $\alpha_1, \ldots, \alpha_n$ be algebraic numbers which are linearly independent over $\mathbb{Q}$. Then $e^{\alpha_1}, \ldots, e^{\alpha_n}$ are *algebraically independent* over $\mathbb{Q}$, i.e. there does not exist a nonzero polynomial $f(x_1, \ldots, x_n)$ with rational coefficients (or even with algebraic coefficients) with $f(e^{\alpha_1}, \ldots, e^{\alpha_n}) = 0$.

It is a simple standard fact from field theory (cf. Exercise III.9.5.2.) that if $\gamma_1, \ldots, \gamma_n$ are complex numbers which are algebraically independent over $\mathbb{Q}$, then they are algebraically independent over the field of algebraic numbers (the algebraic closure of $\mathbb{Q}$).

To obtain III.9.3.4. from III.9.3.3., let $\alpha_1, \ldots, \alpha_n$ be algebraic numbers which are linearly independent over $\mathbb{Q}$. If $f(x_1, \ldots, x_n)$ is a nonzero polynomial with rational coefficients, then f is a sum of terms of the form $c_{(k_1,\ldots,k_n)} x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}$, where the k_i are nonnegative integers. So $f(e^{\alpha_1}, \ldots, e^{\alpha_n})$ is a sum of terms of the form

$$c_{(k_1,\ldots,k_n)} e^{k_1\alpha_1} e^{k_2\alpha_2} \cdots e^{k_n\alpha_n} = c_{(k_1,\ldots,k_n)} e^{k_1\alpha_1 + \cdots + k_n\alpha_n}$$

and the exponents are distinct integer linear combinations of $\alpha_1, \ldots, \alpha_n$, hence distinct algebraic numbers by linear independence. Thus by III.9.3.3., $f(e^{\alpha_1}, \ldots, e^{\alpha_n}) \neq 0$.

Conversely, suppose III.9.3.4. holds, and let $\alpha_1, \ldots, \alpha_n$ be distinct algebraic numbers. If the rational span of $\{\alpha_1, \ldots, \alpha_n\}$ has dimension m over $\mathbb{Q}$, then there are algebraic numbers $\gamma_1, \ldots, \gamma_m$ which are linearly independent over $\mathbb{Q}$ such that, for each k, $1 \leq k \leq n$,

$$\alpha_k = \sum_{j=1}^{m} a_{jk} \gamma_j$$

with $a_{jk} \in \mathbb{Q}$; we may assume each $a_{jk} \in \mathbb{Z}$ by replacing γ_j by γ_j / r_j for suitable $r_j \in \mathbb{Z}$. Now let $\beta_1, \ldots, \beta_n$ be nonzero algebraic numbers, and set

$$f(x_1, \ldots, x_n) = \sum_{k=1}^{n} \beta_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_m^{a_{mk}} \ .$$

Then f is a nonzero rational function with algebraic coefficients (the monomials in the terms are all different since the α_k are distinct), hence a quotient of polynomials with algebraic coefficients (the denominator can be taken to be a monomial), so $f(e^{\gamma_1}, \ldots, e^{\gamma_m}) \neq 0$ by III.9.3.4.. But

$$f(e^{\gamma_1}, \ldots, e^{\gamma_m}) = \sum_{k=1}^{n} \beta_k e^{a_{1k}\gamma_1 + \cdots + a_{mk}\gamma_m} = \sum_{k=1}^{n} \beta_k e^{\alpha_k} \neq 0 \ .$$

The special case of III.9.3.4. with $n = 1$ is:

III.9.3.5. COROLLARY. If α is a nonzero algebraic number, then e^{α} is transcendental.

In these theorems, the α_k and β_k do not have to be real numbers (i.e. they can be complex algebraic numbers).

III.9.3.6. Here are some consequences of III.9.3.5.:

(i) This result also implies that π is transcendental, since if π, and hence πi, were algebraic, then $e^{\pi i} = -1$ would be transcendental.

(ii) If α is a positive algebraic number, $\alpha \neq 1$, then $\log \alpha$ is transcendental, since if it were algebraic we would have $\alpha = e^{\log \alpha}$ transcendental.

(iii) If α is a nonzero real algebraic number, then $e^{i\alpha} = \cos \alpha + i \sin \alpha$ is transcendental. If $\cos \alpha$ were algebraic, $\sin \alpha = \pm\sqrt{1 - \cos^2 \alpha}$ would also be algebraic, so $e^{i\alpha}$ would be algebraic. Thus $\cos \alpha$ and $\sin \alpha$ are transcendental. We then also have $\tan \alpha$, $\arcsin \alpha$, $\arctan \alpha$ transcendental by the argument of III.9.2.5..

III.9.4. Twentieth Century Results

III.9.4.1. In his famous 1900 list of problems (#7), Hilbert asked to prove that if α is an algebraic number other than 0 or 1, and β is an irrational algebraic number, then α^β is transcendental. In a 1920 lecture he expressed the opinion that both the Riemann Hypothesis and Fermat's Last Theorem would be solved first, and that this problem would not be solved within the lifetime of anyone in the room. Actually, the problem was solved independently in 1934 by Gelfond and Schneider, well within Hilbert's own lifetime, using key ideas of Gelfond (1929) and Siegel (1929). (Fermat's Last Theorem was proved in 1995, and the Riemann Hypothesis remains unproved, showing the difficulty in trying to predict the future of mathematics.)

III.9.4.2. Theorem. [Gelfond-Schneider] Let α and β be algebraic complex numbers, with α not 0 or 1 and $\beta \notin \mathbb{Q}$. Then α^β is transcendental.

Actually, in general α^β has many possible values depending on the branch of logarithm chosen; the theorem asserts that all these possible values are transcendental. See [Niv56, Chapter 10] or [Lan02, Appendix] for a proof, or III.9.5.10. for an outline.

III.9.4.3. For example, numbers such as $2^{\sqrt{2}}$ and $\sqrt{2}^{\sqrt{3}}$ are transcendental (these were settled by Kuzmin in 1930). Also, we obtain that e^π is transcendental, since it is one of the values of i^{-2i}. Similarly, $e^{\pi\alpha}$ is transcendental for any nonzero real algebraic α.

An especially interesting example is the "Ramanujan Constant" (actually discovered by Hermite in 1859; the name stems from an April Fool joke by Martin Gardner in 1975. See http://en.wikipedia.org/wiki/Ramanujan_constant#Almost_integers_and_Ramanujan.27s_constant or http://mathworld.wolfram.com/RamanujanConstant.html for more information.) Calculation shows that

$$e^{\pi\sqrt{163}} = 262537412640768744.00000000000$$

correct to eleven decimal places; nonetheless, this number is not an integer, and in fact is transcendental. Powers of this number are also remarkably close to integers. The underlying reason is that 163, which is prime, is the largest squarefree integer d such that the ring of integers in the field $\mathbb{Q}(\sqrt{-d})$ has unique factorization.

We also have the following corollary, which was stated but not proved by Euler:

III.9.4.4. Corollary. Let α and β be positive algebraic numbers, with $\alpha \neq 1$. If β is not a rational power of α (i.e. if $\log_\alpha \beta$ is irrational; cf. III.9.5.3.), then $\log_\alpha \beta$ is transcendental.

Proof: If $\gamma = \log_\alpha \beta$ is irrational but algebraic, then $\beta = \alpha^\gamma$ is transcendental, a contradiction. 👍

A significant advance arose from the fundamental work of A. Baker in the late 1960's. One of his theorems was:

III.9.4.5. Theorem. If $\alpha_1, \ldots, \alpha_n$ are nonzero algebraic numbers such that $\log \alpha_1, \ldots, \log \alpha_n$ (any possible values of these are allowed) are linearly independent over $\mathbb{Q}$, then $1, \log \alpha_1, \ldots, \log \alpha_n$ are linearly independent over the algebraic numbers.

III.9.4.6. The case $n = 2$ implies III.9.4.2.: if α and β are algebraic numbers with α not 0 or 1 (so $\log \alpha$ is defined and nonzero), and α^β is algebraic, set $\alpha_1 = \alpha$, $\beta_1 = \beta$, $\alpha_2 = \alpha^\beta$, and $\beta_2 = -1$. Then $\beta_1 \log \alpha_1 + \beta_2 \log \alpha_2 = 0$, so by III.9.4.5. $\{\log \alpha_1, \log \alpha_2\} = \{\log \alpha, \beta \log \alpha\}$ is linearly dependent over $\mathbb{Q}$, i.e. $\beta \in \mathbb{Q}$. (The converse implication is not quite true: III.9.4.2. implies that if α_1 and α_2 are algebraic numbers with $\{\log \alpha_1, \log \alpha_2\}$ linearly independent over $\mathbb{Q}$, then $\{\log \alpha_1, \log \alpha_2\}$ is linearly independent over the algebraic numbers, but it does not imply that $\{1, \log \alpha_1, \log \alpha_2\}$ is linearly independent over the algebraic numbers.)

III.9.4.7. Corollary.

(i) If $\alpha_1, \ldots, \alpha_n, \beta_0, \ldots, \beta_n$ are nonzero algebraic numbers, then

$$\beta_0 + \beta_1 \log \alpha_1 + \cdots + \beta_n \log \alpha_n \neq 0 \ .$$

(ii) If $\alpha_1, \ldots, \alpha_n, \beta_0, \ldots, \beta_n$ are nonzero algebraic numbers, then

$$e^{\beta_0} \alpha_1^{\beta_1} \cdots \alpha_n^{\beta_n}$$

is transcendental.

(iii) If $\alpha_1, \ldots, \alpha_n$ are algebraic numbers other than 0 or 1, and $\beta_1, \ldots, \beta_n$ are algebraic numbers with $\{1, \beta_1, \ldots, \beta_n\}$ linearly independent over $\mathbb{Q}$, then

$$\alpha_1^{\beta_1} \cdots \alpha_n^{\beta_n}$$

is transcendental.

III.9.4.8. These results imply that $\pi + \log \alpha$ is transcendental for any nonzero algebraic number α, and that $e^{\alpha\pi+\beta}$ is transcendental for any algebraic α and β with $\beta \neq 0$.

III.9.4.9. Siegel (1929) began a general study of the values of certain transcendental functions satisfying linear differential equations, and he and subsequent authors obtained far-reaching generalizations of the results of III.9.3.6.. See [Shi89] for the full story.

III.9.4.10. We have only scratched the surface of the extensive research done in this subject. Books such as () give a much more complete picture. But despite all the advances, there are still basic questions remaining open. For example, it has been difficult to get algebraic independence results for numbers of the form α^β, where α and β are algebraic, analogous to III.9.3.4.. III.9.4.7.(ii)–(iii) are weak analogs, but a better result would be

III.9.4.11. Conjecture. Let $\alpha_1, \ldots, \alpha_n$ be algebraic numbers not 0 or 1, let $\beta_1, \ldots, \beta_n$ be algebraic numbers with $\{1, \beta_1, \ldots, \beta_n\}$ linearly independent over $\mathbb{Q}$, and let $\gamma_1, \ldots, \gamma_m$ be algebraic numbers linearly independent over $\mathbb{Q}$. Then $\{\alpha_1^{\beta_1}, \ldots, \alpha_n^{\beta_n}, e^{\gamma_1}, \ldots, e^{\gamma_m}\}$ is algebraically independent over $\mathbb{Q}$.

The γ_j and e^{γ_j} can be eliminated from this conjecture. Another conjecture, due to Schanuel, which would generalize III.9.3.4. (the case where the α_j are algebraic):

III.9.4.12. Conjecture. Let $\alpha_1, \ldots, \alpha_n$ be complex numbers which are linearly independent over $\mathbb{Q}$. Then the field $\mathbb{Q}(\alpha_1, \ldots, \alpha_n, e^{\alpha_1}, \ldots, e^{\alpha_n})$ has transcendence degree $\geq n$ over $\mathbb{Q}$.

See [Rib00a, p. 323-327] for more related conjectures.

III.9.4.13. As an example of the incomplete state of the theory of transcendental numbers, it is not known whether π and e are algebraically independent (this is apparently universally believed to be true, but remains unproved). It is not even known whether $\pi + e$ and/or πe is irrational! It is known from the following simple proposition that at least one of these numbers is transcendental, but nothing is known about either number individually. It is also not known whether e^e, π^e, π^π, or the Euler-Mascheroni constant γ (IV.2.16.26.) are rational or irrational, algebraic or transcendental (they are likely all transcendental with probability 1).

III.9.4.14. Proposition. Let α and β be numbers.

(i) If α is transcendental, then either $\alpha + \beta$ or $\alpha\beta$ (or both) is transcendental.

(ii) If α is not a quadratic number, then either $\alpha + \beta$ or $\alpha\beta$ (or both) is irrational.

Proof: (i) If $\alpha + \beta$ and $\alpha\beta$ are both algebraic, then α is a root of the quadratic polynomial

$$f(x) = (x - \alpha)(x - \beta) = x^2 - (\alpha + \beta)x + \alpha\beta$$

with algebraic coefficients, so α would be algebraic. The proof of (ii) is a slight variation left to the reader. 👍

III.9.5. Exercises

III.9.5.1. (http://math.stackexchange.com/questions/654232/if-cosa-pi-frac13-then-a-is-irrational?noredirect=1&lq=1)

(a) Let $p, q \in \mathbb{N}$. Then $e^{\pm \frac{\pi i p}{q}}$ is a root of $X^{2q} - 1$, hence an algebraic integer. Thus $2\cos\left(\frac{\pi p}{q}\right) = e^{\frac{\pi i p}{q}} + e^{-\frac{\pi i p}{q}}$ is an algebraic integer.

(b) If $r \in \mathbb{Q}$ and $\cos \pi r \in \mathbb{Q}$, then $2\cos \pi r$ is an integer (cf. III.9.5.5.); hence $|\cos \pi r| = 0, \frac{1}{2}$, or 1. Thus if $0 < r < \frac{1}{2}$, then $r = \frac{1}{3}$.

(c) If $0 < \alpha < \frac{\pi}{2}$ and $\cos \alpha$ is rational, then either $\alpha = \frac{\pi}{3}$ or $\frac{\alpha}{\pi}$ is irrational.

In light of the fact that π is irrational, even transcendental, this result is almost completely complementary to III.9.2.4. or III.9.3.6.(iii). The proof outlined is quite simple, but uses one nontrivial fact from abstract algebra or number theory: that the sum of two algebraic integers is an algebraic integer. A more elementary proof is given in [HD64] using only III.11.10.33. for $n = 2$ and some trigonometric identities; it is also noted there that the result implies that the acute angles in any Pythagorean triangle are irrational multiples of π.

III.9.5.2. This exercise requires knowledge of the basics of field theory in abstract algebra. See books on abstract algebra such as [DF04] for details, and for an explanation of the standard notation and terminology used in this problem.

Suppose K is a field, F a subfield of K, and E a subfield of K containing F.

(a) If E is a finite extension of F and $\alpha \in K$, then $E(\alpha)$ is a finite extension of $F(\alpha)$ and

$$[E(\alpha) : F(\alpha)] \leq [E : F] \ .$$

(b) If E is an algebraic extension of F and $\alpha \in K$, then $E(\alpha)$ is an algebraic extension of $F(\alpha)$.

(c) If E is an algebraic extension of F and $\alpha \in K$ is algebraic over E, then α is algebraic over F.

(d) If $\gamma_1, \ldots, \gamma_n \in K$, then $\{\gamma_1, \ldots, \gamma_n\}$ is algebraically independent over F if and only if γ_1 is transcendental over F and γ_{k+1} is transcendental over $F(\gamma_1, \ldots, \gamma_k)$ for $1 \leq k < n$.

(e) Show inductively using (b), (c), and (d) that if $\{\gamma_1, \ldots, \gamma_n\} \subseteq K$ is algebraically independent over F and E is algebraic over F, then $\{\gamma_1, \ldots, \gamma_n\}$ is algebraically independent over E.

III.9.5.3. (a) Prove that $\log_2 3$ is irrational. [2^p is even and 3^q is odd for all $p, q \in \mathbb{N}$.]

(b) Prove that if m, n are integers > 1, and $\log_m n$ is rational, then m and n are both integer powers of the same natural number.

(c) Give a similar criterion for rationality of $\log_\alpha \beta$ for positive rational numbers α, β with $\alpha \neq 1$.

III.9.5.4. This problem shows that if $n, d \in \mathbb{N}$, then $\sqrt[n]{d}$ is either a natural number or irrational.

(a) Suppose $\sqrt[n]{d} = \frac{a}{b}$ for natural numbers a, b. Then $a^n = b^n d$. Write

$$a = p_1^{r_1} p_2^{r_2} \cdots p_m^{r_m}$$

$$b = p_1^{s_1} p_2^{s_2} \cdots p_m^{s_m}$$

$$d = p_1^{t_1} p_2^{t_2} \cdots p_m^{t_m}$$

where $p_1, \ldots, p_m$ are distinct prime numbers and the exponents are nonnegative integers (some may be zero).
(b) From $a^n = b^n d$ obtain that

$$p_1^{nr_1} p_2^{nr_2} \cdots p_m^{nr_m} = p_1^{ns_1+t_1} p_2^{ns_2+t_2} \cdots p_m^{ns_m+t_m}$$

and conclude that t_k is divisible by n for all k, i.e. that $d = c^n$ for some $c \in \mathbb{N}$.

This argument uses the unique factorization of natural numbers into products of prime numbers (cf. ()), hence for $n = 2$ is not quite as elementary as the one in III.1.12.6.. See III.1.14.7.(e) for an argument not requiring unique factorization.

III.9.5.5. The purpose of this problem is to give a purely algebraic proof of the following theorem:
THEOREM. Let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a monic polynomial with integer coefficients. Then any rational number which is a root of f is an integer (i.e. a rational number which is an algebraic integer is an ordinary integer).

This theorem is a special case of *Gauss's Lemma*. See books on abstract algebra such as [DF04] for the statement and proof of the general form of Gauss's Lemma. (This is only one of several results commonly called "Gauss's Lemma"; he produced a lot of important mathematics.)
If $n, d \in \mathbb{N}$, applying this theorem to $f(x) = x^n - d$, we obtain (see Exercise III.9.5.4. for another proof):
COROLLARY. If $n, d \in \mathbb{N}$, then $\sqrt[n]{d}$ is either a natural number or irrational.

These results are special cases of the more general result (also a special case of Gauss's Lemma):
THEOREM. If $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ is a polynomial with integer coefficients and $a_0 \neq 0$, $a_n \neq 0$, and $\frac{c}{b}$ is a (nonzero) fraction in lowest terms which is a root, then b divides a_n, c divides a_0, and $f(x) = (bx - c)k(x)$ for a polynomial k with coefficients in $\mathbb{Z}$.

(a) Show that if $g(x)$ and $h(x)$ are polynomials with integer coefficients, and p is a prime number that divides every coefficient of $g(x)h(x)$, then either p divides every coefficient of $g(x)$ or it divides every coefficient of $h(x)$. [If not, let r be the largest number for which the coefficient of x^r in g is not divisible by p, s the largest number for which the coefficient of x^s in h is not divisible by p, and consider the coefficient of x^{r+s} in gh.]
(b) If $q = \frac{c}{b} \in \mathbb{Q}$, show that $f(x) = (x - q)g(x)$ for a polynomial g with rational coefficients.
(c) Show there is a $d \in \mathbb{N}$ such that

$$bd \cdot f(x) = (bx - c)h(x)$$

where h is a polynomial with integer coefficients.
(d) If p is a prime number dividing bd, use (a) to show that p divides every coefficient of h.
(e) Dividing out prime factors of bd repeatedly, show that

$$f(x) = (bx - c)k(x)$$

for a polynomial k with integer coefficients. Multiply out to obtain the divisibility conclusion.
(f) In fact, $\mathbb{Z}[X]$ is a *unique factorization domain* or *UFD*: any polynomial with integer coefficients can be written as a product of irreducible polynomials with integer coefficients, uniquely up to the order and signs of the factors.

III.9.5.6. Show that there is a unique real number a such that $e^a = \frac{1}{a}$, and that this number is transcendental. (We have $a = W(1) \approx 0.567$, where W is the *Lambert W function*.)

III.9.5.7. Are the following formulas correct?

$$\pi^4 + \pi^5 = e^6 \ .$$

$$e^\pi - \pi = 20 \ .$$

III.9.5.8. (a) Some other good approximations to π:

$$\sqrt{51} - 4$$

$$\sqrt{\frac{40}{3} - 2\sqrt{3}}$$

$$4 - \sqrt{\frac{14}{19}}$$

$$\frac{71}{28} + \sqrt{\frac{29}{79}}$$

(Y. Baruch; cf. https://mathoverflow.net/questions/351400/approximations-to-pi)
(b) Fix $n \in \mathbb{N}$ (say $n = 100$). Find the integers p, q, r, s with $0 \leq p, q, r, s \leq n$ such that

$$\pm\frac{p}{q} \pm \sqrt{\frac{r}{s}}$$

most closely approximates π. For $n = 100$, the best approximation is apparently $\frac{71}{28} + \sqrt{\frac{29}{79}}$. If $n = 10$, is there a better approximation than $4 - \sqrt{\frac{3}{4}}$? If $n = 20$, is there a better one than $4 - \sqrt{\frac{14}{19}}$?

III.9.5.9. [Sin13] Use a calculator or computer algebra program to calculate

$$\sqrt[12]{3987^{12} + 4365^{12}} \ .$$

Does this provide a counterexample to Fermat's Last Theorem?

III.9.5.10. [Lan02, Appendix] If f is an entire function (), say that f is *of order* $\leq \rho$ if there is a $C > 1$ such that $|f(z)| \leq C^{R^\rho}$ for all sufficiently large R and all $z \in \mathbb{C}$ with $|z| \geq R$, and that a meromorphic function on $\mathbb{C}$ is of order $\leq \rho$ if it is a quotient of two entire functions of order $\leq \rho$.

(a) Prove the following theorem:

Theorem. Let $\mathbb{K}$ be a finite extension of $\mathbb{Q}$. Let $f_1, \ldots, f_n$ be meromorphic functions of degree $\leq \rho$ such that the field $\mathbb{K}(f_1, \ldots, f_n)$ has transcendence degree ≥ 2 over $\mathbb{K}$, and that the derivative of every function in the ring $R = \mathbb{K}[f_1, \ldots, f_n]$ is again in R. Let $w_1, \ldots, w_m$ be complex numbers not lying among the poles of the f_j, such that $f_j(w_k) \in \mathbb{K}$ for $1 \leq j \leq n$, $1 \leq k \leq m$. Then

$$m \leq 10\rho[\mathbb{K} : \mathbb{Q}] \ .$$

(b) Deduce III.9.3.5. as a consequence. [If α and e^α are algebraic, set $\mathbb{K} = \mathbb{Q}(\alpha, e^\alpha)$ and $f_1(z) = z$ and $f_2(z) = e^z$. The functions take algebraic values at α, 2α, $\ldots$, $m\alpha$ for any m, a contradiction.]

(c) Deduce the Gelfond-Schneider Theorem III.9.4.2.. [If α^β is algebraic, set $\mathbb{K} = \mathbb{Q}(\alpha, \beta, \alpha^\beta)$, $f_1(z) = e^z$, $f_2(z) = e^{\beta z}$. The functions take algebraic values at $\log\alpha$, $2\log\alpha$, $\ldots$, $m\log\alpha$ for any m, contradiction.]

(d) Assuming III.9.4.12. is true, show that π and e are algebraically independent. [Consider $\{1, \pi i, e, e^{\pi i}\}$.]

If 666 is the embodiment of evil, is $\sqrt{666} = 25.8069758\cdots$ the root of all evil?

III.10. Hyperreal Numbers and Nonstandard Analysis

> "When we have grasped the spirit of the infinitesimal method, and have verified the exactness of its results either by the geometrical method of prime and ultimate ratios, or by the analytical method of derived functions, we may employ infinitely small quantities as a sure and valuable means of shortening and simplifying our proofs."
>
> *Lagrange*[8]

In the early years, calculus, especially in the Leibniz formulation, was based on the notion of an "infinitesimal" change; indeed, the name "calculus" is a shortened form of "infinitesimal calculus." The logical foundations of calculus were vigorously, and sometimes rightly, challenged, and appeared to be quite shaky even to some of the developers and defenders of the theory, although no one could deny the apparent accuracy and usefulness of the theory in describing and explaining analytic and physical phenomena. (Leibniz himself regarded infinitesimals as a "fiction, although a well-founded one," and considered them to be convenient symbolic shorthand for limit arguments, although his pronouncements on the subject were somewhat inconsistent. Newton felt much the same way. See [Jah03, p. 97-98].) When it became more important to mathematicians in the nineteenth century to carefully develop a rigorous theory of calculus and analysis, the intuitive approach via infinitesimals was deemed incapable of direct logical justification (Cantor was particularly firmly convinced that infinitesimals were a logical fiction, calling them the "cholera bacillus of mathematics"[9]), and was bypassed by the familiar $\epsilon - \delta$ formulations of Cauchy and Weierstrass, which remain the standard approach to this day. See [Gra08, p. 328-346] for a lucid and fascinating discussion of the attitudes about measurement, infinitesimals, and the Archimedean axiom around the turn of the 20th century.

Intuitive arguments involving infinitesimals are far from dead, however, and are quite useful in many instances, particularly to scientists using calculus in their work (in [Gri21], such arguments are called "vigorous," as opposed to "rigorous"). The dream among some mathematicians of a rigorous theory of infinitesimals never faded away.

Finally, in 1960, Abraham Robinson (partially anticipated by C. Schmieden and D. Laugwitz [SL58]) found a way to use fairly sophisticated ideas and results from mathematical logic to give a rigorous direct approach to analysis on $\mathbb{R}$ using infinitesimals. The subject growing out of his work is called *nonstandard analysis* (Robinson's name).

Nonstandard analysis is one of the most controversial parts of mathematics. There are some "true believers" who think the nonstandard approach is *the* way to do analysis, and that eventually the mathematical world will see the light; on the other hand, nonstandard methods are definitely unfashionable among most mainstream analysts, although some aspects of probability and mathematical economics seem to have more natural nonstandard formulations (). It should be noted that the controversy has little if anything to do with the correctness or legitimacy of nonstandard analysis as a part of mathematics; no one today except perhaps for a few constructivists (who also reject substantial parts of "classical" analysis) questions the soundness of nonstandard analysis as a mathematical theory.

My own view of nonstandard analysis is that the views on both sides of the controversy are overblown. It is a remarkable theoretical achievement that the infinitesimal approach to analysis can be made completely rigorous. And this is of more than theoretical importance: users of calculus can be assured (if indeed they need such assurance) that arguments involving infinitesimals can be carefully justified logically. While it

[8] *Méchanique Analytique*; cf. [Mor93, 1906].
[9] cf. [Dau88, p. xiii].

can be shown that (assuming some form of the Axiom of Choice) exactly the same theorems about analysis on $\mathbb{R}$ can be proved using nonstandard analysis as by traditional methods, it can happen (and in fact *has* happened) that nonstandard arguments are easier to find in some cases, so nonstandard analysis is a tool which should not be rejected or set aside as merely a curiosity. On the other hand, I think it is unlikely in the foreseeable future that nonstandard analysis will supplant the traditional approach. Proponents of nonstandard analysis often understate or underappreciate the degree of sophistication in mathematical logic needed to prove even basic theorems about limits via nonstandard analysis. (For example, the book [Gol98] is one of the most readable and appealing treatments of nonstandard analysis; nevertheless, it should be noted that this book does not contain a single complete rigorous nonstandard analysis proof of any theorem of real analysis since the transfer principle is not proved in any form. This is not necessarily a criticism of the book, but does tend to compromise the credibility of the author's forcefully held belief that nonstandard methods are the future of analysis.)

The basic idea of nonstandard analysis is a two-part one. First, the real numbers $\mathbb{R}$ are embedded in a larger ordered field ${}^*\mathbb{R}$ called the *hyperreal numbers*, with particular properties, notably including infinitesimal elements. Second, there is a systematic procedure, called the *transfer principle*, to convert any appropriately formulated statement P about $\mathbb{R}$ into a corresponding statement ${}^*P$ about ${}^*\mathbb{R}$, with the property that ${}^*P$ is true in ${}^*\mathbb{R}$ if and only if P is true in $\mathbb{R}$. One then proves P by proving ${}^*P$. Transfer is used in both directions since the transfer of statements already known to be true in $\mathbb{R}$ can be used to prove ${}^*P$. It is also often possible to rephrase ${}^*P$ as a simpler or more natural statement than P (e.g. the transfer of some statements can be rephrased as a precise version of a classical statement involving infinitesimals). Stating and proving the validity of the transfer principle is the place where mathematical logic comes into nonstandard analysis. The transfer principle is one of Robinson's fundamental theoretical contributions. (There is another, somewhat more esoteric, feature of more advanced nonstandard analysis, also introduced by Robinson: *concurrence*, which is discussed in ().)

As pointed out in [Pot04], nonstandard analysis is based on a delicate interplay between first- and second-order logic. One begins with the first-order theory of $\mathbb{R}$ (the theory of real-closed fields, although it suffices to just use the first-order theory of ordered fields), and then adds the second-order Axiom of Completeness; then restricts to the first-order fragment of the second-order theory thus obtained and applies the Compactness Theorem to obtain a non-Archimedean extension of the real numbers that is nonetheless elementarily equivalent to it.

To avoid the logical technicalities needed to define the hyperreal numbers and prove the transfer principle, the hyperreals and the transfer principle are sometimes presented axiomatically. It is argued that in many ordinary real analysis courses and texts, the real numbers are also presented axiomatically without a rigorous construction, and this approach to nonstandard analysis is no different. This argument is superficially valid, but on closer examination its persuasiveness begins to evaporate:

1. Rigorous mathematics can in fact be done in both settings, although the statements of the theorems proved are really different from the way they are usually expressed. For example, in an ordinary real analysis class in which $\mathbb{R}$ is presented axiomatically, the theorems are properly of the form

> "If there is a model for $\mathbb{R}$, then statement P is true in this model."

(Here P is a statement about real numbers, sets of real numbers, functions on $\mathbb{R}$, ...). One must then be convinced of the existence (and essential uniqueness) of a model for $\mathbb{R}$ for such a theorem to assume its usual form.

Similarly, theorems in a nonstandard analysis course in which ${}^*\mathbb{R}$ (as well as $\mathbb{R}$) is presented axiomatically will be of the form

> "If a model for $\mathbb{R}$ exists, *and a model for* ${}^*\mathbb{R}$ *with a valid transfer principle exists for this model of* $\mathbb{R}$, then statement P is true in this model of $\mathbb{R}$."

This is quite a different theorem until one is convinced of the existence of models of the hyperreal numbers.

I can write down any list of axioms I want and prove theorems from them, but until it is shown that there is a model for the axioms, or preferably that the axioms are satisfied in some natural setting with real applications, the theory is an empty one. (This is one of my principal objections to formalist mathematics, although far from my only one.)

2. By the time they begin calculus, most students are prepared, and even comfortable, to accept on faith the existence of a model for $\mathbb{R}$ satisfying the usual axioms. But few students, even (one might even say *especially*) good ones, are prepared to accept on faith the existence of a model for the hyperreals. This point is elucidated at length in the introduction to [Gol98]. I generally agree with GOLDBLATT on this issue, but I would go farther in that I do not share his expectation that eventually beginning calculus students will be as willing to accept the hyperreals as they are to accept $\mathbb{R}$ today, primarily because of the next point.

3. There are rigorous explicit models for $\mathbb{R}$, e.g. via Dedekind cuts, which are sufficiently simple and concrete that they can be followed and understood by almost any undergraduate mathematics major or even honors calculus student. But all (known) models for the hyperreals are at least an order of magnitude more difficult. Even the construction via ultrapowers, which is probably the simplest and most easily understood of the (known) models, is beyond the ability of most undergraduates to fully understand and appreciate. (To begin with, a free ultrafilter on $\mathbb{N}$ is needed. There is no explicit description of any such ultrafilter; indeed, some form of the Axiom of Choice is needed to even prove existence. And even after the construction of ${}^*\mathbb{R}$ some nontrivial mathematical logic is needed to justify the transfer principle.)

In this section, we will not take an axiomatic approach. We will first construct (a model for) the hyperreals via ultrapowers, and then discuss and justify the transfer principle.

Every serious mathematician needs to be aware of the basics of nonstandard analysis, and to have a tolerant if not enthusiastic attitude about its place in the world of mathematics.

III.10.1. Ordered Fields Containing $\mathbb{R}$

The basic object of study in nonstandard analysis is an ordered field ${}^*\mathbb{R}$ of *hyperreal numbers* which contains $\mathbb{R}$ as a proper subfield. Let us first examine what such an ordered field must look like.

III.10.1.1. Recall (III.1.2.9.) that an *ordered field* is a set F with two binary operations satisfying the field axioms (III.1.1.2.), along with a total ordering which can be either specified as satisfying (Ord), or by giving a positive cone satisfying (Pos). The smallest subset of F containing 1 and closed under addition, multiplication, negation, and reciprocals can be naturally identified with $\mathbb{Q}$ as in III.1.11.2.. Each element of F has an absolute value: $|x| = x$ if $x = 0$ or $x \in P$, and $|x| = -x$ if $-x \in P$.

Throughout this section, F will denote an ordered field containing $\mathbb{R}$ (more precisely, with a specified subfield order-isomorphic to $\mathbb{R}$).

III.10.1.2. DEFINITION. Let F be an ordered field containing $\mathbb{R}$, and let $x \in F$.

x is *limited* if $|x| < r$ for some $r \in \mathbb{R}$.

x is *unlimited* if it is not limited.

x is *infinitesimal* if $|x| < r$ for all $r \in \mathbb{R}$, $r > 0$.

The terms *finite* and *infinite* (or *infinitely large*) are sometimes used instead of *limited* and *unlimited* respectively, but these terms can be confusing in certain contexts. According to our definition, 0 is an infinitesimal; this is not entirely standard.

Every element of $\mathbb{R}$ is clearly limited; the only infinitesimal element of $\mathbb{R}$ is 0.

The field F has the Archimedean property () if and only if F has no unlimited or nonzero infinitesimal elements.

III.10.1.3. PROPOSITION. Let F be an ordered field containing $\mathbb{R}$, $x, y \in F$. Then

(i) If x and y are limited, so are $x + y$, xy, and $-x$.

(ii) If x and y are infinitesimal, then $x + y$ and $-x$ are infinitesimal.

(iii) If x is infinitesimal and y is limited, then xy is infinitesimal. In particular, if x is infinitesimal and $r \in \mathbb{R}$, then rx is infinitesimal.

(iv) If $x \neq 0$, then x is infinitesimal if and only if x^{-1} is unlimited.

III.10.1.4. PROPOSITION. Let F be an ordered field properly containing $\mathbb{R}$. Then

(i) F contains unlimited elements and nonzero infinitesimals.

(ii) If $x \in F$ is limited, then there is a unique $s \in \mathbb{R}$ such that $x - s$ is infinitesimal. (This s is called the *shadow* of x, written $s = sh(x)$.)

PROOF: (ii) Replacing x by $-x$ if necessary, we may assume $x \geq 0$. Let $A = \{r \in \mathbb{R} : x < r\}$. By assumption, A is nonempty, and it is bounded below (e.g. by 0). Let $s = \inf(A)$ (infimum in $\mathbb{R}$). If $x = s$, we are done. If $x < s$, then $0 < s - x$. If $s - x > r$ for $r \in \mathbb{R}$, $r > 0$, then $x < s - r$, contradicting the definition of s. Thus $s - x$ is infinitesimal. If $x > s$, a similar proof shows that $x - s$ is infinitesimal. Uniqueness is obvious from III.10.1.3.(ii)–(iii) [if $x - s$ and $x - t$ are infinitesimal, so is $s - t = (x - t) - (x - s)$].
(i): Let $x \in F \setminus \mathbb{R}$. If x is unlimited, then x^{-1} is infinitesimal. If x is limited, let $s = sh(x)$. Then $x - s$ is a nonzero infinitesimal, and $(x - s)^{-1}$ is unlimited. 👍

The next result is also a corollary of (), but we give a direct proof.

III.10.1.5. COROLLARY. Let F be an ordered field properly containing $\mathbb{R}$. Then F is not order-complete.

PROOF: Let I be the set of infinitesimals in F. Then I is nonempty and bounded (e.g. by 1). Suppose I has a supremum x. Since F contains a positive infinitesimal by III.10.1.4., $x > 0$. Also, every positive $r \in \mathbb{R}$ is an upper bound for I, so x is infinitesimal. But then $2x$ is infinitesimal and $2x > x$, a contradiction. 👍

III.10.1.6. It is obvious that if x, y are limited, then $sh(x+y) = sh(x)+sh(y)$ and $sh(xy) = sh(x)sh(y)$, and that x is infinitesimal if and only if $sh(x) = 0$; $sh(x) = sh(y)$ if and only if $x - y$ is infinitesimal.

III.10.1.7. DEFINITION. If F is an ordered field containing $\mathbb{R}$ and $x, y \in F$, then $x \simeq y$ if $x - y$ is infinitesimal. The *halo* of x is

$$hal(x) = \{y \in F : y \simeq x\} = \{x + z : z \text{ infinitesimal}\} .$$

III.10.1.8. It is immediate from III.10.1.3.(ii)–(iii) that $\simeq$ is an equivalence relation; thus, if $x, y \in F$, then either $hal(x) = hal(y)$ or $hal(x)$ and $hal(y)$ are disjoint. If x and y are limited, then $hal(x) = hal(y)$ if and only if $sh(x) = sh(y)$. Each halo of limited elements contains exactly one real number ($hal(x) = hal(sh(x))$ if x is limited).

III.10.1.9. One can also define a weaker equivalence relation: $x \sim y$ if $x-y$ is limited. The $\sim$-equivalence class of x is called the *galaxy* of x. The galaxy of any limited element is the set of all limited elements.

III.10.2. The Hyperreal Numbers via Ultrafilters

It is not obvious that there exist ordered fields properly containing $\mathbb{R}$. A simple example (the simplest example) is given in Exercise III.5.3.1.. In this subsection we describe a somewhat more complicated example with additional properties which will allow us to do nonstandard analysis.

III.10.2.10. Recall () that an *ultrafilter* on a set X is a collection $\mathcal{U}$ of subsets of X such that

(i) $\mathcal{U}$ is closed under finite intersections.

(ii) If $A \in \mathcal{U}$ and $A \subseteq B$, then $B \in \mathcal{U}$.

(iii) $\emptyset \notin \mathcal{U}$.

(iv) If $A \subseteq X$, then either $A \in \mathcal{U}$ or $A^c \in \mathcal{U}$.

(the two possibilities in (iv) are mutually exclusive by (i) and (iii).)
The ultrafilter $\mathcal{U}$ is *free* if the intersection of the sets in $\mathcal{U}$ is $\emptyset$ (equivalently, if every set with finite complement is in $\mathcal{U}$).

We think of an ultrafilter as giving a coherent way of declaring which subsets of X are "large" (those in the ultrafilter) and which are "negligibly small" (those not in the ultrafilter). Every subset is either large or negligibly small, but never both. If the ultrafilter is free, every finite subset is negligibly small.

III.10.2.11. Using the Axiom of Choice, it can be shown () that every infinite set has a free ultrafilter on it; in fact, there are $2^{2^{|X|}}$ free ultrafilters on an infinite set X.

Throughout the rest of this subsection, $\mathcal{U}$ will denote a free ultrafilter on $\mathbb{N}$ which is specified once and for all. It does not matter for nonstandard analysis purposes which of the $2^{2^{\aleph_0}}$ free ultrafilters on $\mathbb{N}$ we choose. While we use $\mathbb{N}$ as an index set to make things as simple and concrete as possible, the construction works identically for any index set; and for some purposes a much larger index set is needed (III.10.8.13.), although not too large (III.10.8.14.).

III.10.2.12. DEFINITION. Let $\mathbb{S}$ be the set of all sequences of real numbers. Define an equivalence relation on $\mathbb{S}$ as follows. If $x = (x_m)$ and $y = (y_m)$ are in $\mathbb{S}$, set $x \sim y$ if

$$\{m \in \mathbb{N} : x_m = y_m\} \in \mathcal{U} \ .$$

Denote the equivalence class of x by $[[(x_m)]]$, and let ${}^*\mathbb{R}$ be the set of equivalence classes. ${}^*\mathbb{R}$ is called the set of *hyperreal numbers* (with respect to $\mathcal{U}$).

It is easy to check that $\sim$ is an equivalence relation. We think of two sequences as being equivalent if they agree "almost everywhere" (i.e. except on a "negligibly small" set of indices).

III.10.2.13. In order to specify an element of ${}^*\mathbb{R}$, it is only necessary to specify a "large" set of coordinates (a set in $\mathcal{U}$). The other coordinates can be chosen arbitrarily and do not affect the element of ${}^*\mathbb{R}$ specified. This observation will be quite useful in avoiding the annoying and unnecessary step of extending definitions to all of $\mathbb{N}$.

III.10.2.14. $\mathbb{S}$ has natural algebraic operations of addition, subtraction, and multiplication defined coordinatewise. These drop to well-defined operations on ${}^*\mathbb{R}$ by the following proposition:

III.10.2.15. PROPOSITION. If $(x_m) \sim (z_m)$ and $(y_m) \sim (w_m)$, then $(x_m + y_m) \sim (z_m + w_m)$, $(x_m - y_m) \sim (z_m - w_m)$, and $(x_m y_m) \sim (z_m w_m)$.

PROOF: We have

$$\{m \in \mathbb{N} : x_m + y_m = z_m + w_m\} \supseteq \{m \in \mathbb{N} : x_m = z_m\} \cap \{m \in \mathbb{N} : y_m = w_m\} \ .$$

By assumption, the sets on the right are in $\mathcal{U}$, hence so is their intersection by III.10.2.10.(i); therefore the set on the left is also in $\mathcal{U}$ by III.10.2.10.(ii). The proof in the other two cases is almost identical. 👍

III.10.2.16. PROPOSITION. The coordinatewise operations on $\mathbb{S}$ give well-defined operations of addition, subtraction, and multiplication on ${}^*\mathbb{R}$ making it into a unital commutative ring (). The embedding of $\mathbb{R}$ into ${}^*\mathbb{R}$ as equivalence classes of constant sequences respects addition, subtraction, and multiplication.

PROOF: It is immediate from III.10.2.15. that addition, subtraction, and multiplication are well defined on ${}^*\mathbb{R}$, and then almost a triviality that addition and multiplication are associative and commutative, and that the distributive law holds. The additive identity is the equivalence class of the constant sequence 0, and the multiplicative unit the equivalence class of the constant sequence 1. The last statement is also obvious. 👍

If $r \in \mathbb{R}$, we will identify r with the equivalence class of the constant sequence (r), and hence regard $\mathbb{R}$ as a subset of ${}^*\mathbb{R}$; the operations on ${}^*\mathbb{R}$ extend the usual ones on $\mathbb{R}$.

III.10.2.17. Reciprocals and division can also be defined with only a slight bit of care. Suppose $x = [[(x_m)]] \in {}^*\mathbb{R}$, $x \neq 0$. Then $\{m \in \mathbb{N} : x_m \neq 0\} \in \mathcal{U}$, so we can define $x^{-1} \in {}^*\mathbb{R}$ as the equivalence class of (x_m^{-1}) (this sequence is only defined on a set of indices in $\mathcal{U}$, but this is sufficient by the remark in III.10.2.13..) It is clear that $xx^{-1} = 1$, so x^{-1} is a multiplicative inverse for x. Thus ${}^*\mathbb{R}$ is a field, and $\mathbb{R}$ is a subfield.

III.10.2.18. We can also define a total ordering on ${}^*\mathbb{R}$. If $x \in {}^*\mathbb{R}$, $x = [[(x_m)]]$, we would like to say that $x > 0$ if $\{m \in \mathbb{N} : x_m > 0\} \in \mathcal{U}$. It is easily seen that this definition is independent of the representing sequence (x_m). It gives a good definition due to the next proposition.

III.10.2.19. PROPOSITION. If $(x_m) \in \mathbb{S}$, then exactly one of the following sets is in $\mathcal{U}$:

$$A = \{m \in \mathbb{N} : x_m > 0\}, \quad B = \{m \in \mathbb{N} : x_m = 0\}, \quad C = \{m \in \mathbb{N} : x_m < 0\} .$$

PROOF: By III.10.2.10.(iv), either A or $A^c = \{m \in \mathbb{N} : x_m \leq 0\}$ is in $\mathcal{U}$, but not both. If $A \in \mathcal{U}$, then $A^c \notin \mathcal{U}$, so $B \notin \mathcal{U}$ and $C \notin \mathcal{U}$ by III.10.2.10.(ii) since $B, C \subseteq A^c$. Suppose $A^c \in \mathcal{U}$; then $A \notin \mathcal{U}$. Either $C \in \mathcal{U}$ or $C^c \in \mathcal{U}$ but not both. If $C \in \mathcal{U}$, then $B \notin \mathcal{U}$ as above. If $C^c \in \mathcal{U}$, then $B = A^c \cap C^c \in \mathcal{U}$ and $A, C \notin \mathcal{U}$. 👍

We can then define the positive cone in ${}^*\mathbb{R}$. The construction is a special case of taking the nonstandard extension of a subset of $\mathbb{R}$ ().

III.10.2.20. COROLLARY. Let ${}^*P$ be the set of $x = [[(x_m)]] \in {}^*\mathbb{R}$ such that $\{m \in \mathbb{N} : x_m > 0\} \in \mathcal{U}$. Then ${}^*P$ satisfies the positive cone axioms () and thus defines a total order on ${}^*\mathbb{R}$ extending the usual order on $\mathbb{R}$.

PROOF: By III.10.2.19., ${}^*\mathbb{R}$ is the disjoint union of ${}^*P$, $\{0\}$, and $-({}^*P)$. We only need to check that ${}^*P$ is closed under addition and multiplication, which follows from an argument almost identical to the proof of III.10.2.15., and that $({}^*P) \cap \mathbb{R}$ is the usual positive cone of $\mathbb{R}$, which is obvious. 👍

We summarize the construction in a theorem:

III.10.2.21. THEOREM. ${}^*\mathbb{R}$ with the above algebraic operations and ordering is an ordered field containing $\mathbb{R}$.

One can even extend the algebraic operations on $\mathbb{R}$ such as exponentiation. This is also a special case of a general procedure for extending functions from $\mathbb{R}$ to ${}^*\mathbb{R}$ ().

III.10.2.22. DEFINITION. Let $\boldsymbol{x} = [[(x_m)]]$ and $\boldsymbol{a} = [[(a_m)]]$ be elements of ${}^*\mathbb{R}$, with $\boldsymbol{x} > 0$. Define

$$\boldsymbol{x^a} = [[(x_m^{a_m})]]$$

(note that $x_m^{a_m}$ is defined for almost all m).

III.10.2.23. PROPOSITION. Exponentiation is well defined, and satisfies the usual rules of exponents: if $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b} \in {}^*\mathbb{R}$ with $\boldsymbol{x}, \boldsymbol{y} \geq 0$, then $\boldsymbol{x^a x^b = x^{a+b}}$, $\boldsymbol{\frac{x^a}{x^b} = x^{a-b}}$, $\boldsymbol{(x^a)^b = x^{ab}}$, $\boldsymbol{(xy)^a = x^a y^a}$. If $\boldsymbol{x < y}$ and $\boldsymbol{a > 0}$, then $\boldsymbol{x^a < y^a}$; if $\boldsymbol{a < 0}$, then $\boldsymbol{x^a > y^a}$. If $\boldsymbol{x, a} \in \mathbb{R}$, exponentiation in ${}^*\mathbb{R}$ agrees with usual exponentiation in $\mathbb{R}$.

The proof is routine along the usual lines. See Exercise ().

III.10.2.24. Corollary. Let $\boldsymbol{x} \in {}^*\mathbb{R}$, $\boldsymbol{x} > 0$, and $\boldsymbol{a} \in {}^*\mathbb{R}$, $\boldsymbol{a} \neq 0$. Then there is a unique $\boldsymbol{y} \in {}^*\mathbb{R}$, $\boldsymbol{y} > 0$, such that $\boldsymbol{y}^{\boldsymbol{a}} = \boldsymbol{x}$. In particular, if $n \in \mathbb{N}$, then $\boldsymbol{x}$ has a unique nonnegative n'th root in ${}^*\mathbb{R}$.

III.10.2.25. We can similarly define logarithms on ${}^*\mathbb{R}$. If $0 < \boldsymbol{a} = [[(a_m)]] \in {}^*\mathbb{R}$ and $0 < \boldsymbol{x} = [[(x_m)]] \in {}^*\mathbb{R}$, define

$$\log_{\boldsymbol{a}} \boldsymbol{x} = [[(\log_{a_m} x_m)]]$$

(this is well defined by the usual argument). We then have $\boldsymbol{y} = \log_{\boldsymbol{a}} \boldsymbol{x}$ if and only if $\boldsymbol{a}^{\boldsymbol{y}} = \boldsymbol{x}$.

III.10.2.26. It may not yet be obvious that ${}^*\mathbb{R}$ is larger than $\mathbb{R}$. But it is in fact much larger. We can easily define an unlimited element of ${}^*\mathbb{R}$: set $x_m = m$ and $\boldsymbol{x} = [[(x_m)]] \in {}^*\mathbb{R}$. If $r \in \mathbb{R}$, then $\{m \in \mathbb{N} : x_m \leq r\}$ is finite, so $\boldsymbol{x} > r$. Similarly, if $y_m = m^{-1}$ and $\boldsymbol{y} = [[(y_m)]] \in {}^*\mathbb{R}$, then $\boldsymbol{y}$ is a positive infinitesimal in ${}^*\mathbb{R}$. Actually, if (x_m) is any sequence of distinct real numbers, then $[[(x_m)]] \in {}^*\mathbb{R} \setminus \mathbb{R}$.

Note, however, that ${}^*\mathbb{R}$ is not larger than $\mathbb{R}$ in a cardinality sense. We have $|\mathbb{S}| = \left(2^{\aleph_0}\right)^{\aleph_0} = 2^{\aleph_0}$, so $|{}^*\mathbb{R}| \leq 2^{\aleph_0}$, and the opposite inequality is obvious, so $|{}^*\mathbb{R}| = 2^{\aleph_0} = |\mathbb{R}|$. (This is a result of taking an ultrafilter on $\mathbb{N}$ for the construction of ${}^*\mathbb{R}$. If a free ultrafilter is used on a set of large cardinality, the cardinality of ${}^*\mathbb{R}$ can be larger than $2^{\aleph_0}$. See Exercise III.10.8.10..)

III.10.2.27. Like any totally ordered set, ${}^*\mathbb{R}$ has an order topology (). However, this topology has pathological properties (Exercise III.10.8.9.), and is of little if any use in nonstandard analysis.

III.10.3. Internal and External Sets and Functions

Since ${}^*\mathbb{R}$ is considerably larger than $\mathbb{R}$, general subsets of ${}^*\mathbb{R}$ and general functions from ${}^*\mathbb{R}$ to ${}^*\mathbb{R}$ are very complicated (and can depend on the specific construction of ${}^*\mathbb{R}$ used). Relatively few subsets of ${}^*\mathbb{R}$ or functions from ${}^*\mathbb{R}$ to ${}^*\mathbb{R}$ are sufficiently well describable in standard terms that they can be considered "transferable" from $\mathbb{R}$ using a transfer principle (in a way independent of the construction used for ${}^*\mathbb{R}$). In this subsection we describe the ones which can be used, called *internal sets and functions*.

The Nonstandard Extension of a Subset of $\mathbb{R}$

The simplest subsets of ${}^*\mathbb{R}$ for transfer purposes are ones directly definable from a subset of $\mathbb{R}$.

III.10.3.1. Definition. Let $A \subseteq \mathbb{R}$. Then the *nonstandard extension of* A is the set

$${}^*A = \{[[(x_m)]] \in {}^*\mathbb{R} : x_m \in A \text{ for almost all } m\}$$

where "$x_m \in A$ for almost all m" means "$\{m \in \mathbb{N} : x_m \in A\} \in \mathcal{U}$."

Note that if $[[(x_m)]] = [[(y_m)]]$, i.e. $x_m = y_m$ for almost all m, then $x_m \in A$ for almost all m if and only if $y_m \in A$ for almost all m (cf. the proof of III.10.2.15.), so membership in ${}^*A$ is well defined. ${}^*A$ could also be defined to be simply the set of equivalence classes of sequences in A.

III.10.3.2. PROPOSITION. Let $A \subseteq \mathbb{R}$. Then

(i) ${}^*A \cap \mathbb{R} = A$ (in particular, $A \subseteq {}^*A$).

(ii) ${}^*A = A$ if and only if A is finite.

(iii) ${}^*(\mathbb{R} \setminus A) = {}^*\mathbb{R} \setminus {}^*A$.

PROOF: (i) is obvious: if $r \in \mathbb{R}$, then $[[(r)]] \in {}^*A$ if and only if $r \in A$. For (ii), if (x_m) is a sequence of distinct elements of A, then $[[(x_m)]] \in {}^*A \setminus \mathbb{R} = {}^*A \setminus A$, so if A is infinite we have ${}^*A \neq A$. On the other hand, if $A = \{r^{(1)}, \dots, r^{(n)}\}$, and $[[(x_m)]] \in {}^*A$, set $N_k = \{m \in \mathbb{N} : x_m = r^{(k)}\}$ for $1 \leq k \leq n$. Then $N_1 \cup \cdots \cup N_n \in \mathcal{U}$, so $N_k \in \mathcal{U}$ for some (exactly one) k (), i.e. $[[(x_m)]] = r^{(k)}$. (Note also that ${}^*\emptyset = \emptyset$.) For (iii), note that if $[[(x_m)]] \in {}^*\mathbb{R}$, then exactly one of the sets $\{m \in \mathbb{N} : x_m \in A\}$ and $\{m \in \mathbb{N} : x_m \notin A\}$ is in $\mathcal{U}$. 👍

III.10.3.3. PROPOSITION. Let $A, B \subseteq \mathbb{R}$. Then ${}^*(A \cup B) = {}^*A \cup {}^*B$, ${}^*(A \cap B) = {}^*A \cap {}^*B$, and ${}^*(A \setminus B) = {}^*A \setminus {}^*B$.

PROOF: Note that if $\boldsymbol{x} = [[(x_m)]] \in {}^*\mathbb{R}$, then

$$\{m \in \mathbb{N} : x_m \in A\} \cap \{m \in \mathbb{N} : x_m \in B\} = \{m \in \mathbb{N} : x_m \in A \cap B\}$$

and the set on the right is in $\mathcal{U}$ if and only if both sets on the left are in $\mathcal{U}$, so $\boldsymbol{x} \in {}^*(A \cap B)$ if and only if $\boldsymbol{x} \in {}^*A \cap {}^*B$. The result for unions and symmetric differences follows from this and III.10.3.2.(iii) by taking complements. 👍

III.10.3.4. Note that this statement is *false* in general for infinite unions and intersections, even countable ones. (Any infinite set is a union of finite sets; use III.10.3.2.(ii).)

The proof of the next proposition is routine along the same lines:

III.10.3.5. PROPOSITION. Let $a, b \in \mathbb{R}$, $a < b$, and denote by $[a, b]$, (a, b), $(a, b]$, $[a, b)$ the usual intervals in $\mathbb{R}$. Then

$$\begin{aligned} {}^*[a, b] &= \{\boldsymbol{x} \in {}^*\mathbb{R} : a \leq \boldsymbol{x} \leq b\} \\ {}^*(a, b) &= \{\boldsymbol{x} \in {}^*\mathbb{R} : a < \boldsymbol{x} < b\} \\ {}^*(a, b] &= \{\boldsymbol{x} \in {}^*\mathbb{R} : a < \boldsymbol{x} \leq b\} \\ {}^*[a, b) &= \{\boldsymbol{x} \in {}^*\mathbb{R} : a \leq \boldsymbol{x} < b\} \ . \end{aligned}$$

Similar statements hold for ${}^*(-\infty, a]$, ${}^*(a, +\infty)$, etc.

III.10.3.6. PROPOSITION. Let $A \subseteq \mathbb{R}$. Then the following are equivalent:

(i) ${}^*A$ is bounded above in ${}^*\mathbb{R}$.

(ii) Every positive element of ${}^*A$ is limited.

(iii) A is bounded above in $\mathbb{R}$.

If these conditions hold, any upper bound for A in $\mathbb{R}$ is an upper bound for ${}^*A$ in ${}^*\mathbb{R}$.

PROOF: (ii) $\Rightarrow$ (i) is trivial: any unlimited positive element of ${}^*\mathbb{R}$ would be an upper bound for ${}^*A$. For (iii) $\Rightarrow$ (ii), and for the last statement, if b is an upper bound for A in $\mathbb{R}$, then ${}^*A \subseteq {}^*(-\infty, b]$, so by III.10.3.5. b is an upper bound for ${}^*A$ and (ii) holds. To prove (i) $\Rightarrow$ (iii), suppose A is not bounded above in $\mathbb{R}$. Let $\boldsymbol{x} = [[(x_m)]] \in {}^*\mathbb{R}$. For each m let $y_m \in A$ with $y_m \geq x_m$. Then $\boldsymbol{y} = [[(y_m)]] \in {}^*A$ and $\boldsymbol{x} \leq \boldsymbol{y}$. ☝

III.10.3.7. Perhaps the most important nonstandard extension set is ${}^*\mathbb{N}$, the *hypernatural numbers*, equivalence classes of sequences of natural numbers. We also have ${}^*\mathbb{Z}$, the *hyperintegers*, and ${}^*\mathbb{Q}$, the *hyperrational numbers*. We have that ${}^*\mathbb{N}$ is closed under addition and multiplication, ${}^*\mathbb{Z}$ is a subring of ${}^*\mathbb{R}$, and ${}^*\mathbb{Q}$ is a subfield of ${}^*\mathbb{R}$. Every element of ${}^*\mathbb{Z}$ is a difference of two elements of ${}^*\mathbb{N}$, and every element of ${}^*\mathbb{Q}$ is a quotient of an element of ${}^*\mathbb{Z}$ by an element of ${}^*\mathbb{N}$. All of these sets turn out to be uncountable (III.10.8.3.). See Exercises III.10.8.3.–III.10.8.8. for some properties of ${}^*\mathbb{N}$.

Nonstandard Extensions of Functions

III.10.3.8. Let $A \subseteq \mathbb{R}$, and let f be a function from A to $\mathbb{R}$. We can extend f to a function ${}^*f$ from ${}^*A$ to ${}^*\mathbb{R}$ as follows. Let $\boldsymbol{x} = [[(x_m)]] \in {}^*A$. Then $x_m \in A$ for almost all m, so the element $[[(f(x_m)]]$ is well defined as an element of ${}^*\mathbb{R}$. It is easily seen in the usual way that this element of ${}^*\mathbb{R}$ is independent of the choice of the representative (x_m) chosen for x. Write ${}^*f(\boldsymbol{x}) = [[(f(x_m)]]$.

III.10.3.9. If f is a polynomial or rational function with real coefficients, then ${}^*f$ defined this way agrees with the usual algebraic extension of f to a function on ${}^*\mathbb{R}$. More generally, if $a \in \mathbb{R}$, and $f(x) = x^a$ on $[0, +\infty)$, then ${}^*f$ is the function on ${}^*[0, +\infty)$ defined by nonstandard exponentiation (III.10.2.22.). Similarly, if $0 < a \in \mathbb{R}$, the nonstandard extension of the exponential function $f(x) = a^x$ is given by nonstandard exponentiation. If $0 < a \in \mathbb{R}$, then the nonstandard extension of the function $f(x) = \log_a x$ is the nonstandard logarithm function of III.10.2.25..

III.10.3.10. Another common instance of nonstandard extension is for trigonometric functions. The functions ${}^* \sin x$ and ${}^* \cos x$ are defined for all nonstandard real numbers; the domain of ${}^* \tan \boldsymbol{x}$ is the set

$$ {}^* \left(\mathbb{R} \setminus \left\{ \frac{\pi}{2} + n\pi : n \in \mathbb{N} \right\} \right) = {}^*\mathbb{R} \setminus \left\{ \frac{\pi}{2} + n\pi : n \in {}^*\mathbb{N} \right\} $$

and the other extended trig functions have similar domains.

III.10.3.11. It is usual to omit the * on the nonstandard extension of a function except in the relatively rare case (e.g. in III.10.3.14.–III.10.3.15.) where a distinction must be made between the original function and its extension. Thus, for example, we will usually just write sin, $\log_a$, etc., for the extended functions.

The proof of the next result is routine by the usual arguments:

III.10.3.12. PROPOSITION. All standard trigonometric identities hold for the nonstandard extensions. For example:

$\sin^2 \boldsymbol{x} + \cos^2 \boldsymbol{x} = 1$ for all $\boldsymbol{x} \in {}^*\mathbb{R}$.

$\sin(\boldsymbol{x} + \boldsymbol{y}) = \sin \boldsymbol{x} \cos \boldsymbol{y} + \cos \boldsymbol{x} \sin \boldsymbol{y}$ for all $\boldsymbol{x}, \boldsymbol{y} \in {}^*\mathbb{R}$.

III.10.3.13. Inequalities valid for standard functions also hold for their extensions. For example:

$e^{\boldsymbol{x}} > 0$ for all $\boldsymbol{x} \in {}^*\mathbb{R}$; $e^{\boldsymbol{x}} < 1$ for $\boldsymbol{x} < 0$ and $e^{\boldsymbol{x}} > 1$ for $\boldsymbol{x} > 0$.

$\log \boldsymbol{x} < 0$ for $\boldsymbol{x} \in {}^*\mathbb{R}$, $0 < \boldsymbol{x} < 1$; $\log \boldsymbol{x} > 0$ for $\boldsymbol{x} \in {}^*\mathbb{R}$, $\boldsymbol{x} > 1$.

$-1 \leq \sin \boldsymbol{x} \leq 1$ and $-1 \leq \cos \boldsymbol{x} \leq 1$ for all $\boldsymbol{x} \in {}^*\mathbb{R}$.

If the domain of f is A, then the domain of ${}^*f$ is by definition ${}^*A$. We also have:

III.10.3.14. PROPOSITION. If $f : A \to \mathbb{R}$ and $B \subseteq A$, then ${}^*f({}^*B) = {}^*[f(B)]$. In particular, the range of ${}^*f$ is ${}^*[f(A)]$.

PROOF: If $\boldsymbol{x} = [[(x_m)]] \in {}^*B$, then $x_m \in B$ for almost all m, so $f(x_m) \in f(B)$ for almost all m and $[[(f(x_m))]] \in {}^*[f(B)]$. Conversely, if $\boldsymbol{y} = [[(y_m)]] \in {}^*[f(B)]$, then for almost all m we have $y_m \in f(B)$, so there is an $x_m \in B$ with $y_m = f(x_m)$. The element $\boldsymbol{x} = [[(x_m)]]$ is well defined and in ${}^*B$, and ${}^*f(\boldsymbol{x}) = \boldsymbol{y}$, so $\boldsymbol{y} \in {}^*f({}^*B)$.

III.10.3.15. If $f : A \to \mathbb{R}$, $g : B \to \mathbb{R}$, and $f(A) \subseteq B$, then $g \circ f : A \to \mathbb{R}$ is defined. It is routine to check that ${}^*(g \circ f) = {}^*g \circ {}^*f$ as a function from ${}^*A$ to ${}^*\mathbb{R}$. In particular, if f is one-to-one on A, then ${}^*(f^{-1}) \circ {}^*f$ is the identity on ${}^*A$ and ${}^*f \circ {}^*(f^{-1})$ is the identity on ${}^*[f(A)]$, so ${}^*f$ is a bijection from ${}^*A$ to ${}^*[f(A)]$ and $({}^*f)^{-1} = {}^*(f^{-1})$.

Nonstandard Euclidean Space

III.10.3.16. For fixed $n \in \mathbb{N}$, there is a natural one-one correspondence between $\mathbb{S}^n$, the set of n-tuples of sequences of real numbers, and the set of sequences of n-tuples of real numbers:

$$((x_m^{(1)}), (x_m^{(2)}), \ldots, (x_m^{(n)})) \longleftrightarrow ((x_m^{(1)}, x_m^{(2)}, \ldots, x_m^{(n)}))$$

and we have that

$$(x_m^{(1)}, x_m^{(2)}, \ldots, x_m^{(n)}) = (y_m^{(1)}, y_m^{(2)}, \ldots, y_m^{(n)}) \text{ for almost all } m$$

if and only if, for each k, $1 \leq k \leq n$, we have $x_m^{(k)} = y_m^{(k)}$ for almost all m. Thus there is an unambiguous equivalence relation on the set of sequences of n-tuples of real numbers, and the set ${}^*(\mathbb{R}^n)$ of equivalence classes can be naturally identified with $({}^*\mathbb{R})^n$, the set of n-tuples of hyperreal numbers. We will denote this set by ${}^*\mathbb{R}^n$ and freely describe it both ways. Write $[[((x_m^{(1)}, \ldots, x_m^{(n)}))]]$ for the equivalence class of the sequence $((x_m^{(1)}, \ldots, x_m^{(n)}))$, thought of as a typical representative of an element of ${}^*\mathbb{R}^n$.

The Nonstandard Extension of a Relation

The nonstandard extensions of sets and functions can be subsumed in the general notion of the nonstandard extension of an n-ary relation.

III.10.3.17. Recall that an *n-ary relation* on $\mathbb{R}$ is simply a subset of $\mathbb{R}^n$. In particular, a usual (binary) relation is a subset of $\mathbb{R}^2$; recall that a function is officially a binary relation (its graph). A unary relation is just a subset of $\mathbb{R}$.

III.10.3.18. DEFINITION. If A is a subset of $\mathbb{R}^n$, i.e. an n-ary relation on $\mathbb{R}$, we define its *nonstandard extension* to be

$$^*A = \{[[((x_m^{(1)}, \ldots, x_m^{(n)}))]] : (x_m^{(1)}, \ldots, x_m^{(n)}) \in A \text{ for almost all } m\} \subseteq {}^*\mathbb{R}^n \ .$$

If $n = 1$, this agrees with the previous definition of the nonstandard extension of a subset of $\mathbb{R}$. It is also easy to verify that the nonstandard extension of a function on $\mathbb{R}$ coincides with the nonstandard extension of the corresponding binary relation (the graph of the function).

III.10.3.19. The binary relation $<$ on $\mathbb{R}$ has a nonstandard extension $^*<$, which coincides with the ordering $<$ on $^*\mathbb{R}$ defined in III.10.2.20.. Even the binary relation $=$ on $\mathbb{R}$ has ordinary equality on $^*\mathbb{R}$ as its nonstandard extension.

III.10.3.20. A function from $\mathbb{R}^n$ to $\mathbb{R}^k$ can be regarded as an $(n+k)$'ary relation on $\mathbb{R}$, and thus has a nonstandard extension, which can be defined similarly to the nonstandard extension of a function from $\mathbb{R}$ to $\mathbb{R}$. For some simple examples, addition and multiplication may be regarded as functions from $\mathbb{R}^2$ to $\mathbb{R}$; their nonstandard extensions are nothing but addition and multiplication on $^*\mathbb{R}$ as we have defined them. Another example: $f(x, y) = \log_x y$ is a real-valued function defined on $(0, \infty)^2$, and its nonstandard extension is the logarithm function defined in III.10.2.25..

III.10.4. The Transfer Principle

By now we have shown that a great many statements about the real numbers have nonstandard analogs, and the proofs of the nonstandard statements from the standard ones all seem pretty much the same. A few classes of examples (phrased informally) are:

> $\mathbb{R}$ satisfies the ordered field axioms.
>
> The rules of exponents hold in $\mathbb{R}$.
>
> Trigonometric identities hold in $\mathbb{R}$.

We might suspect that there is a metaprinciple that *every* statement about $\mathbb{R}$ has an analog in $^*\mathbb{R}$ and that there is a single scheme for simultaneously deriving proofs of the nonstandard versions of all true standard statements. With due care, there is: this is the *transfer principle*.

However, we must be careful. Not every true statement about $\mathbb{R}$ can have a nonstandard version. It is easy to find counterexamples: a simple one is "$\mathbb{R}$ contains no nonzero infinitesimals." A more important

example is the Completeness Property of $\mathbb{R}$; for somewhat deeper reasons (III.10.1.5.) this statement cannot be true verbatim about ${}^*\mathbb{R}$.

So we must instead say that every *appropriately formulated* statement about $\mathbb{R}$ has a nonstandard analog, which is true if and only if the standard version is true. Precisely describing what "appropriately formulated" means is where mathematical logic comes into the picture.

The transfer principle has two levels: *first-order transfer* and *higher-order transfer*. First-order transfer is relatively simple to formulate and even not too hard to prove, and only requires a modest amount of mathematical logic; and first-order transfer is adequate for most elementary applications of nonstandard analysis. Higher-order transfer is more sophisticated logically, but also more powerful. We will concentrate on first-order transfer and only give a brief survey of higher-order transfer.

Simple or First-Order Transfer

III.10.4.1. The setup for first-order transfer will be to begin with a relational structure $\mathcal{S} = (\Omega, \mathcal{R}, \mathcal{F})$ (II.1.8.1.), which will be called the *standard structure*. In almost all applications $\mathcal{R}$ and $\mathcal{F}$ will be the sets of all relations and functions on Ω respectively; in basic nonstandard analysis Ω will be $\mathbb{R}$ with $\mathcal{R}$ and $\mathcal{F}$ the sets of all relations and functions. At first the reader may want to just assume we are working with this structure.

A subset of Ω^n can be regarded as an n-ary relation on Ω; for technical reasons we will assume that the domain of any function in $\mathcal{F}$ is a relation in $\mathcal{R}$.

In addition, we will fix an index set I and an ultrafilter $\mathcal{U}$ on I. For basic nonstandard analysis we take $I = \mathbb{N}$ and $\mathcal{U}$ any free ultrafilter on $\mathbb{N}$.

III.10.4.2. The first step is to define a nonstandard version of $\mathcal{S}$. We let ${}^*\Omega$ be the set of equivalence classes of functions from I to Ω, where two functions are equivalent if they agree "almost everywhere" (abbreviated *a.e.*), or "for almost all i," i.e. on a set in $\mathcal{U}$. In the basic nonstandard analysis case where $\Omega = \mathbb{R}$ and $I = \mathbb{N}$ the set ${}^*\Omega$ is just ${}^*\mathbb{R}$ as defined in ().

III.10.4.3. If R is an n-ary relation in $\mathcal{R}$, we define a nonstandard version ${}^*R$, which will be an n-ary relation on ${}^*\Omega$, by letting ${}^*R$ be the set of all n-tuples $(\omega_1, \ldots, \omega_n)$ for which there is for each k a representation $\omega_k = [[(\omega_{ki})]]$ such that

$$(\omega_{1i}, \ldots, \omega_{ni}) \in R$$

for almost all $i \in I$. A routine argument along the lines of () shows that if this condition is satisfied for one n-tuple of representatives of $\omega_1, \ldots, \omega_n$, it is satisfied for any other n-tuple of representatives. In the basic nonstandard analysis case, ${}^*R$ is just the relation defined in III.10.3.18..

Similarly, if $f \in \mathcal{F}$ is a function from a subset of Ω^n to Ω, we define ${}^*f$ to be the function from a subset of $({}^*\Omega)^n$ to ${}^*\Omega$ given by

$${}^*f(\omega_1, \ldots, \omega_n) = [[(f(\omega_{1i}, \ldots, \omega_{ni})]]$$

if $\omega_k = [[(\omega_{ki})]]$ for each k and $(\omega_{1i}, \ldots, \omega_{ni})$ is in the domain of f for almost all i. It is again routine to check that ${}^*f$ is well defined independent of the choice of representatives, and that the domain of ${}^*f$ is ${}^*(dom\ f)$ (recall that $dom\ f$ is a relation in $\mathcal{R}$, so ${}^*(dom\ f)$ is a defined subset of $({}^*\Omega)^n$). In the basic nonstandard analysis case, ${}^*f$ is just the function defined in III.10.3.8..

III.10.4.4. DEFINITION. The relational structure ${}^*\mathcal{S} = ({}^*\Omega, {}^*\mathcal{R}, {}^*\mathcal{F})$, where ${}^*\mathcal{R} = \{{}^*R : R \in \mathcal{R}\}$ and ${}^*\mathcal{F} = \{{}^*f : f \in \mathcal{F}\}$, is called the *(first-order) nonstandard structure* for $\mathcal{S}$ (with respect to I and $\mathcal{U}$).

Although the notation does not reflect it, ${}^*\mathcal{S}$ depends on I and $\mathcal{U}$ as well as on $\mathcal{S}$. However, for nonstandard analysis purposes the dependence on I and $\mathcal{U}$ is mostly irrelevant and can be largely ignored (but see III.10.8.13. and III.10.8.14.). Note that even if $\mathcal{R}$ is the set of all relations on Ω, ${}^*\mathcal{R}$ will be far from containing all relations on ${}^*\Omega$ in general, and similarly for ${}^*\mathcal{F}$. The relations and functions in ${}^*\mathcal{R}$ and ${}^*\mathcal{F}$ will be the relations and functions on ${}^*\Omega$ which are definable by transfer of definitions in Ω.

III.10.4.5. Next we fix a simple language $\mathcal{L}$ for the relational structure $\mathcal{S}$ (), and a language ${}^*\mathcal{L}$ for the relational structure ${}^*\mathcal{S}$. We do not require that ${}^*\mathcal{L}$ be a simple language for ${}^*\mathcal{S}$, e.g. it could be a simple language for a larger relational structure such as $({}^*\Omega, \tilde{\mathcal{R}}, \tilde{\mathcal{F}})$, where $\tilde{\mathcal{R}}$ and $\tilde{\mathcal{F}}$ are the sets of all relations and all functions on ${}^*\Omega$. But the only formulas in ${}^*\mathcal{L}$ we will consider are ones from the simple sublanguage for ${}^*\mathcal{S}$, so we may restrict to this sublanguage when convenient. We may assume that the variable names in ${}^*\mathcal{L}$ contain all the variable names in $\mathcal{L}$ (in fact, since we only ever use a countable set of variables, we could always use a common set of variable names in all languages).

We then define a "transfer" of each symbol, term, and formula of $\mathcal{L}$ into a corresponding symbol, term, or formula in ${}^*\mathcal{L}$:

III.10.4.6. DEFINITION. (i) If c is the name of an element $\omega \in \Omega$, then the *transfer* ${}^*c$ of c is a name of the element of ${}^*\Omega$ defined by the constant function ω from I to Ω. (We will notationally identify the element $c \in \mathcal{L}$ with ${}^*c \in {}^*\mathcal{L}$, i.e. we will regard $\Omega \subseteq {}^*\Omega$, to simplify notation.)

(ii) If $R \in \mathcal{R}$, the *transfer* of R is ${}^*R \in {}^*\mathcal{R}$.

(iii) If $f \in \mathcal{F}$, the *transfer* of f is ${}^*f \in {}^*\mathcal{F}$.

(iv) The *transfer* of any variable or logical symbol $(\forall, \Rightarrow)$ is itself.

(v) The *transfer* of any term is defined inductively: the transfer of a constant or variable has already been defined, and the transfer of a term $f(t_1, \ldots, t_n)$ is ${}^*f({}^*t_1, \ldots, {}^*t_n)$, where ${}^*t_k$ is the transfer of t_k for each k.

Transfer of formulas is then defined inductively:

III.10.4.7. DEFINITION. (i) The *transfer* of the formula $R(t_1, \ldots, t_n)$ is ${}^*R({}^*t_1, \ldots, {}^*t_n)$, where ${}^*t_k$ is the transfer of the term t_k for each k.

(ii) The *transfer* of the formula $\forall x_1 \cdots \forall x_n(P)$ is $\forall x_1 \cdots \forall x_n({}^*P)$, where ${}^*P$ is the transfer of P.

(iii) The *transfer* of the formula $P_1 \wedge \cdots \wedge P_n$ is $({}^*P_1) \wedge \cdots \wedge ({}^*P_n)$, where ${}^*P_k$ is the transfer of P_k for each k.

(iv) The *transfer* of the formula $P_1 \Rightarrow P_2$ is $({}^*P_1) \Rightarrow ({}^*P_2)$, where ${}^*P_k$ is the transfer of P_k for each k.

(v) The *transfer* of any formula of level $n+1$ is defined by application of the appropriate one of (i)–(iv) to the constituent formulas of level n.

A routine induction argument verifies:

III.10.4.8. Proposition. The transfer of any symbol, term, or formula in $\mathcal{L}$ is a well-defined symbol, term, or formula in ${}^*\mathcal{L}$. For any formula P, the variables appearing in ${}^*P$ are the same as the ones appearing in P; and a variable is quantified in ${}^*P$ if and only if it is quantified in P. In particular, the transfer of any statement in $\mathcal{L}$ is a statement in ${}^*\mathcal{L}$.

The Transfer Principle is:

III.10.4.9. Theorem. [Simple or First-Order Transfer Principle] Let $\mathcal{S}$ be a relational structure, ${}^*\mathcal{S}$ a first-order nonstandard structure for $\mathcal{S}$, $\mathcal{L}$ a simple language for $\mathcal{S}$, and ${}^*\mathcal{L}$ a language for ${}^*\mathcal{S}$. If P is any statement of $\mathcal{L}$, with transfer ${}^*P$ as defined in III.10.4.7., then ${}^*P$ holds in ${}^*\mathcal{S}$ if and only if P holds in $\mathcal{S}$.

III.10.4.10. This result is symmetric in the sense that transfer goes both ways due to the "if and only if." But it is important to note that transfer is not quite symmetric in $\mathcal{L}$ and ${}^*\mathcal{L}$: every statement in $\mathcal{L}$ has a transfer in ${}^*\mathcal{L}$, but not every statement in ${}^*\mathcal{L}$ is the transfer of a statement in $\mathcal{L}$ in general. If ${}^*\mathcal{L}$ is a simple language for ${}^*\mathcal{S}$, then transfer is truly symmetric, since the only formulas of ${}^*\mathcal{L}$ will be transfers of formulas in $\mathcal{L}$.

III.10.4.11. We can extend transfer a little to make it more useful. If the relational structure $\mathcal{S}$ is nondegenerate (II.1.8.8.), we can fix a false statement Φ in $\mathcal{L}$ containing no variables. Then ${}^*\Phi$ is also false in ${}^*\mathcal{S}$, and contains no variables. We can extend the simple language $\mathcal{L}$ to an extended simple language $\bar{\mathcal{L}}$ by adding the symbols $\neg, \vee, \exists, \iff$ as in II.1.8.10., with a standard translation as in II.1.8.10. using Φ; similarly, if ${}^*\mathcal{L}$ is simple, or more generally does not contain the extra logical symbols, we can extend ${}^*\mathcal{L}$ to ${}^*\bar{\mathcal{L}}$ in the same manner, using ${}^*\Phi$. We then routinely check:

III.10.4.12. Proposition. Each formula P of $\bar{\mathcal{L}}$ has a well-defined transfer to a formula ${}^*P$ of ${}^*\bar{\mathcal{L}}$. If Q is the standard translation of P to $\mathcal{L}$, then the standard translation of ${}^*P$ to ${}^*\mathcal{L}$ is ${}^*Q$. The transfer of a statement is a statement.

Using this, we can extend Transfer:

III.10.4.13. Theorem. [Extended Simple or First-Order Transfer Principle] Let $\mathcal{S}$ be a relational structure, ${}^*\mathcal{S}$ a first-order nonstandard structure for $\mathcal{S}$, $\mathcal{L}$ a simple language for $\mathcal{S}$, ${}^*\mathcal{L}$ a simple language for ${}^*\mathcal{S}$, and $\bar{\mathcal{L}}$ and ${}^*\bar{\mathcal{L}}$ the corresponding extended languages. If P is any statement of $\bar{\mathcal{L}}$, with transfer ${}^*P$ as defined in III.10.4.12., then ${}^*P$ holds in ${}^*\mathcal{S}$ if and only if P holds in $\mathcal{S}$.

Proof: Let Q be the standard translation of P in $\mathcal{L}$. Then

$$P \text{ holds in } \mathcal{S}$$
$$\iff Q \text{ holds in } \mathcal{S}$$
$$\iff {}^*Q \text{ holds in } {}^*\mathcal{S}$$

$\iff {}^*P$ holds in ${}^*\mathcal{S}$

where the first and third equivalences are by definition (III.10.4.12. is needed for the third) and the second equivalence is by III.10.4.9.. 👍

Proof of the Simple Transfer Principle

The proof of the Simple Transfer Principle, by induction on level, is actually quite straightforward and routine. It is a nearly immediate corollary of the simple version of Łoś's Theorem:

III.10.4.14. Theorem. [Łoś's Theorem, Simple or First-Order Version] Let $\mathcal{S}$ be a relational structure, ${}^*\mathcal{S}$ a first-order nonstandard structure for $\mathcal{S}$, $\mathcal{L}$ a simple language for $\mathcal{S}$, and ${}^*\mathcal{L}$ a language for ${}^*\mathcal{S}$. Let P be any formula of $\mathcal{L}$, with transfer ${}^*P$ as defined in III.10.4.7., with unquantified variables among $x_1, \ldots, x_n$; let $c_1, \ldots, c_n$ be names in ${}^*\mathcal{L}$ for elements of ${}^*\Omega$ with (c_{ki}) a representative for c_k with c_{ki} the name of an element of Ω for each (k, i). Then ${}^*P$ holds for $(c_1, \ldots, c_n)$ in ${}^*\mathcal{S}$ if and only if, for almost all i, P holds for $(c_{1i}, \ldots, c_{ni})$ in $\mathcal{S}$.

Superstructures

III.10.4.15. Definition. If X is a set, we define

$$V_0(X) = X$$

$$V_{n+1}(X) = V_n(X) \cup \mathcal{P}(V_n(X))$$

$$V(X) = \cup_{n=0}^{\infty} V_n(X)$$

$V(X)$ is called the *superstructure* of X.

Note that if $y \in V(X) \setminus X$, then $y \subseteq V(X)$. The converse is false; however, if $y \subseteq V_n(X)$ for some n, then $y \in V_{n+1}(X) \subseteq V(X)$.

III.10.4.16. There is a similarity between the construction of superstructures and II.8.9.7.. However, $V(X)$ is not a model for set theory (for example, $\mathcal{P}(V(X))$ is not an element of $V(X)$); to get a model for set theory, the construction has to be continued transfinitely for all ordinals (and then the result is a class, not a set). For nonstandard analysis, however, it is adequate to just continue the construction to $\aleph_0$.

III.10.4.17. For technical reasons, we will want to assume that $\emptyset \notin X$ and that no element of $V(X)$ belongs to an element of X, i.e. that $x \cap V(X) = \emptyset$ for all $x \in X$. This is normally accomplished by assuming that the elements of X are *atoms* (or *urelements*), i.e. if $x \in X$, then $x \neq \emptyset$ but that x has no elements. For a general X we can just replace X by an equipotent set consisting of atoms. For basic nonstandard analysis, we simply assume that real numbers are atoms of set theory, which we can do (III.2.2.9.). We will call the elements of X the *elements* of the superstructure $V(X)$, and the elements of $V(X) \setminus X$ the *sets* of the superstructure $V(X)$. If A is a set in $V(X)$, then all the elements of A are either elements or sets in $V(X)$.

III.10.4.18. We can define ordered pairs in $V(X)$: if $a, b \in V_k(X)$, then $\{a\}$ and $\{a, b\}$ are in $V_{k+1}(X)$, so $(a, b) = \{\{a, \{a, b\}\} \in V_{k+2}(X)$. Similarly, ordered n-tuples of elements of $V(X)$ are again elements of $V(X)$. In particular, all ordered n-tuples of elements of X are in $V_2(X)$, and thus $X^n \in V_3(X)$. More generally, the cartesian product of any finite number of subsets of $V_k(X)$ is again an element of $V_{k+3}(X) \subseteq V(X)$.

An n-ary relation on $V_k(X)$ is a subset of $V_k(X)^n$, hence an element of $V_{k+4}(X)$. Thus $V(X)$ contains all relations on $V_k(X)$ for any k. Any such relation is a relation on $V(X)$. But note that not all relations on $V(X)$ are of this form. An n-ary relation on $V(X)$ which is contained in $V_k(X)^n$ for some k is called *limited*; these are precisely the relations on $V(X)$ which are elements of $V(X)$.

Similarly, any function from $V_k(X)^n$ to $V_k(X)^m$ is in $V_{k+4}(X)$, hence in $V(X)$. In particular, any function from X^n to X^m is in $V_4(X)$, hence in $V(X)$. As with relations, not every function on $V(X)$ is in $V(X)$, only ones whose domain and range is contained in $V_k(X)$ for some k (called *limited functions*).

III.10.4.19. We can thus form a relational structure $(V(X), \mathcal{R}_\ell, \mathcal{F}_\ell)$, where $\mathcal{R}_\ell$ and $\mathcal{F}_\ell$ are the sets of limited relations and limited functions in $V(X)$. It is convenient to only consider limited relations and functions and not the unwieldy sets of all relations and functions on $V(X)$. However, there are a few basic relations and functions on $V(X)$ which are not limited, which it is convenient to include: $\in$, $\subseteq$, $\cup$, $\cap$, $\setminus$. These can be avoided at some notational cost: for example, $\in$ can be avoided by only working with the coherent family of limited relations $\in_k$, where $\in_k$ is the membership relation on $V_k(X)$. We will thus include these five relations in $\mathcal{R}_\ell$ and $\cup$, $\cap$, and $\setminus$ in $\mathcal{F}_\ell$, essentially to ease notation. (Actually any relations or functions on $V(X)$ can be thrown in with no cost by regarding them as coherent families of limited relations or functions in the same manner.)

III.10.4.20. We can then fix a simple language (II.1.8.2.) $\mathcal{L}_X$ for the relational structure $\mathcal{X} = (V(X), \mathcal{R}_\ell, \mathcal{F}_\ell)$, which will be called the *language* of the superstructure of X. Since this relational structure is always nondegenerate (even if $X = \emptyset$!), the language $\mathcal{L}_X$ can be extended to an extended simple language $\bar{\mathcal{L}}_X$ adding $\neg$, $\vee$, $\exists$, $\iff$ (II.1.8.10.), which will have their usual meanings. The language $\bar{\mathcal{L}}_X$ is rich enough that we can express any reasonable statement about X, subsets of X, relations and functions on X, etc., in a relatively straightforward way.

III.10.4.21. Some references argue that inclusion of $\mathcal{F}_\ell$ in the relational structure is not necessary: in a relational structure $(\Omega, \mathcal{R}, \mathcal{F})$, the only use of the functions in $\mathcal{F}$ is to define terms; in the case of $(V(X), \mathcal{R}_\ell, \mathcal{F}_\ell)$ this is redundant since the terms defined this way are already constants in $V(X)$. Thus the use of terms, as opposed to constants, is unnecessary. This is technically correct, but terms in a simple language can include variables as well as constants, so inclusion of the functions in $\mathcal{F}_\ell$ serves to expand the simple language of the superstructure in a useful way without causing additional difficulties (sentences in the expanded language can be systematically translated into the simpler language).

Nonstandard Superstructures

III.10.4.22. If we now fix and index set I and an ultrafilter $\mathcal{U}$ on I, we can form the corresponding nonstandard relational structure ${}^*\mathcal{X} = ({}^*V(X), {}^*\mathcal{R}_\ell, {}^*\mathcal{F}_\ell)$. We would like to regard this relational structure as a (generally proper) substructure of $(V({}^*X), \mathcal{R}_\ell^*, \mathcal{F}_\ell^*)$, where $\mathcal{R}_\ell^*$ and $\mathcal{F}_\ell^*$ are the sets of limited relations and functions on $V({}^*X)$. However, there is a logical problem: ${}^*V(X)$ is not a subset of $V({}^*X)$ (although ${}^*X$ is naturally a subset of both). For example, if $A \subseteq X$, then the object ${}^*A$ of ${}^*V_1(X)$ is not a subset of ${}^*X$. Put another way, A may be regarded as a unary relation R_A on X, i.e. as an element of $\mathcal{R}_\ell$, but the element

${}^*A$ of ${}^*V_1(X)$ is not the unary relation ${}^*R_A$ on ${}^*X$ (which is technically a subset of ${}^*X$, i.e. an element of $V_1({}^*X)$).

This observation leads to the solution: identify ${}^*V_1(X) = {}^*X \cup {}^*\mathcal{P}(X)$ with a subset of $V_1({}^*X) = {}^*X \cup \mathcal{P}({}^*X)$ by identifying ${}^*x \in {}^*X$ with itself and $A \subseteq X$ with ${}^*R_A \in \mathcal{P}({}^*X)$. The embedding is well defined since $X \cap \mathcal{P}(X) = \emptyset$ and hence ${}^*X \cap {}^*\mathcal{P}(X) = \emptyset$, and is indeed an embedding (injective) since ${}^*X \cap \mathcal{P}({}^*X) = \emptyset$. Call the embedding μ_0.

This procedure can be continued inductively, although with a slight complication. The embedding μ_1 of ${}^*V_1(X)$ into $V_1({}^*X)$ defines an embedding of $\mathcal{P}({}^*V_1(X))$ into $\mathcal{P}(V_1({}^*X))$, so we can map ${}^*V_2(X) = {}^*V_1(X) \cup {}^*\mathcal{P}(V_1(X))$ into $V_2({}^*X) = V_1({}^*X) \cup \mathcal{P}(V_1({}^*X))$ as follows. The map μ_2 is defined on ${}^*V_1(X)$ as μ_1. If $B \subseteq V_1(X)$, so ${}^*B \in {}^*\mathcal{P}(V_1(X))$, regard B as a unary relation R_B on $V_1(X)$; then ${}^*R_B$ is a unary relation on ${}^*V_1(X)$, hence defines an element of $\mathcal{P}({}^*V_1(X))$ which in turn defines an element C of $\mathcal{P}(V_1({}^*X))$. Set $\mu_2({}^*B) = C$. It is not entirely obvious that μ_2 is well defined, since $V_1(X) \cap \mathcal{P}(V_1(X)) \neq \emptyset$; but in fact $V_1(X) \cap \mathcal{P}(V_1(X)) = \{\emptyset\}$ and ${}^*\emptyset = \emptyset$ unambiguously. It is unclear that μ_2 is injective in general.

Iterating this construction, we get a well-defined map μ_n from ${}^*V_n(X)$ to $V_n({}^*X)$ for each n (the proof that the map is well defined becomes successively more complicated as n increases), and hence a map $\mu : {}^*V(X) \to V({}^*X)$, called the *Mostowski collapsing map*. We also need to identify relations in ${}^*\mathcal{R}_\ell$ with relations in $\mathcal{R}_\ell^*$: since every relation in ${}^*\mathcal{R}_\ell$ is an element of

III.10.5. Other Nonstandard Structures

III.10.6. Proofs and Applications of Nonstandard Analysis

III.10.7. Loeb Measures and a Nonstandard Approach to Measure Theory

III.10.8. Exercises

III.10.8.1. (a) Show that ${}^*\mathbb{R}$ has no countable cofinal subset: if $(\boldsymbol{a}^{(n)})$ is a sequence in ${}^*\mathbb{R}$, then there is a $\boldsymbol{b} \in {}^*\mathbb{R}$ with $\boldsymbol{a}^{(n)} < \boldsymbol{b}$ for all n. [For each $m \in \mathbb{N}$, set $b_m = 1 + \max\{a_m^{(1)}, \ldots, a_m^{(m)}\}$.]

(b) Show that if $(\boldsymbol{a}^{(n)})$ and $(\boldsymbol{b}^{(n)})$ are sequences in ${}^*\mathbb{R}$ with $\boldsymbol{a}^{(n)} < \boldsymbol{b}^{(r)}$ for all n and r, there are $\boldsymbol{c}, \boldsymbol{d} \in {}^*\mathbb{R}$ with $\boldsymbol{a}^{(n)} < \boldsymbol{c} < \boldsymbol{d} < \boldsymbol{b}^{(n)}$ for all n. [For almost all $m \in \mathbb{N}$, choose $c_m, d_m \in \mathbb{R}$ with $a_m^{(n)} < c_m < d_m < b_m^{(r)}$ for all $n, r \leq m$.]

III.10.8.2. (a) Show that if F is any ordered field and $a, b, c, d \in F$ with $a < b$ and $c < d$, the open intervals (a, b) and (c, d) are order-isomorphic. [Use a "linear" function.]

(b) Show that any two open intervals in ${}^*\mathbb{R}$ [of the form (a, b), $(a, +\infty)$, or $(-\infty, b)$] are order-isomorphic. [Use the extensions () of the tangent and arctangent functions to give an order-isomorphism between ${}^*\mathbb{R}$ and $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.] What about intervals which are just open in the order topology (cf. III.10.8.9.), i.e. unions of intervals of the above form?

III.10.8.3. Let $\boldsymbol{n} \in {}^*\mathbb{N} \setminus \mathbb{N}$. Show that $\boldsymbol{n} - 1 \in {}^*\mathbb{N} \setminus \mathbb{N}$. Show that the intersection $\gamma(\boldsymbol{n})$ of the galaxy of $\boldsymbol{n}$ with ${}^*\mathbb{N}$ is a copy of $\mathbb{Z}$ (consisting of $\{\boldsymbol{n} + k : k \in \mathbb{Z}\}$). Show that ${}^*\mathbb{N}$ consists of $\mathbb{N}$ and uncountably many disjoint such $\gamma(\boldsymbol{n})$, and between any two such $\gamma(\boldsymbol{n})$, or between $\mathbb{N}$ and any $\gamma(\boldsymbol{n})$, there are uncountably many $\gamma(\boldsymbol{k})$. In particular, ${}^*\mathbb{N}$ is uncountable. Give similar results for ${}^*\mathbb{Z}$ and ${}^*\mathbb{Q}$. [Argue as in III.10.8.1..]

III.10.8.4. If $\boldsymbol{x}, \boldsymbol{a} \in {}^*\mathbb{N}$, show that the nonstandard exponential $\boldsymbol{x}^{\boldsymbol{a}}$ is in ${}^*\mathbb{N}$.

III.10.8.5. Let $\boldsymbol{n} = [[(n_m)]] \in {}^*\mathbb{N}$. Either $\{m \in \mathbb{N} : n_m \text{ is odd }\}$ or $\{m \in \mathbb{N} : n_m \text{ is even }\}$ is in $\mathcal{U}$. Say $\boldsymbol{n}$ is *odd* or *even* in the two cases. Show that if $\boldsymbol{n} \in {}^*\mathbb{N}$, then either $\boldsymbol{n}$ or $\boldsymbol{n}+1$ is even, and $\boldsymbol{n}$ is even if and only if $\boldsymbol{n}/2 \in {}^*\mathbb{N}$.

III.10.8.6. If $\boldsymbol{k}, \boldsymbol{n} \in {}^*\mathbb{N}$, say $\boldsymbol{k}$ *divides* $\boldsymbol{n}$ (written $\boldsymbol{k}|\boldsymbol{n}$) if $\boldsymbol{n}/\boldsymbol{k} \in {}^*\mathbb{N}$.

(a) If $\boldsymbol{k} = [[(k_m)]]$ and $\boldsymbol{n} = [[(n_m)]]$ with $k_m, n_m \in \mathbb{N}$ for almost all m, then $\boldsymbol{k}|\boldsymbol{n}$ if and only if $k_m|n_m$ for almost all m.

(b) If $\boldsymbol{k}, \boldsymbol{n} \in {}^*\mathbb{N}$, show that $\boldsymbol{k}$ and $\boldsymbol{n}$ have a unique *greatest common divisor* in ${}^*\mathbb{N}$, an element $\boldsymbol{d} \in {}^*\mathbb{N}$ such that $\boldsymbol{d}|\boldsymbol{k}$, $\boldsymbol{d}|\boldsymbol{n}$, and for any $\boldsymbol{c} \in {}^*\mathbb{N}$ such that $\boldsymbol{c}|\boldsymbol{k}$ and $\boldsymbol{c}|\boldsymbol{n}$, we have that $\boldsymbol{c}|\boldsymbol{d}$. Similarly show that $\boldsymbol{k}$ and $\boldsymbol{n}$ have a *least common multiple* in ${}^*\mathbb{N}$.

III.10.8.7. Let $\boldsymbol{n} = [[(m!)]] \in {}^*\mathbb{N}$. Show that $\boldsymbol{n}$ is divisible by every $k \in \mathbb{N}$.

III.10.8.8. Let $\boldsymbol{p} \in {}^*\mathbb{N}$. Say $\boldsymbol{p}$ is *prime* if whenever $\boldsymbol{p} = \boldsymbol{k}\boldsymbol{n}$ with $\boldsymbol{k}, \boldsymbol{n} \in {}^*\mathbb{N}$, then either $\boldsymbol{k}$ or $\boldsymbol{n}$ is 1.

(a) If $\boldsymbol{p} = [[(p_m)]]$ with $p_m \in \mathbb{N}$ for almost all m, then $\boldsymbol{p}$ is prime in ${}^*\mathbb{N}$ if and only if p_m is prime in $\mathbb{N}$ for almost all m.

(b) If Π is the set of primes in $\mathbb{N}$, then ${}^*\Pi$ is the set of primes in ${}^*\mathbb{N}$.

(c) If $\boldsymbol{p}$ is prime in ${}^*\mathbb{N}$, and $\boldsymbol{k}, \boldsymbol{n} \in {}^*\mathbb{N}$, then $\boldsymbol{p}|\boldsymbol{k}\boldsymbol{n}$ if and only if $\boldsymbol{p}|\boldsymbol{k}$ or $\boldsymbol{p}|\boldsymbol{n}$.

(d) If $\boldsymbol{p}$ is prime in ${}^*\mathbb{N}$, and $\boldsymbol{n} \in {}^*\mathbb{N}$ with $\boldsymbol{p}|\boldsymbol{n}$, then there is a unique $\boldsymbol{r} \in {}^*\mathbb{N}$ for which $\boldsymbol{n} = \boldsymbol{p}^{\boldsymbol{r}}\boldsymbol{k}$ for some $\boldsymbol{k} \in {}^*\mathbb{N}$ such that $\boldsymbol{p}$ does not divide $\boldsymbol{k}$.

(e) Let $\boldsymbol{n}_1, \boldsymbol{n}_2 \in {}^*\mathbb{N}$. Show that $\boldsymbol{n}_1 = \boldsymbol{n}_2$ if and only if, for every prime $\boldsymbol{p}$ in ${}^*\mathbb{N}$, when we write $\boldsymbol{n}_1 = \boldsymbol{p}^{\boldsymbol{r}}\boldsymbol{k}_1$ and $\boldsymbol{n}_2 = \boldsymbol{p}^{\boldsymbol{s}}\boldsymbol{k}_2$ as in (d), we have $\boldsymbol{r} = \boldsymbol{s}$.

(f) Can every $\boldsymbol{n} \in {}^*\mathbb{N}$ be written as a (finite) product of prime powers? [See III.10.8.7..]

III.10.8.9. Let $\mathcal{T}$ be the order topology () on ${}^*\mathbb{R}$, i.e. the topology generated by the open intervals.

(a) Show that the halo and the galaxy of any element are both open and closed in ${}^*\mathbb{R}$.

(b) Show that the relative topology from $\mathcal{T}$ on $\mathbb{R}$ is the discrete topology.

(c) Show that $({}^*\mathbb{R}, \mathcal{T})$ is totally disconnected. [Use III.10.8.2..]

(d) Show that no point of ${}^*\mathbb{R}$ has a countable local base for $\mathcal{T}$. [Use III.10.8.1..]

III.10.8.10. Let X be an uncountable set. Mimic the construction of ${}^*\mathbb{R}$ using an ultrafilter on X containing the complements of all subsets of X with cardinality less than $|X|$. Show that ordered fields containing $\mathbb{R}$ of arbitrarily large cardinality can be constructed. (See [GJ76, 12E].) Are such fields a satisfactory model for the hyperreal numbers for the purpose of nonstandard analysis? See Exercise III.10.8.14..

III.10.8.11. (a) Repeat the construction of the hyperreal numbers using ${}^*\mathbb{R}$ in place of $\mathbb{R}$: let $\mathcal{U}$ be a free ultrafilter on $\mathbb{N}$, and ${}^*\mathbb{S}$ the set of sequences of elements of ${}^*\mathbb{R}$. Define an equivalence relation on ${}^*\mathbb{S}$ via $\mathcal{F}$ as before, and let ${}^{**}\mathbb{R}$ be the set of equivalence classes. Show that ${}^{**}\mathbb{R}$ is an ordered field containing ${}^*\mathbb{R}$. Is it essentially larger? Is it isomorphic to ${}^*\mathbb{R}$?

(b) Repeat the construction transfinitely, taking a "union" (direct limit) at limit ordinals, to construct ordered fields of arbitrary cardinality containing $\mathbb{R}$. Is there a transfer principle for these fields?

III.10.8.12. [Dav77] **A Nonstandard Construction of** $\mathbb{R}$. Construct the nonstandard rational numbers ${}^*\mathbb{Q}$. Let L be the set of all limited elements of ${}^*\mathbb{Q}$, and I the set of infinitesimals in ${}^*\mathbb{Q}$.

(a) Show that L is a commutative unital ring, and that I is a maximal ideal in L. Thus $R = L/I$ is a field.

(b) Show that the ordering on ${}^*\mathbb{Q}$ induces an ordering on R making it an ordered field. If $\mathbb{Q}$ is regarded as a subring of L in the usual way, then the image of $\mathbb{Q}$ in R is order-dense.

(c) If A and B are nonempty subsets of ${}^*\mathbb{Q}$ with $a < b$ for all $a \in A$, $b \in B$, then there is a $c \in {}^*\mathbb{Q}$ with $a < c < b$ for all $a \in A$, $b \in B$. [Use the argument of III.10.8.1.(b).]

(d) Conclude that R is a complete ordered field. [If A is a nonempty subset of R with no largest element which is bounded above, let B be the set of upper bounds for A. Consider the preimages in L, and apply (c).]

This is a very efficient construction of $\mathbb{R}$, but unlike the Dedekind cut or Cauchy sequence constructions it requires some form of AC (existence of a free ultrafilter on $\mathbb{N}$).

III.10.8.13. (a) Show that (assuming AC) using a countable index set (e.g. $\mathbb{N}$) in the construction of ${}^*\mathbb{R}$ is not sufficient to obtain concurrence (); an index set I of cardinality at least $2^{\aleph_0}$ is needed. [Consider a well ordering of $\mathbb{R}$ according to the smallest ordinal of cardinality $2^{\aleph_0}$, which is a concurrent relation on $\mathbb{R}$. Every element of ${}^*\mathbb{R}$ is determined by a function from I to $\mathbb{R}$, hence by at most $|I|$ elements of $\mathbb{R}$.]

(b) Show similarly that (under AC) to obtain concurrence in a nonstandard model for a structure of cardinality κ, an index set of cardinality at least κ is necessary. (Even then the ultrafilter needs to be carefully chosen.)

III.10.8.14. (a) Show that if an ultrafilter on some index set I is used for the construction of ${}^*\mathbb{R}$ which is not closed under countable intersections (such an ultrafilter is necessarily free), we obtain ${}^*\mathbb{R} \neq \mathbb{R}$ so that ${}^*\mathbb{R}$ is a true nonstandard model for $\mathbb{R}$. [Show there is a sequence (V_n) of subsets of I with $\cup_{n=1}^{\infty} V_n = I$ and $V_n \notin \mathcal{U}$ for every n. Use this sequence to construct an element of ${}^*\mathbb{N} \setminus \mathbb{N}$.]

(b) Conversely, if an ultrafilter is used which is closed under countable intersections (this can only happen if the ultrafilter is fixed or if I is huge, of cardinality at least as large as a measurable cardinal), then ${}^*\mathbb{R} = \mathbb{R}$ so the construction is degenerate and not suitable for nonstandard analysis. [First show that if $U_1, U_2, \ldots$ is a sequence of subsets of I and $\cup_{n=1}^{\infty} U_n \in \mathcal{U}$, then some $U_n \in \mathcal{U}$. Then show that ${}^*\mathbb{N} = \mathbb{N}$ and that ${}^*\mathbb{R}$ has no unlimited elements, and apply III.10.1.4..]

III.10.8.15. Theorem. The following theorems are logically equivalent:

(i) Tikhonov's Theorem for Hausdorff spaces.

(ii) Tikhonov's Theorem for copies of $[0, 1]$.

(iii) Existence of Stone-Čech compactification for discrete spaces.

(iv) Existence of Stone-Čech compactification for arbitrary completely regular spaces.

Proof: (i) ⇒ (ii) and (iv) ⇒ (iii) are trivial. (ii) ⇒ (iv) is the standard construction of the Stone-Čech compactification (). So we need only prove (iii) ⇒ (i).

Let $\{X_i : i \in I\}$ be an indexed collection of compact Hausdorff spaces, and let $X = \prod_i X_i$ with the product topology. Let Y be the set X with the discrete topology, and βY its Stone-Čech compactification. For each i, the coordinate map $\pi_i : X \to X_i$, regarded as a map from Y to X_i, extends uniquely to a continuous function f_i from βY to X_i, by the universal property of βY. The f_i define a continuous function from βY to X (), which is surjective since its restriction to Y is the identity map from Y to X. Thus X is a continuous image of the compact space βY, hence compact by ().

Note that (ii) $\Rightarrow$ (i) is a simple consequence of the fact that every compact Hausdorff space can be embedded in a product of intervals, but we did not need to use this argument. 👍

It is a direct consequence of (iii) that every filter on a set X is contained in an ultrafilter. Thus this statement, along with the ones from the theorem, is logically strictly weaker than the Axiom of Choice.

III.11. Linear Algebra

Linear algebra is one of the most important and fundamental parts of elementary (university-level) mathematics. Although linear algebra is not directly part of the subject of this book, it is an indispensable tool in analysis, particularly vector analysis and functional analysis, as well as throughout mathematics in general. We will not give a full treatment of the subject here (there are any number of good texts available, as well as some that are not so good), but simply review the parts of the subject most relevant to analysis, often from the point of view of the analysis applications.

The linear algebra done in most texts is *finite-dimensional linear algebra*. While finite-dimensional linear algebra is important in analysis, especially vector analysis, much of advanced analysis uses infinite-dimensional vector spaces, so we will place more emphasis than is usual in linear algebra on aspects of the subject relevant in the infinite-dimensional case.

Abstract linear algebra is done over an arbitrary (but fixed) field $\mathbb{F}$. In 75% of the analysis applications (at least in this book), $\mathbb{F}$ will be the real numbers $\mathbb{R}$, and in 99% of the applications in which $\mathbb{F}$ is not $\mathbb{R}$ it is the complex numbers $\mathbb{C}$. The reader may just take the symbol $\mathbb{F}$ to stand for either $\mathbb{R}$ or $\mathbb{C}$ without essential loss of generality for analysis purposes. (In the remaining applications, $\mathbb{F}$ could be the rational numbers $\mathbb{Q}$, the hyperreal numbers ${}^*\mathbb{R}$ (III.10.2.21.), a field of rational functions (III.4.3.5.(c)) or Laurent series (III.5.3.3.), the p-adic numbers (III.5.3.9.), or occasionally something else.)

III.11.1. Matrix Algebra

Matrices can be regarded simply as an extremely useful shorthand notation for certain complicated algebraic systems. They are so useful, however, that they are really more than just a shorthand, becoming important mathematical objects in their own right.

III.11.1.1. DEFINITION. Let $m, n \in \mathbb{N}$. An *m-by-n ($m \times n$) matrix over* $\mathbb{F}$ is an array

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

of elements $\{a_{ij}\} \in \mathbb{F}$, with m rows and n columns. The element a_{ij} in the i'th row and j'th column is called the (i, j) entry of the matrix A, sometimes written A_{ij}. A is sometimes denoted $[a_{ij}]$ or $[a_{ij}]_{mn}$.

Actually, most of the theory of matrices works equally well over any ring R, i.e. the a_{ij} are all elements of R. For some of the theory it is necessary to restrict to a commutative ring (we will note these instances).

Addition of Matrices

III.11.1.2. Only matrices of the same size can be added. Addition is done entrywise:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{bmatrix}$$

The sum is another matrix of the same size.

III.11.1.3. EXAMPLE. Let

$$A = \begin{bmatrix} 1 & -2 & 4 \\ 0 & 1 & -2 \end{bmatrix}, B = \begin{bmatrix} -2 & 1 & 0 \\ 5 & -3 & 3 \end{bmatrix}, C = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix} .$$

Then

$$A + B = \begin{bmatrix} -1 & -1 & 4 \\ 5 & -2 & 1 \end{bmatrix}$$

and $A + C$ and $B + C$ are not defined since A and B are 2×3 and C is 2×2.

The next result is obvious:

III.11.1.4. PROPOSITION. Let $m, n \in \mathbb{N}$. Addition of $m \times n$ matrices over $\mathbb{F}$ is associative and commutative. There is an $m \times n$ zero matrix $0 = 0_{m \times n}$ such that $0 + A = A + 0 = A$ for every $m \times n$ matrix A. For every such A there is a unique $m \times n$ matrix $-A$ such that $A + (-A) = 0$.

In fact, $0_{m \times n}$ is the $m \times n$ matrix whose entries are all 0, and if $A = [a_{ij}]$, then $-A = [-a_{ij}]$.

III.11.1.5. We can also define scalar multiplication of matrices. If $A = [a_{ij}]$ is an $m \times n$ matrix over $\mathbb{F}$ and $r \in \mathbb{F}$, then define $rA = [ra_{ij}]$. Scalar multiplication can be regarded as a special case of matrix multiplication (), and thus satisfies all the usual algebraic rules () (this is easily verified directly). We have $-A = (-1)A$ for any A.

Row and Column Vectors

III.11.1.6. A $1 \times n$ matrix over $\mathbb{F}$ is of the form

$$\begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$$

and is usually called a *row n-vector*. The set of row n-vectors can be naturally identified with $\mathbb{F}^n$.

An $m \times 1$ matrix over $\mathbb{F}$ is of the form

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

and is usually called a *column m-vector*. The set of column m-vectors can be naturally identified with $\mathbb{F}^m$.

Under both identifications, matrix addition and scalar multiplication corresponds with usual vector addition and scalar multiplication.

Matrix Multiplication

III.11.1.7. The normal way to multiply matrices is more complicated and less obvious. Coordinatewise multiplication of matrices of the same size turns out not to be the useful product for most purposes (although this multiplication, called *Schur product*, does have some limited applications). The most useful product is hard to motivate until some further linear algebra is developed, so we will just give the definition and then see why it is important and has nice properties.

Multiplication of matrices is not necessarily defined for matrices of the same size. Instead, if A is an $m \times n$ matrix and B is a $q \times p$ matrix, then AB is defined if and only if $n = q$, and the product is then an $m \times p$ matrix. Here is the formal definition:

III.11.1.8. DEFINITION. Let $A = [a_{ij}]$ be an $m \times n$ matrix and $B = [b_{ij}]$ a $q \times p$ matrix. Then the matrix product AB is defined to be the $m \times p$ matrix $C = [c_{ij}]$, where

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

if $n = q$, and AB is undefined if $n \neq q$.

The idea is that to get the (i, j) entry of AB, the i'th row of A is "multiplied" times the j'th column of B (in a manner strongly reminiscent of the dot product in $\mathbb{R}^n$; in fact, this "multiplication" is exactly the dot product of the i'th row vector of A and the j'th column vector of B if $\mathbb{F} = \mathbb{R}$.) The product of the matrices is then defined if and only if the row vectors of A and the column vectors of B have the same length.

If A and B are *square matrices* of the same size (i.e. $n \times n$ for some n), then AB is defined and is again an $n \times n$ matrix.

III.11.1.9. EXAMPLES. Let

$$A = \begin{bmatrix} 1 & -2 & 4 \\ 0 & 1 & -2 \end{bmatrix}, B = \begin{bmatrix} -2 & 1 \\ 0 & 5 \\ -3 & 3 \end{bmatrix}, C = \begin{bmatrix} 2 & 3 \\ -1 & 2 \end{bmatrix}.$$

Then

$$AB = \begin{bmatrix} -14 & 3 \\ 6 & -1 \end{bmatrix}, BA = \begin{bmatrix} -2 & 5 & -10 \\ 0 & 5 & -10 \\ -3 & 9 & -6 \end{bmatrix}$$

$$CA = \begin{bmatrix} 2 & -1 & 2 \\ -1 & 4 & 0 \end{bmatrix}, BC = \begin{bmatrix} -5 & -4 \\ -5 & 10 \\ -9 & -3 \end{bmatrix}, C^2 = CC = \begin{bmatrix} 1 & 12 \\ -4 & 1 \end{bmatrix}$$

and AC, CB, A^2, and B^2 are undefined.

III.11.1.10. It is obvious from this example that AB and BA are usually not the same. In fact, one of these products can be defined and the other undefined, and even if both are defined they need not even be the same size. Even if A and B are square matrices of the same size, so AB and BA are both defined and the same size, they need not be equal (in almost any example one tries randomly they will be different). Thus matrix multiplication is *noncommutative* in a very fundamental sense.

It may then be surprising that the other usual rules of algebra hold for matrix multiplication:

III.11.1.11. PROPOSITION. Matrix multiplication, when defined, is associative: if A, B, C are matrices and $(AB)C$ is defined, then $A(BC)$ is also defined and $A(BC) = (AB)C$.

PROOF: This is a straightforward calculation. The products are defined if and only if A is $m \times n$, B is $n \times p$, and C is $p \times q$ for some $m, n, p, q \in \mathbb{N}$. In this case, for $1 \leq i \leq m$, $1 \leq j \leq q$ we have

$$[(AB)C]_{ij} = \sum_{r=1}^{p} (AB)_{ir} C_{rj} = \sum_{r=1}^{p} \left(\sum_{k=1}^{n} A_{ik} B_{kr} \right) C_{rj} = \sum_{r=1}^{p} \sum_{k=1}^{n} A_{ik} B_{kr} C_{rj}$$

$$= \sum_{k=1}^{n} \sum_{r=1}^{p} A_{ik} B_{kr} C_{rj} = \sum_{k=1}^{n} A_{ik} \left(\sum_{r=1}^{p} B_{kr} C_{rj} \right) = \sum_{k=1}^{n} A_{ik} (BC)_{kj} = [A(BC)]_{ij} \ .$$

III.11.1.12. PROPOSITION. The distributive laws hold for matrix addition and multiplication:

(i) If A, B, C are matrices with B and C the same size, and AB (and hence also AC) is defined, then $A(B + C)$ is defined and $A(B + C) = AB + AC$.

(ii) If A, B, C are matrices with A and B the same size, and AC (and hence also BC) is defined, then $(A + B)C$ is defined and $(A + B)C = AC + BC$.

The proof is a straightforward calculation. Note that both distributive laws must be stated and proved because of the noncommutativity of matrix multiplication.

The Identity Matrix

III.11.1.13. For $n \in \mathbb{N}$, define the *identity matrix of size* n to be the (square) $n \times n$ matrix

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

i.e. the (i, j) entry is 1 if $i = j$ and 0 if $i \neq j$.

A simple calculation shows that the identity matrices are "multiplicative identities":

III.11.1.14. PROPOSITION. Let A be an $m \times n$ matrix. Then $I_m A = A I_n = A$.

Rings of Square Matrices

III.11.1.15. If $n \in \mathbb{N}$, the set $M_n(\mathbb{F})$ of $n \times n$ matrices over $\mathbb{F}$ is closed under addition and multiplication, and the ring axioms () are satisfied. Since $I_n \in M_n(\mathbb{F})$, $M_n(\mathbb{F})$ is a unital ring under this addition and multiplication; this ring is noncommutative if $n > 1$. More generally, if R is any ring, then the set $M_n(R)$ of $n \times n$ matrices over R is a ring, which is unital if R is unital, and noncommutative if R is unital and $n > 1$.

Invertible Matrices

III.11.1.16. If A is an $m \times n$ matrix and B is $n \times m$, and $AB = I_m$, then A is a *left inverse* for B and B is a *right inverse* for A, and A and B are *right* and *left invertible* respectively. It turns out that this is possible only if $n \geq m$ (). If $n = m$, left and right inverses, if they exist, are unique and are actually two-sided inverses, i.e. if A and B are $n \times n$ matrices and $AB = I_n$, then automatically $BA = I_n$ also (). (These results are valid only for matrices over commutative rings.) A square matrix A with a (necessarily two-sided) inverse is called *invertible*, or *nonsingular*, and its inverse is denoted A^{-1}.

III.11.1.17. Examples. The identity matrix I_n is invertible, with $I_n^{-1} = I_n$. The zero matrix is obviously not invertible.

Not every nonzero square matrix is invertible: a simple calculation shows, for example, that

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

is not invertible.

We will later () give criteria for when a square matrix is invertible.

Transpose

III.11.1.18. If A is an $m \times n$ matrix, the *transpose* of A, denoted A^t, is the $n \times m$ matrix obtained by interchanging the rows and columns of A, i.e. if $A = [a_{ij}]_{mn}$, then $A^t = [a_{ji}]_{nm}$. The transpose of a row vector is a column vector, and conversely. If A is any matrix, then $[A^t]^t = A$. If A and B are matrices of the same size, then $(A + B)^t = A^t + B^t$.

Transpose reverses multiplication:

III.11.1.19. Proposition. Let A be $m \times n$ and B $n \times p$. Then $(AB)^t = B^t A^t$.

The proof is a simple calculation left to the reader. (This result is only valid over a commutative ring.)

Since $I_n^t = I_n$, we have:

III.11.1.20. Corollary. If A is an $n \times n$ matrix, then A^t is invertible if and only if A is invertible, and $(A^t)^{-1} = (A^{-1})^t$.

Multiplication by Blocks

III.11.1.21. A very useful symbolic procedure for multiplying matrices, frequently overlooked in linear algebra texts, is by decomposition into blocks.

Let $m, n, m_1, m_2, n_1, n_2 \in \mathbb{N}$ with $m = m_1 + m_2$, $n = n_1 + n_2$. If A is an $m \times n$ matrix over $\mathbb{F}$, then A can be decomposed or partitioned symbolically as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{ij} is an $m_i \times n_j$ matrix (A_{11} and A_{12} comprise the first m_1 rows of A, etc.) Sometimes we draw dotted lines in an array to denote a partition of the matrix. (We have described this in the case where the rows and columns are each partitioned into two blocks, but any finite numbers of blocks, including one, can be used independently for the rows and columns.)

The proof of the next key proposition is a straightforward but ugly calculation along the lines of the proof of III.11.1.11..

III.11.1.22. PROPOSITION. Let $m, n, p, m_1, m_2, n_1, n_2, p_1, p_2 \in \mathbb{N}$ with $m = m_1 + m_2$, $n = n_1 + n_2$, $p = p_1 + p_2$. Let A be an $m \times n$ matrix and B an $n \times p$ matrix, and set $C = AB$. Partition A, B, C according to the m_i, n_i, p_i. Then for any i and j,

$$C_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j}$$

(C_{ij} is an $m_i \times p_j$ matrix, and so are the two terms on the right). A similar result holds for partitions into other numbers of blocks (provided only that the columns of A and the rows of B are partitioned in the same way).

Symbolically, AB can be computed by multiplying

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

according to the usual rule for multiplying two 2×2 matrices, i.e. the matrix product can be "computed by blocks."

If the matrices are square and all blocks are the same size, as a special case we obtain:

III.11.1.23. COROLLARY. If $m, n \in \mathbb{N}$, then $M_m(M_n(\mathbb{F}))$ is naturally isomorphic to $M_{mn}(\mathbb{F})$.

This result holds over any ring. For fields (or commutative rings), it can also be proved by tensor products ().

The technique of multiplication by blocks is especially useful if one of the matrices has a rectangular block of zeroes. Here is a sample application, which will be used in the proof of the Implicit Function Theorem of vector calculus ():

III.11.1.24. PROPOSITION. Let $n, n_1, n_2 \in \mathbb{N}$ with $n = n_1 + n_2$, A an $n_1 \times n_1$ matrix, and B an $n_1 \times n_2$ matrix. Then the $n \times n$ matrix

$$C = \begin{bmatrix} A & B \\ 0 & I \end{bmatrix}$$

(where the 0 is an $n_2 \times n_1$ block and I is $n_2 \times n_2$ identity matrix) is invertible if and only if A is invertible, and in this case

$$C^{-1} = \begin{bmatrix} A^{-1} & -A^{-1}B \\ 0 & I \end{bmatrix}.$$

PROOF: It is easily checked by multiplication by blocks that if A is invertible, then the formula gives an inverse for C. Conversely, suppose C is invertible and

$$C^{-1} = \begin{bmatrix} X & Y \\ Z & W \end{bmatrix}$$

with blocks of the same size as in C. We then have

$$\begin{bmatrix} A & B \\ 0 & I \end{bmatrix} \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} AX + BZ & AY + BW \\ Z & W \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

and comparing entries, we get $Z = 0$, $W = I$, $X = A^{-1}$, $Y = -A^{-1}B$. 👍

Systems of Linear Equations

As an example of the utility of using matrix notation, we briefly review the theory of systems of linear equations.

III.11.1.25. A system of m linear equations in n unknowns over $\mathbb{F}$ is of the form

$$\begin{array}{c} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{array}$$

where the a_{ij} and the b_j are given elements of $\mathbb{F}$, and $x_1, \ldots, x_n$ are unknowns. The system is *homogeneous* if all the b_j are 0. A *solution* is, of course, a set of values for the x_i in $\mathbb{F}$ satisfying the equations.

III.11.1.26. This system can be conveniently abbreviated in matrix form. If

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

we can then compactly represent the system as the matrix equation

$$A\boldsymbol{x} = \boldsymbol{b}$$

where A is a given $m \times n$ matrix, $\boldsymbol{b}$ a given column m-vector, and $\boldsymbol{x}$ an unknown column n-vector. This clean formulation alone gives some justification for why matrix multiplication as we have defined it is useful. But this is only the beginning of the utility.

III.11.1.27. We would expect to have a unique solution to the system if and only if $m = n$; if $m < n$ the system is "underdetermined" and we would expect many solutions, and if $m > n$ we would expect that the system is "overdetermined" and that there is no solution in general. These answers are "usually" correct, but can fail if there are degeneracies in the system.

In the case $m = n$, if the matrix A is invertible, the matrix form gives a way to find a solution (which turns out to be the unique solution for any $\boldsymbol{b}$):

$$\boldsymbol{x} = A^{-1}\boldsymbol{b} \ .$$

We will examine the general theory after developing some linear algebra machinery (III.11.7.).

III.11.2. Vector Spaces

A vector space is simply an abstraction of Euclidean space $\mathbb{R}^n$ with its essential algebraic structure. The two basic algebraic operations in $\mathbb{R}^n$ are vector addition and scalar multiplication. Other algebraic operations such as the dot product will not be used at this point, but will be incorporated later.

Let $\mathbb{F}$ be a field (usually $\mathbb{R}$ or $\mathbb{C}$ in analysis applications), called the *field of scalars*, which is fixed throughout the discussion. A vector space over $\mathbb{F}$ will be a set equipped with an addition and a scalar multiplication satisfying a list of natural algebraic properties. Specifically:

III.11.2.1. Definition. A *vector space over* $\mathbb{F}$ is a set $\mathcal{V}$ with two binary operations $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ (written $+(x, y) = x + y$) and $\cdot : \mathbb{F} \times \mathcal{V} \to \mathcal{V}$ (written $\cdot(\alpha, x) = \alpha \cdot x$ or usually just αx) satisfying:

(A1) The binary operation $+$ is *associative*: $x + (y + z) = (x + y) + z$ for all $x, y, z \in \mathcal{V}$.

(A2) The binary operation $+$ is *commutative*: $x + y = y + x$ for all $x, y \in \mathcal{V}$.

(A3) There is an element $\mathbf{0} \in \mathcal{V}$ such that $\mathbf{0} + x = x$ for all $x \in \mathcal{V}$.

(A4) For every $x \in \mathcal{V}$ there is an element $(-x) \in \mathcal{V}$ with $x + (-x) = \mathbf{0}$.

(M1) The binary operation $\cdot$ is *associative*: $\alpha(\beta x) = (\alpha\beta)x$ for all $\alpha, \beta \in \mathbb{F}$, $x \in \mathcal{V}$.

(M2) The binary operation $\cdot$ is *distributive*: $\alpha(x + y) = \alpha x + \alpha y$ and $(\alpha + \beta)x = \alpha x + \beta x$ for all $\alpha, \beta \in \mathbb{F}$, $x, y \in \mathcal{V}$.

(M3) $1 \cdot x = x$ for all $x \in \mathcal{V}$.

If R is a ring, a set with two operations satisfying these axioms is called a *left R-module*. Thus a vector space is just a left module over a field.

III.11.2.2. Examples. (i) If $n \in \mathbb{N}$, then the set $\mathbb{F}^n$ of n-tuples of elements of $\mathbb{F}$, with coordinatewise addition and scalar multiplication, is a vector space over $\mathbb{F}$. This is the prototype example. It will turn out that any finite-dimensional vector space over $\mathbb{F}$ is essentially identical to this example for a unique n.

(ii) Let $m, n \in \mathbb{N}$. The set $M_{m,n}(\mathbb{F})$ of $m \times n$ matrices over $\mathbb{F}$, with matrix addition and scalar multiplication, is a vector space over $\mathbb{F}$. This vector space has a natural identification with $\mathbb{F}^{mn}$ by writing the entries of the matrix in a single string in a fixed order.

(iii) Here is a general example including (i) and (ii) as a special case: let X be a set and $\mathcal{V}$ the set of all functions from X to $\mathbb{F}$. Define addition and scalar multiplication pointwise: if $f, g \in \mathcal{V}$ and $\alpha \in \mathbb{F}$, set $(f+g)(x) = f(x)+g(x)$ and $(\alpha f)(x) = \alpha f(x)$ (where the operations on the right-hand sides are the operations in $\mathbb{F}$). Example (i) is the case where $X = \{1, \ldots, n\}$ and example (ii) is the case $X = \{1, \ldots, m\} \times \{1, \ldots, n\}$.

(iv) Let $\mathcal{P}$ be the set of all polynomials with coefficients in $\mathbb{F}$, expressions of the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

where the $a_k \in \mathbb{F}$ and the n may vary from polynomial to polynomial. Add polynomials formally in the usual way by collecting together terms with the same power of x, and multiply by scalars termwise. Then $\mathcal{P}$ is a vector space over $\mathbb{F}$.

(v) Let $[a, b]$ be an interval in $\mathbb{R}$, and $C([a, b])$ the set of continuous functions from $[a, b]$ to $\mathbb{R}$. Add functions and multiply by scalars pointwise in the usual way, as in (iii) (a theorem () is needed to show that $C([a, b])$ is closed under these operations). Then $C([a, b])$ is a vector space over $\mathbb{R}$.

Example (v) has many variations which are important in analysis.

III.11.2.3. In some of these examples, there is a natural multiplication defined on pairs of elements of the vector space which is compatible (bilinear) with scalar multiplication. This multiplication is not part of the vector space structure. A vector space over $\mathbb{F}$ with a compatible multiplication is called an *algebra* over $\mathbb{F}$. $M_n(\mathbb{F})$ and $\mathcal{P}$ are typical examples of algebras over $\mathbb{F}$. $C([a, b])$ is an algebra over $\mathbb{R}$.

III.11.2.4. There are two important points which must be emphasized about the definition of a vector space. First, the operations of addition and scalar multiplication are part of the definition and must be specified; thus it is technically incorrect to make a statement such as "Let $\mathcal{V}$ be a vector space over $\mathbb{F}$." But we do this all the time when the operations in $\mathcal{V}$ are understood, such as in these examples or variations of them.

Secondly, the field of scalars must be specified; it is incorrect to say "Let $\mathcal{V}$ be a vector space." (We often do this too, but the field of scalars must be understood.) There is a variation of this point which must be kept in mind: if $\mathcal{V}$ is a vector space over $\mathbb{C}$, then since $\mathbb{R} \subseteq \mathbb{C}$ we can restrict scalar multiplication to $\mathbb{R}$ and obtain a real vector space (vector space over $\mathbb{R}$). Some references thus state:

> "A complex vector space is a real vector space."

From a logical standpoint, however, this statement is meaningless: real vector spaces and complex vector spaces are distinct mathematical structures and it does not make sense to mix them. A more careful and logically correct way of phrasing this statement is:

> "A complex vector space has a unique underlying structure as a real vector space."

The situation is similar to the corresponding relation between metric spaces and topological spaces (). The same considerations apply to any other field and subfield (e.g. $\mathbb{Q} \subseteq \mathbb{R}$).

III.11.2.5. Elements of a vector space are typically called *points* or *vectors*. This terminology does not carry with it any necessary geometric interpretation. In fact, general vector spaces do not have any natural geometric interpretation. But the quasi-geometric language used is well established, and can be useful in motivating results, although care must be exercised in proofs not to use anything beyond what simply follows from the axioms.

The standard algebraic properties similar to those of () are consequences of the axioms. Here are some of the most important ones:

III.11.2.6. PROPOSITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. Then

(i) $0x = \mathbf{0}$ for all $x \in \mathcal{V}$.

(ii) If $\alpha \in \mathbb{F}$ and $x \in \mathcal{V}$, then $\alpha x = \mathbf{0}$ if and only if $\alpha = 0$ or $x = \mathbf{0}$.

(iii) If $x, y, z \in \mathcal{V}$ and $x + z = y + z$, then $x = y$.

(iv) If $\alpha \in \mathbb{F}$, $\alpha \neq 0$, and $x, y \in \mathcal{V}$, and $\alpha x = \alpha y$, then $x = y$.

(v) $(-1)x = -x$ for all $x \in \mathcal{V}$.

Isomorphism of Vector Spaces

As is usual in mathematics, we need a precise specification of when two vector spaces are abstractly the "same." The obvious criterion is:

III.11.2.7. DEFINITION. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$. A function $T : \mathcal{V} \to \mathcal{W}$ is an *isomorphism* (of vector spaces over $\mathbb{F}$) if T is a bijection which is *linear*, i.e. respects addition and scalar multiplication:

$$T(x + y) = T(x) + T(y)$$
$$T(\alpha x) = \alpha T(x)$$

for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{F}$. $\mathcal{V}$ and $\mathcal{W}$ are *isomorphic* (as vector spaces over $\mathbb{F}$) if there is an isomorphism from $\mathcal{V}$ to $\mathcal{W}$.

III.11.2.8. If $T : \mathcal{V} \to \mathcal{W}$ is an isomorphism, then $T^{-1} : \mathcal{W} \to \mathcal{V}$ is automatically linear, hence an isomorphism; thus the situation is symmetric in $\mathcal{V}$ and $\mathcal{W}$, i.e. the relation of isomorphism is symmetric. A composition of two isomorphisms (when defined) is an isomorphism, so the isomorphism relation is transitive. And the identity map on any vector space is an isomorphism, so the isomorphism relation is reflexive. Thus isomorphism is an equivalence relation on vector spaces over $\mathbb{F}$.

Vector Subspaces

III.11.2.9. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. A subset $\mathcal{W}$ of $\mathcal{V}$ is a *vector subspace* or *linear subspace* (often just called a *subspace*) if $\mathcal{W}$ is itself a vector space under the same operations of $\mathcal{V}$.

If $\mathcal{W}$ is a vector subspace of $\mathcal{V}$, the definition entails that $\mathcal{W}$ is closed under addition and scalar multiplication, i.e. if $x, y \in \mathcal{W}$ and $\alpha \in \mathbb{F}$, then $x + y$ and αx are in $\mathcal{W}$. $\mathcal{W}$ is also nonempty since $\mathbf{0} \in \mathcal{W}$. These are sufficient:

III.11.2.10. PROPOSITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $\mathcal{W} \subseteq \mathcal{V}$. Then $\mathcal{W}$ is a vector subspace of $\mathcal{V}$ if and only if $\mathcal{W}$ is nonempty and closed under addition and scalar multiplication.

PROOF: Necessity has been observed. Conversely, suppose $\mathcal{W}$ is nonempty and closed under addition and scalar multiplication. Axioms (A1), (A2), and (M1)–(M3) are automatic in $\mathcal{W}$ since they hold in $\mathcal{V}$. If $x \in \mathcal{W}$, then $0x = \mathbf{0} \in \mathcal{W}$, so (A3) holds. And if $x \in \mathcal{W}$, then $-x = (-1)x \in \mathcal{W}$, so (A4) holds. 👍

III.11.2.11. EXAMPLES. (i) If $\mathcal{V}$ is any vector space, then $\mathcal{V}$ and $\{\mathbf{0}\}$ are vector subspaces of $\mathcal{V}$. These are called the *trivial subspaces*.

(ii) Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $x \in \mathcal{V}$. Then

$$\{\alpha x : \alpha \in \mathbb{F}\}$$

is a vector subspace of $\mathcal{V}$. If $x \neq \mathbf{0}$, this is called the *one-dimensional subspace of $\mathcal{V}$ spanned by x*. This example will be generalized in III.11.2.14..

(iii) Let $\mathcal{V}$ be the set of all functions from $[a, b]$ to $\mathbb{R}$. Then $\mathcal{V}$ is a real vector space as in III.11.2.2.(iii). Then $C([a, b])$ (III.11.2.2.(v)) is a vector subspace of $\mathcal{V}$ ().

Span of a Subset, Linear Combinations

III.11.2.12. It is obvious that an arbitrary intersection of vector subspaces of a vector space $\mathcal{V}$ over $\mathbb{F}$ is a vector subspace (the intersection is nonempty since every subspace contains $\mathbf{0}$). Thus, if S is any subset of $\mathcal{V}$, the intersection of all subspaces of $\mathcal{V}$ containing S is a subspace of $\mathcal{V}$ which is the smallest subspace of $\mathcal{V}$ containing S. (Note that there is at least one subspace of $\mathcal{V}$ containing S, $\mathcal{V}$ itself.) This smallest subspace is called the *span* of S, denoted $span(S)$. We say S *spans* $\mathcal{V}$ if $span(S) = \mathcal{V}$.

There is a concrete description of $span(S)$ using the notion of linear combination:

III.11.2.13. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $x_1, \ldots, x_n \in \mathcal{V}$. A *linear combination* ($\mathbb{F}$-*linear combination*) of $\{x_1, \ldots, x_n\}$ is a vector of the form

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

where the α_k are in $\mathbb{F}$.

If S is an arbitrary nonempty subset of $\mathcal{V}$, then a *linear combination of S* is a vector which is a linear combination of a set of vectors in S. The set of linear combinations of $\emptyset$ is defined to be $\{\mathbf{0}\}$.

Note that even if S is infinite, a linear combination of S is a linear combination of a finite subset of S, i.e. an individual linear combination involves only finitely many elements of S.

III.11.2.14. PROPOSITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $S \subseteq \mathcal{V}$. Then $span(S)$ is precisely the set $Lin(S)$ of linear combinations of S.

PROOF: Any vector in S is trivially in $Lin(S)$, and every linear combination of S is in any vector subspace containing S. Thus $S \subseteq Lin(S) \subseteq span(S)$. So it suffices to note that $Lin(S)$ is a vector subspace of $\mathcal{V}$. It is obvious that $\mathbf{0} \in Lin(S)$, and that a sum or scalar multiple of linear combinations of S is again a linear combination of S. 👍

III.11.2.15. PROPOSITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $S \subseteq \mathcal{V}$. If $x_0 \in S$ and x_0 can be written as a linear combination of other vectors in S, then $span(S \setminus \{x_0\}) = span(S)$.

PROOF: Suppose $x_0 = \alpha_1 x_1 + \cdots + \alpha_n x_n$ with $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$ and $x_1, \ldots, x_n \in S \setminus \{x_0\}$. Let $x \in span(S)$. Then $x = \beta_1 y_1 + \cdots + \beta_m y_m$ for some $y_1, \ldots, y_m \in S$ and $\beta_1, \ldots, \beta_m \in \mathbb{F}$. By collecting together terms, we may assume all the y_i are distinct. If none of the y_i are equal to x_0, then x is already written as a linear combination of $S \setminus \{x_0\}$. If $y_i = x_0$ for some i, by reordering we may assume $i = 1$. Then

$$x = \beta_1(\alpha_1 x_1 + \cdots + \alpha_n x_n) + \beta_2 y_2 + \cdots + \beta_m y_m$$

and x is again in $span(S \setminus \{x_0\})$. 👍

III.11.2.16. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. $\mathcal{V}$ is *finite-dimensional* if there is a finite subset of $\mathcal{V}$ which spans $\mathcal{V}$.

III.11.2.17. EXAMPLES. (i) $\mathbb{F}^n$ is spanned by $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_k = (0, \ldots, 0, 1, 0, \ldots, 0)$, with the 1 in the k'th place. (This set of vectors is called the *standard basis* of $\mathbb{F}^n$). Thus $\mathbb{F}^n$ is finite-dimensional.

(ii) The vector space $\mathcal{P}$ of all polynomials over $\mathbb{F}$ (III.11.2.2.(iv) is not finite-dimensional, since a finite set of polynomials has a maximal degree and no linear combination of the finite set of polynomials can have greater degree. The countable set $\{1, x, x^2, \ldots\}$ does span $\mathcal{P}$.

It is a little harder to show that Example III.11.2.2.(v), and III.11.2.2.(iii) for infinite X, are not finite-dimensional (this follows from III.11.2.30.(iv) in both cases; cf. III.11.2.21.(iv)).

Linear Independence

III.11.2.18. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and S a subset of $\mathcal{V}$. Then S is *linearly independent* (over $\mathbb{F}$) if no vector in S is a linear combination of the other vectors in S.

In other words, S is linearly independent if it has no "redundancy," i.e. no proper subset of S has the same span as S; every vector in S makes an essential contribution to the span. Since the zero vector is in the span of $\emptyset$, no linearly independent set can contain the zero vector.

The empty set is vacuously linearly independent.

A convenient rephrasing is:

III.11.2.19. PROPOSITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $S \subseteq \mathcal{V}$. The following are equivalent:

(i) S is linearly independent.

(ii) Whenever $x_1, \ldots, x_n \in S$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$, and

$$\alpha_1 x_1 + \cdots + \alpha_n x_n = \mathbf{0}$$

then $\alpha_1 = \cdots = \alpha_n = 0$.

(iii) Every nonzero vector in $span(S)$ can be written as a linear combination of vectors in S with nonzero coefficients in a unique way.

The uniqueness in (iii) must be expressed carefully; additional terms with coefficients 0 may be added to the sum without changing the vector represented.

PROOF: (i) $\Rightarrow$ (ii): Suppose $x_1, \ldots, x_n \in S$, $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$, and $\alpha_1 x_1 + \cdots + \alpha_n x_n = \mathbf{0}$ with not all $\alpha_k = 0$. By reordering, we may assume $\alpha_1 \neq 0$. Then we have

$$x_1 = -\frac{\alpha_2}{\alpha_1} x_2 - \cdots - \frac{\alpha_n}{\alpha_1} x_n$$

so S is not linearly independent.

(ii) $\Rightarrow$ (iii): Suppose, for some $x \in span(S)$,

$$x = \alpha_1 x_1 + \cdots + \alpha_n x_n = \beta_1 x_1 + \cdots + \beta_n x_n$$

with $\alpha_k \neq \beta_k$ for some k. Then

$$(\alpha_1 - \beta_1) x_1 + \cdots + (\alpha_n - \beta_n) x_n = \mathbf{0}$$

with $\alpha_k - \beta_k \neq 0$, so (ii) does not hold.

(iii) $\Rightarrow$ (i): If $x \in S$ were a linear combination $x = 1x = \alpha_1 x_1 + \cdots + \alpha_n x_n$ with $x_k \neq x$ for all k, then x would be a linear combination of vectors in S in a nonunique way. 👍

III.11.2.20. A sum $\alpha_1 x_1 + \cdots + \alpha_n x_n = \mathbf{0}$ as in (ii) with $x_1, \ldots, x_n \in S$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$ not all zero is called a *linear dependence* (or *linear relation*) in S (over $\mathbb{F}$), and a set with a linear dependence (i.e. a set which is not linearly independent) is called *linearly dependent*.

III.11.2.21. EXAMPLES. (i) A set $\{x\}$ of one vector in a vector space $\mathcal{V}$ is linearly independent if and only if $x \neq \mathbf{0}$.

(ii) If $\{x, y\}$ is a subset of $\mathcal{V}$ with $x \neq y$, then $\{x, y\}$ is linearly independent in $\mathcal{V}$ if and only if x is nonzero and y is not a scalar multiple of x. (The condition is symmetric in x and y.)

(iii) It is harder to tell by inspection whether a set of three or more vectors is linearly independent, since the set can be linearly dependent even if no vector in the set is a scalar multiple of another. For example, it is easily seen that

$$\{(1,0), (0,1), (1,1)\}$$

is linearly dependent in $\mathbb{R}^2$ since the third vector is the sum of the first two.

(iv) In Example III.11.2.2.(iv) or (v), let $f_n(x) = x^n$. It is easy to verify that $\{f_n : n \in \mathbb{N}\}$ is linearly independent in each case. In case (iv) this set spans the vector space, and in (v) it does not.

Bases

III.11.2.22. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. A *basis* for $\mathcal{V}$ (over $\mathbb{F}$) is a linearly independent set of vectors in $\mathcal{V}$ which spans $\mathcal{V}$.

A basis is a set of vectors which is "just large enough" to span $\mathcal{V}$ yet small enough to contain no redundancies. If B is a basis for $\mathcal{V}$, then no smaller subset of $\mathcal{V}$ spans $\mathcal{V}$ and no larger subset of $\mathcal{V}$ is linearly independent, i.e. B is a maximal linearly independent subset of $\mathcal{V}$.

The plural of *basis* is *bases*. The plural of *base* is spelled the same way but pronounced differently.

III.11.2.23. Bases are especially important in finite-dimensional vector spaces, where they can be regarded as giving a "coordinate system." They are of less crucial importance in infinite-dimensional vector spaces. Certain infinite-dimensional topological vector spaces have other basis-like sets (where "linear combinations" involve infinite series) which are a more useful substitute (). In infinite-dimensional vector spaces a basis in the linear algebra sense (i.e. in the sense of III.11.2.22.) is often called a *Hamel basis* to distinguish the concept from the other one.

III.11.2.24. PROPOSITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. A subset B of $\mathcal{V}$ is a basis for $\mathcal{V}$ if and only if B is a maximal linearly independent subset.

PROOF: It has already been observed that a basis is a maximal linearly independent subset. On the other hand, suppose B is a maximal linearly independent set in $\mathcal{V}$, and $x \in \mathcal{V}$. If $x \notin span(B)$, then $S \cup \{x\}$ is linearly independent, contradicting maximality of S. Thus $span(B) = \mathcal{V}$ and B is a basis for $\mathcal{V}$. 👍

III.11.2.25. EXAMPLES. (i) Let $\mathcal{V} = \mathbb{F}^n$ as a vector space over $\mathbb{F}$. Then the set of standard basis vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is a basis for $\mathbb{F}^n$: if $\mathbf{x} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{F}^n$, then

$$(\alpha_1, \ldots, \alpha_n) = \alpha_1 \mathbf{e}_1 + \cdots + \alpha_n \mathbf{e}_n$$

is the unique expression of $\mathbf{x}$ as a linear combination of $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$.

(ii) One can give completely different bases for $\mathbb{F}^n$. For example, a simple calculation shows that

$$\mathbf{x}_1 = (1, -1, 0, \ldots, 0)$$

$$\mathbf{x}_2 = (0, 1, -1, \ldots, 0)$$

$$\cdots$$

$$\mathbf{x}_{n-1} = (0, 0, \ldots, 1, -1)$$

$$\mathbf{x}_n = (-1, 0, 0, \ldots, 0, 1)$$

forms a basis for $\mathbb{F}^n$. (There are also *many* other bases, especially if $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$.)

(iii) The set $\{f_n : n \in \mathbb{N}\}$ (III.11.2.21.(iv)) is a basis for the vector space $\mathcal{P}$ over $\mathbb{R}$ (III.11.2.2.(iv)).

III.11.2.26. Examples (i) and (ii) show that there is nothing unique about bases for a vector space; two bases may have no vectors in common. But it turns out (III.11.2.28.(iv)) that the *number* of vectors in any two bases for a vector space $\mathcal{V}$ is always the same; this number is called the *dimension* of $\mathcal{V}$.

III.11.2.27. Elaborating further on Example III.11.2.25., if $\mathcal{V}$ is a vector space over $\mathbb{F}$ with basis $B = \{x_1, \ldots, x_n\}$, then the (ordered) basis B induces a natural one-one correspondence with $\mathbb{F}^n$:

$$(\alpha_1, \ldots, \alpha_n) \longleftrightarrow \alpha_1 x_1 + \cdots + \alpha_n x_n$$

which respects vector addition and scalar multiplication (a *vector space isomorphism*). Thus an ordered basis for $\mathcal{V}$ can be regarded as a "coordinate system" in $\mathcal{V}$. Indeed, a finite-dimensional abstract vector space over $\mathbb{F}$ can just be regarded as a copy of $\mathbb{F}^n$ for some n with the "natural" coordinate system stripped away; choice of a basis specifies a coordinate system, with any such coordinate system as good as any other for most linear algebra purposes (if a linear operator is being considered, some coordinate systems may be more compatible with the operator than others; this will be discussed later).

> "$\mathbb{R}^n$ (or $\mathbb{C}^n$) is not a vector space: it is a vector space *plus* a basis."
>
> *J. Renault*

Fundamental Theorems of Linear Algebra

It is not obvious that every vector space has a basis (in fact, in general, the Axiom of Choice is needed to prove this). The following blockbuster theorem shows that the situation is as nice as could be hoped for:

III.11.2.28. THEOREM. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. Then

(i) $\mathcal{V}$ has a basis.

(ii) Every linearly independent subset of $\mathcal{V}$ can be expanded to a basis for $\mathcal{V}$.

(iii) Every spanning subset of $\mathcal{V}$ contains a basis for $\mathcal{V}$.

(iv) Any two bases for $\mathcal{V}$ have the same cardinality (if $\mathcal{V}$ is finite-dimensional, any two bases for $\mathcal{V}$ have the same finite number of elements).

This theorem is elementary to prove if $\mathcal{V}$ is finite-dimensional, but more delicate (requiring the AC) in the general case. We give the finite-dimensional proof here and defer the general proof until III.11.2.32..
PROOF: (Finite-dimensional case.) Clearly (ii) $\Rightarrow$ (i) (begin with $\emptyset$). To (partly) prove (ii), let $\{y_1, \ldots, y_m\}$ be a finite set which spans $\mathcal{V}$, and let $\{x_1, \ldots, x_r\}$ be a finite linearly independent set in $\mathcal{V}$. Set

$$S_1 = \{x_1, \ldots, x_r, y_1, \ldots, y_m\}$$

regarded as an ordered set. S_1 spans $\mathcal{V}$, and is linearly dependent since each x_i can be written as a linear combination of the y_j. Working from left to right, let z_1 be the first vector which can be written as a linear combination of the preceding ones. There must be such a z_1, and it must be one of the y_j since the x_i are linearly independent. Set $S_2 = S_1 \setminus \{z_1\}$. Then $span(S_2) = span(S_1) = \mathcal{V}$ by III.11.2.15.. If S_2 is

linearly independent, it is a basis containing $\{x_1, \ldots, x_r\}$; otherwise, repeat the process, letting z_2 be the first vector in the ordered set S_2 which can be written as a linear combination of the preceding vectors, and $S_3 = S_2 \setminus \{z_2\}$. Since S_1 is finite, we must reach a linearly independent set in a finite number of steps; this set will be a basis for $\mathcal{V}$ consisting of $x_1, \ldots, x_r$ and (in general) some, but not all, of the y_j. This argument proves (ii) for a *finite* linearly independent set, and thus the special case (i).

The argument for (iv) is similar. We will show a stronger statement: if $\{y_1, \ldots, y_n\}$ is a basis for $\mathcal{V}$, and $\{x_1, \ldots, x_r\}$ is a linearly independent set in $\mathcal{V}$, then $r \leq n$. This will show (iv) by symmetry, and also complete the proof of (ii) since it implies that any linearly independent set in $\mathcal{V}$ is finite. Suppose $r > n$. Set $S_1 = \{x_1, y_1, \ldots, y_n\}$, regarded as an ordered set. This set spans $\mathcal{V}$ and is linearly dependent. Let z_1 be the first vector which can be written as a linear combination of the previous ones; then z_1 must be one of the y_j. Set

$$S_2 = \{x_2, x_1, y_1, \ldots, y_n\}$$

with $z_1 = y_j$ removed. Then $span(S_2) = span(S_1) = \mathcal{V}$, and S_2 is linearly dependent since $x_2 \in span(S_1 \setminus \{z_1\}) = \mathcal{V}$. Repeat the process. At each stage another x_k is added and a y_j is removed; thus after n steps we have $S_n \setminus \{z_n\} = \{x_n, \ldots, x_1\}$, and this set spans $\mathcal{V}$, contradicting that x_{n+1} is not a linear combination of $\{x_1, \ldots, x_n\}$. Thus $r \leq n$.

Finally, we show (iii). Let S be a set which spans $\mathcal{V}$, and $\{y_1, \ldots, y_n\}$ a basis for $\mathcal{V}$. Each y_j can be written as a linear combination of a finite number of vectors in S; hence there is a finite subset of S whose span includes all the y_j, hence is $\mathcal{V}$. Thus we may assume S is finite. List the elements of S and successively eliminate any vector which can be written as a linear combination of the preceding ones. Each reduced set will still span $\mathcal{V}$ by III.11.2.15.. In a finite number of steps a linearly independent set will be reached, which is a basis for $\mathcal{V}$. 👍

III.11.2.29. Definition. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. The *(linear) dimension* of $\mathcal{V}$, denoted $dim(\mathcal{V})$, is the cardinality of a basis of $\mathcal{V}$.

The dimension of $\mathcal{V}$ is well defined by III.11.2.28.(iv), and is finite (a natural number) if and only if $\mathcal{V}$ is finite-dimensional in the sense of III.11.2.16.. We sometimes write $dim_{\mathbb{F}}(\mathcal{V})$ to emphasize the field we are working over (cf. III.11.2.33.).

We note the following corollary of the finite-dimensional version of III.11.2.28.:

III.11.2.30. Corollary. Let $\mathcal{V}$ be a finite-dimensional vector space over $\mathbb{F}$, with $dim(\mathcal{V}) = n$. Then

(i) If $\{x_1, \ldots, x_r\}$ is a linearly independent set of vectors in $\mathcal{V}$, then $r \leq n$.

(ii) If $S = \{x_1, \ldots, x_n\}$ is a linearly independent set of vectors in $\mathcal{V}$, then S spans $\mathcal{V}$ (i.e. is a basis for $\mathcal{V}$).

(iii) If $\{x_1, \ldots, x_r\}$ spans $\mathcal{V}$, then $r \geq n$.

(iv) If $S = \{x_1, \ldots, x_n\}$ is a set of vectors in $\mathcal{V}$ which spans $\mathcal{V}$, then S is linearly independent (i.e. is a basis for $\mathcal{V}$).

(v) If $\mathcal{W}$ is a vector subspace of $\mathcal{V}$, then $dim(\mathcal{W}) \leq dim(\mathcal{V})$; $dim(\mathcal{W}) = dim(\mathcal{V})$ if and only if $\mathcal{W} = \mathcal{V}$. In particular, any subspace of $\mathcal{V}$ is finite-dimensional.

PROOF: (i)–(iv) are obvious from III.11.2.28.. For (v), suppose $dim(\mathcal{W}) > n$ (or $\mathcal{W}$ is not finite-dimensional). Then (without any form of AC) we can choose a set $S = \{x_1, \ldots, x_{n+1}\}$ in $\mathcal{W}$ which is linearly independent [let x_1 be any nonzero vector in $\mathcal{W}$, and successively choose x_{k+1} to be a vector not in the span of $\{x_1, \ldots, x_k\}$]. Then S is a linearly independent subset of $\mathcal{V}$, contradicting (i). If $dim(\mathcal{W}) = n$, apply (ii) to a basis for $\mathcal{W}$. 👍

III.11.2.31. THEOREM. Two vector spaces over $\mathbb{F}$ are isomorphic if and only if they have the same dimension. In particular, any n-dimensional vector space over $\mathbb{F}$ is isomorphic to $\mathbb{F}^n$.

PROOF: Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$ of the same dimension, and let B and C be bases for $\mathcal{V}$ and $\mathcal{W}$ respectively. Let $\psi : B \to C$ be a bijection. Extend ψ to a bijection $\phi : \mathcal{V} \to \mathcal{W}$ as follows. If $x \in \mathcal{V}$, then x can be written uniquely as

$$x = \alpha_1 x_1 + \cdots + \alpha_n x_n$$

with $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$, all nonzero, and $x_1, \ldots, x_n \in B$. Define

$$\phi(x) = \alpha_1 \psi(x_1) + \cdots + \alpha_n \psi(x_n) \ .$$

Then ϕ is well defined and one-to-one since C is linearly independent. If $y \in \mathcal{W}$, then

$$y = \alpha_1 y_1 + \cdots + \alpha_m y_m$$

for some $\alpha_1, \ldots, \alpha_m \in \mathbb{F}$ and $y_1, \ldots, y_m \in C$; if $y_j = \psi(x_j)$ for each j, then $y = \phi(x)$, where

$$x = \alpha_1 x_1 + \cdots + \alpha_m x_m \ .$$

So ϕ is surjective. It is routine to check that ϕ is linear. 👍

III.11.2.32. We now prove the general version of Theorem III.11.2.28.. This proof requires the Axiom of Choice (in fact, (i)–(iii) may each be equivalent to AC; (iv) only needs a weaker version of Choice).
PROOF: We again begin by proving (ii), which clearly implies (i) as before. Let S be a linearly independent subset of $\mathcal{V}$. Let $\mathcal{X}$ be the collection of all linearly independent subsets of $\mathcal{V}$ containing S, ordered by inclusion. If $\{X_i : i \in I\}$ is a chain in $\mathcal{X}$, then $X = \cup_i X_i$ is a linearly independent subset of $\mathcal{V}$, since a linear dependence involves only finitely many vectors and any finite subset of X is contained in some X_i. Thus $X \in \mathcal{X}$, and X is an upper bound for $\{X_i : i \in I\}$. By Zorn's Lemma, $\mathcal{X}$ contains a maximal element B. As in III.11.2.24., B is a basis for $\mathcal{V}$ containing S.

The proof of (iii) is similar. Let S be a set which spans $\mathcal{V}$. Let $\mathcal{X}$ be the collection of all linearly independent subsets of S, ordered by inclusion. As before, Zorn's Lemma applies to give a maximal element B of $\mathcal{X}$. As in the proof of III.11.2.24., if $span(B)$ is not all of $\mathcal{V}$, it cannot contain all of S since S spans $\mathcal{V}$; if $x \in S \setminus span(B)$, then $B \cup \{x\}$ is linearly independent, contradicting maximality of B. Thus B is a basis for $\mathcal{V}$.

The proof of (iv) is more complicated. Suppose $\{x_i : i \in I\}$ and $\{y_j : j \in J\}$ are bases for $\mathcal{V}$. We may assume I and J are infinite since the finite case has already been proved. Let $\kappa = card(I)$ and $\lambda = card(J)$. Each x_i is a linear combination of finitely many y_j, and every y_j occurs in one of these finite sets, or otherwise all the x_i would be contained in the proper subspace of $\mathcal{V}$ spanned by a proper subset of the y_j. Thus $\{y_j : j \in J\}$ is a union of κ finite subsets, so $\lambda \leq \aleph_0 \kappa = \kappa$ since κ is infinite (II.9.9.4.). Symmetrically, $\kappa \leq \lambda$. Thus $\kappa = \lambda$ by the Schröder-Bernstein Theorem. 👍

III.11.2.33. When the same set is regarded as a vector space over more than one field, the notions of span, linear independence, basis, and dimension are dependent on the choice of base field. For example, $\mathbb{C}^n$ may be regarded as an (n-dimensional) vector space over $\mathbb{C}$, or as a vector space over $\mathbb{R}$ by restriction of scalar multiplication. The standard basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ is indeed a basis for $\mathbb{C}^n$ over $\mathbb{C}$. This set is linearly independent over $\mathbb{R}$, but not a basis for $\mathbb{C}^n$ over $\mathbb{R}$; in fact, its real span is the real vector subspace $\mathbb{R}^n$ of $\mathbb{C}^n$. On the other hand, the set

$$\{\mathbf{e}_1, \ldots, \mathbf{e}_n, i\mathbf{e}_1, \ldots, i\mathbf{e}_n\}$$

is linearly independent over $\mathbb{R}$ but not over $\mathbb{C}$. It is easily checked that this set is a basis for $\mathbb{C}^n$ over $\mathbb{R}$, and thus $dim_{\mathbb{R}}(\mathbb{C}^n) = 2n$.

Complementary Subspaces

III.11.2.34. Definition. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. Two subspaces $\mathcal{W}_1$ and $\mathcal{W}_2$ of $\mathcal{V}$ are *complementary* if $\mathcal{W}_1 \cap \mathcal{W}_2 = \{0\}$ and $span(\mathcal{W}_1 \cup \mathcal{W}_2) = \mathcal{V}$. $\mathcal{W}_2$ is complementary to $\mathcal{W}_1$ and vice versa.

This definition is symmetric in $\mathcal{W}_1$ and $\mathcal{W}_2$. But we often talk about subspaces complementary to a given subspace.

III.11.2.35. There is nothing remotely unique about complementary subspaces to a given subspace. For example, if $\mathcal{W}$ is any one-dimensional subspace of $\mathbb{R}^2$, then any one-dimensional subspace of $\mathbb{R}^2$ not equal to $\mathcal{W}$ is complementary to $\mathcal{W}$.

III.11.2.36. Proposition. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $\mathcal{W}_1$ and $\mathcal{W}_2$ subspaces of $\mathcal{V}$. Then $\mathcal{W}_1$ and $\mathcal{W}_2$ are complementary subspaces of $\mathcal{V}$ if and only if every vector $x \in \mathcal{V}$ can be uniquely written as $x = x_1 + x_2$, with $x_j \in \mathcal{W}_j$.

Proof: If $x \in span(\mathcal{W}_1 \cup \mathcal{W}_2)$, then x can be written as

$$x = \alpha_1 y_1 + \cdots + \alpha_n y_n + \beta_1 z_1 + \cdots + \beta_m z_m$$

with $y_1, \ldots, y_n \in \mathcal{W}_1$ and $z_1, \ldots, z_m \in \mathcal{W}_2$. If $x_1 = \alpha_1 y_1 + \cdots + \alpha_n y_n$ and $x_2 = \beta_1 z_1 + \cdots + \beta_m z_m$, then $x_j \in \mathcal{W}_j$ and $x = x_1 + x_2$. Thus every vector in $\mathcal{V}$ can be so written if and only if $span(\mathcal{W}_1 \cup \mathcal{W}_2) = \mathcal{V}$.
Suppose $\mathcal{W}_1 \cap \mathcal{W}_2 = \{0\}$, and $x = x_1 + x_2 = y_1 + y_2$ with $x_j, y_j \in \mathcal{W}_j$. Then $x_1 - y_1 = x_2 - y_2$; the left side is in $\mathcal{W}_1$ and the right side is in $\mathcal{W}_2$, so both sides are in $\mathcal{W}_1 \cap \mathcal{W}_2$. Hence both sides are zero and the representation of x is unique. Conversely, if $x \in \mathcal{W}_1 \cap \mathcal{W}_2$, then $x = x + 0 = 0 + x$ gives a nonunique decomposition unless $x = 0$. 👍

III.11.2.37. Corollary. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $\mathcal{W}_1$ and $\mathcal{W}_2$ be subspaces of $\mathcal{V}$, B_j a basis for $\mathcal{W}_j$, and $B = B_1 \cup B_2$. Then

(i) $\mathcal{W}_1 \cap \mathcal{W}_2 = \{0\}$ if and only if B is linearly independent.

(ii) $\mathcal{W}_1 \cup \mathcal{W}_2$ spans $\mathcal{V}$ if and only if B spans $\mathcal{V}$.

(iii) $\mathcal{W}_1$ and $\mathcal{W}_2$ are complementary if and only if B is a basis for $\mathcal{V}$.

In particular, if $\mathcal{W}_1$ and $\mathcal{W}_2$ are complementary, then $dim(\mathcal{W}_1) + dim(\mathcal{W}_2) = dim(\mathcal{V})$.

This suggests a method of proof of the following theorem, which is the main result on this topic:

III.11.2.38. THEOREM. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $\mathcal{W}$ a subspace of $\mathcal{V}$. Then there is a subspace of $\mathcal{V}$ complementary to $\mathcal{W}$.

PROOF: This is an easy consequence of Theorem III.11.2.28. (thus it is elementary if $\mathcal{V}$ is finite-dimensional, and requires AC in general). Let S be a basis for $\mathcal{W}$, and B a basis for $\mathcal{V}$ containing S. Then the span of $B \setminus S$ is a subspace of $\mathcal{V}$ complementary to $\mathcal{W}$, as is easily checked.
There is an alternate direct proof of this theorem using Zorn's Lemma (Exercise ()). 👍

III.11.2.39. From a theoretical point of view, this theorem is a fundamental fact; it is the essential feature distinguishing the theory of vector spaces from the theory of modules over rings. In fact, it is precisely the failure of this result for modules over a ring which is not a field that makes module theory considerably more complicated (and more interesting!) than the theory of vector spaces.

Two ways to rephrase this result are:

Every subspace of a vector space is a direct summand.

Every short exact sequence of vector spaces splits.

See books on abstract algebra such as [DF04] for details.

Quotient Spaces

There is one remaining topic about vector spaces, which is somewhat more complicated conceptually than the results so far, and which is not usually treated in linear algebra texts: quotient spaces. This construction is used frequently in functional analysis.

III.11.2.40. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $\mathcal{W}$ a subspace of $\mathcal{V}$. Define a relation $\sim$ on $\mathcal{V}$ by $x \sim y$ if $x - y \in \mathcal{W}$.

III.11.2.41. PROPOSITION.

(i) If $x, y, z \in \mathcal{V}$, $x \sim y$, and $y \sim z$, then $x \sim z$ ($\sim$ is *transitive*).

(ii) If $x_1, x_2, y_1, y_2 \in \mathcal{V}$, $x_1 \sim y_1$, and $x_2 \sim y_2$, then $x_1 + x_2 \sim y_1 + y_2$.

(iii) If $x, y \in \mathcal{V}$, $\alpha \in \mathbb{F}$, and $x \sim y$, then $\alpha x \sim \alpha y$.

PROOF: (i): we have $x - y \in \mathcal{W}$ and $y - z \in \mathcal{W}$, so

$$x - z = (x - y) + (y - z) \in \mathcal{W} .$$

(ii) is similar: $x_1 - y_1 \in \mathcal{W}$ and $x_2 - y_2 \in \mathcal{W}$, so

$$(x_1 + x_2) - (y_1 + y_2) = (x_1 - y_1) + (x_2 - y_2) \in \mathcal{W} .$$

The proof of (iii) is almost identical. 👍

III.11.2.42. By (i), $\sim$ is an equivalence relation () since it is obviously reflexive and symmetric. Thus $\mathcal{V}$ partitions into a disjoint union of equivalence classes (called *cosets* of $\mathcal{W}$ in $\mathcal{V}$). Denote by $\mathcal{V}/\mathcal{W}$ the set of equivalence classes, and for $x \in \mathcal{V}$ let $[x]$ denote the equivalence class $\{y \in \mathcal{V} : y \sim x\}$ of X.

By (ii), addition of equivalence classes by $[x] + [y] = [x + y]$ is well defined, and by (iii) scalar multiplication by $\alpha[x] = [\alpha x]$ is also well defined.

III.11.2.43. PROPOSITION. With this addition and scalar multiplication, $\mathcal{V}/\mathcal{W}$ is a vector space over $\mathbb{F}$.

PROOF: To prove associativity of addition, if $x, y, z \in \mathcal{V}$, we have

$$([x] + [y]) + [z] = [x + y] + [z] = [(x + y) + z] = [x + (y + z)] = [x] + [y + z] = [x] + ([y] + [z]) .$$

The proofs of the other properties are similar. 👍

III.11.2.44. DEFINITION. The vector space $\mathcal{V}/\mathcal{W}$ is called the *quotient vector space* of $\mathcal{V}$ by $\mathcal{W}$. The dimension of $\mathcal{V}/\mathcal{W}$ is called the *codimension* of $\mathcal{W}$ in $\mathcal{V}$.

III.11.2.45. EXAMPLES. (i) Let $\mathbb{F} = \mathbb{R}$, $\mathcal{V} = \mathbb{R}^2$, and $\mathcal{W}$ the one-dimensional space spanned by $\mathbf{e}_2 = (0, 1)$. If $\mathbf{x} = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, then the equivalence class of $\mathbf{x}$ is

$$[\mathbf{x}] = \{(\alpha_1, \beta) : \beta \in \mathbb{R}\} .$$

Thus the equivalence classes are parametrized by their first coordinates, and there is a natural one-one correspondence of $\mathbb{R}^2/\mathcal{W}$ with $\mathbb{R}^1$; this correspondence respects addition and scalar multiplication, so $\mathbb{R}^2/\mathcal{W} \cong \mathbb{R}^1$.

(ii) Here is a more typical example of how quotient spaces are used in analysis. Let X be an infinite set, and $\mathcal{V}$ the vector space of all functions from X to $\mathbb{F}$ (III.11.2.2.(iii)). Let $\mathcal{W}$ be the subspace of all functions which are 0 except on a finite subset of X (which may vary from function to function). It is easy to check that $\mathcal{W}$ is a subspace of $\mathcal{V}$. If $f, g \in \mathcal{V}$, then $f \sim g$ if and only if the functions f and g agree "almost everywhere" (except on a finite subset of X), i.e.

$$\{x \in X : f(x) \neq g(x)\}$$

is a finite subset of X. We can form the quotient vector space $\mathcal{V}/\mathcal{W}$, whose elements are equivalence classes of functions which agree "almost everywhere"; we often by slight abuse of terminology regard elements of $\mathcal{V}/\mathcal{W}$ as functions from X to $\mathbb{F}$, with functions agreeing almost everywhere identified together.

III.11.2.46. There is a systematization of the identification in (i) using complementary subspaces which is sometimes useful. Let $\mathcal{V}$ be a vector space and $\mathcal{W}$ a subspace of $\mathcal{V}$. Let $\mathcal{W}'$ be a complementary subspace to $\mathcal{W}$. For $x \in \mathcal{W}'$, define $\phi(x) = [x] \in \mathcal{V}/\mathcal{W}$.

III.11.2.47. PROPOSITION. The map $\phi : \mathcal{W}' \to \mathcal{V}/\mathcal{W}$ is an isomorphism.

PROOF: If $x \in \mathcal{V}$, then x can be written as $y + z$, with $y \in \mathcal{W}$ and $z \in \mathcal{W}'$. Then $x - z = y \in \mathcal{W}$, so $[x] = [z] = \phi(z)$ and ϕ is surjective. If $x, y \in \mathcal{W}'$ and $\phi(x) = [x] = [y] = \phi(y)$, then $x - y \in \mathcal{W}$; since $x - y \in \mathcal{W}'$, we have $x - y \in \mathcal{W} \cap \mathcal{W}' = \{0\}$, so $x = y$ and ϕ is injective. The fact that ϕ is linear follows immediately from the way addition and scalar multiplication in $\mathcal{V}/\mathcal{W}$ are defined. 👍

III.11.2.48. COROLLARY. If $\mathcal{V}$ is finite-dimensional and $\mathcal{W}$ is a subspace of $\mathcal{V}$, then $\mathcal{V}/\mathcal{W}$ is finite-dimensional and $dim(\mathcal{V}/\mathcal{W}) = dim(\mathcal{V}) - dim(\mathcal{W})$. Thus the dimension of $\mathcal{W}$ plus the codimension of $\mathcal{W}$ in $\mathcal{V}$ equals the dimension of $\mathcal{V}$.

III.11.2.49. Even if $\mathcal{V}$ and $\mathcal{W}$ are infinite-dimensional, the codimension of $\mathcal{W}$ in $\mathcal{V}$ is well defined. Thus the last statement of III.11.2.48. is true in general, and we obtain the nonobvious general fact that any two subspaces of $\mathcal{V}$ complementary to $\mathcal{W}$ have the same dimension.

III.11.2.50. The identification in III.11.2.47. is important in principle, but it does not help much in understanding an example like III.11.2.45.(ii), since in this case there is no obvious explicitly describable subspace of $\mathcal{V}$ which is complementary to $\mathcal{W}$ (in fact, in general the AC must be employed to obtain a complement in this example).

Summary

III.11.3. Linear Transformations

It follows from the results of the last section that the theory of vector spaces over a field $\mathbb{F}$ (especially finite-dimensional ones) is quite simple and not very interesting: the only isomorphism invariant is the dimension, and there are simple standard examples of vector spaces of each dimension.

The primary subject matter of linear algebra is not vector spaces themselves, but linear transformations between them. These form a much more complicated, and interesting, theory, which we outline in this section.

III.11.3.1. DEFINITION. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$. A function $T : \mathcal{V} \to \mathcal{W}$ is a *linear transformation* (*linear function, linear operator*) if T respects addition and scalar multiplication:

$$T(x + y) = T(x) + T(y)$$

$$T(\alpha x) = \alpha T(x)$$

for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{F}$.

III.11.3.2. The terms *linear transformation*, *linear function*, and *linear operator* are synonymous and interchangeable. We will usually use the term *linear transformation*, but often the term *(linear) operator* when $\mathcal{V} = \mathcal{W}$. The "linear" is often left out and understood when the term *operator* is used; nonlinear operators are also sometimes considered, but we will always qualify such operators explicitly as "nonlinear."

Note that the field $\mathbb{F}$ must be specified in the definition of linearity. If there is any potential ambiguity about the field we are working over, we use the more precise term $\mathbb{F}$-*linear* in place of *linear*. For example, if $\mathcal{V}$ and $\mathcal{W}$ are vector spaces over $\mathbb{C}$, we distinguish between real-linear and complex-linear transformations (any complex-linear transformation is real-linear, but not conversely); in this setting, "linear" will normally mean "complex-linear".

It is often customary to eliminate the parentheses around the argument of a linear transformation, i.e. to write Tx in place of $T(x)$. We will try to avoid doing this, at least at first, in the interest of clarity, to emphasize that a linear transformation is a function.

The first simple observation is:

III.11.3.3. Proposition. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$, and $T : \mathcal{V} \to \mathcal{W}$ a linear transformation. Then $T(\mathbf{0}) = \mathbf{0}$.

Proof: By linearity, we have $T(\mathbf{0}) = T(0 \cdot \mathbf{0}) = 0 \cdot T(\mathbf{0}) = \mathbf{0}$. 👍

This is generalized in III.11.3.6..

III.11.3.4. Examples. (i) An isomorphism of vector spaces (III.11.2.7.) is a linear transformation.

(ii) If $\mathcal{V}$ and $\mathcal{W}$ are any vector spaces over $\mathbb{F}$, the constant function 0 from $\mathcal{V}$ to $\mathcal{W}$ is linear.

(iii) Regard $\mathbb{F}$ as a one-dimensional vector space over $\mathbb{F}$. Let $T : \mathbb{F} \to \mathbb{F}$ be linear, and set $m = T(1)$. Then, for any $x \in \mathbb{F}$, we have $T(x) = T(x \cdot 1) = x \cdot T(1) = x \cdot m = mx$. Thus $T(x) = mx$ for some fixed $m \in \mathbb{F}$. Conversely, by the associative, commutative, and distributive laws in $\mathbb{F}$, any such function is linear.

(iv) Let $\mathcal{P}$ be the real vector space of polynomials with real coefficients. Define $D : \mathcal{P} \to \mathcal{P}$ by $D(f) = f'$. Elementary theorems of calculus (),() show that D is a linear transformation.

III.11.3.5. Example (iii) for $\mathbb{F} = \mathbb{R}$ reveals an ambiguity in standard mathematical terminology. It is customary in calculus and precalculus to call a function $f : \mathbb{R} \to \mathbb{R}$ a "linear function" if it is of the form

$$f(x) = mx + b$$

for constants m and b. But such a function is a linear function in the linear algebra sense if and only if $b = 0$. To resolve this ambiguity, we will usually call a function of the form $f(x) = mx + b$ an *affine function* rather than a linear function if $b \neq 0$, reserving the term "linear" for functions which are actually linear in the sense of III.11.3.1..

Null Space and Range

Linear transformations map subspaces to subspaces in the following sense:

III.11.3.6. PROPOSITION. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$, and let $T : \mathcal{V} \to \mathcal{W}$ be a linear transformation.

(i) Let $\mathcal{X}$ be a subspace of $\mathcal{V}$. Set

$$T(\mathcal{X}) = \{T(x) : x \in \mathcal{X}\} .$$

Then $T(\mathcal{X})$ is a subspace of $\mathcal{W}$.

(ii) Let $\mathcal{Y}$ be a subspace of $\mathcal{W}$. Set

$$T^{-1}(\mathcal{Y}) = \{x \in \mathcal{V} : T(x) \in \mathcal{Y}\} .$$

Then $T^{-1}(\mathcal{Y})$ is a subspace of $\mathcal{V}$.

PROOF: (i) $T(\mathcal{X})$ is nonempty since $\mathcal{X}$ is nonempty. Let $y_1, y_2 \in T(\mathcal{X})$. Then $y_1 = T(x_1)$ and $y_2 = T(x_2)$ for some $x_1, x_2 \in \mathcal{X}$. Then $x_1 + x_2 \in \mathcal{X}$, and

$$T(x_1 + x_2) = T(x_1) + T(x_2) = y_1 + y_2 \in T(\mathcal{X})$$

since T is linear. A similar argument shows that $T(\mathcal{X})$ is closed under scalar multiplication.

(ii): Since $T(\mathbf{0}_\mathcal{V}) = \mathbf{0}_\mathcal{W}$ and $\mathbf{0}_\mathcal{W} \in \mathcal{Y}$, $\mathbf{0}_\mathcal{V} \in T^{-1}(\mathcal{Y})$ and $T^{-1}(\mathcal{Y}) \neq \emptyset$. If $x_1, x_2 \in T^{-1}(\mathcal{Y})$, then $T(x_1), T(x_2) \in \mathcal{Y}$ and thus

$$T(x_1 + x_2) = T(x_1) + T(x_2) \in \mathcal{Y}$$

so $x_1 + x_2 \in T^{-1}(\mathcal{Y})$. Similarly, $T^{-1}(\mathcal{Y})$ is closed under scalar multiplication. 👍

There are two especially important subspaces associated with any linear transformation:

III.11.3.7. DEFINITION. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$, and let $T : \mathcal{V} \to \mathcal{W}$ be a linear transformation. The *range* of T is

$$\mathcal{R}(T) = T(\mathcal{V}) = \{T(x) : x \in \mathcal{V}\} \subseteq \mathcal{W} .$$

The *null space* of T is

$$\mathcal{N}(T) = T^{-1}(\{\mathbf{0}\}) = \{x \in \mathcal{V} : T(x) = \mathbf{0}\} \subseteq \mathcal{V} .$$

By III.11.3.6., $\mathcal{R}(T)$ and $\mathcal{N}(T)$ are subspaces of $\mathcal{W}$ and $\mathcal{V}$ respectively.

III.11.3.8. PROPOSITION. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$, and let $T : \mathcal{V} \to \mathcal{W}$ be a linear transformation. Then T is one-to-one if and only if $\mathcal{N}(T) = \{\mathbf{0}\}$.

PROOF: If T is one-to-one, if $x \in \mathcal{N}(T)$, then $T(x) = \mathbf{0} = T(\mathbf{0})$, so $x = \mathbf{0}$. Conversely, suppose $\mathcal{N}(T) = \{\mathbf{0}\}$. If $x_1, x_2 \in \mathcal{V}$ and $T(x_1) = T(x_2)$, then $T(x_1 - x_2) = T(x_1) - T(x_2) = \mathbf{0}$ and $x_1 - x_2 \in \mathcal{N}(T)$, so $x_1 = x_2$ and T is one-to-one. 👍

The next result, a generalization of III.11.3.8., is one of the fundamental facts from linear algebra:

III.11.3.9. THEOREM. [FIRST ISOMORPHISM THEOREM] Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$, and let $T : \mathcal{V} \to \mathcal{W}$ be a linear transformation. Let $\mathcal{Y}$ be a subspace of $\mathcal{V}$ contained in $\mathcal{N}(T)$. Then T induces a well-defined linear transformation $\bar{T}$ from $\mathcal{V}/\mathcal{Y}$ onto $\mathcal{R}(T)$ by $\bar{T}([x]) = T(x)$ for $x \in \mathcal{V}$. The transformation $\bar{T}$ is an isomorphism if and only if $\mathcal{Y} = \mathcal{N}(T)$.

PROOF: We have to show that $\bar{T}$ is well-defined; then it is automatic that $\bar{T}$ is linear and $\mathcal{R}(\bar{T}) = \mathcal{R}(T)$. So suppose $x_1, x_2 \in \mathcal{V}$ and $[x_1] = [x_2]$ in $\mathcal{V}/\mathcal{Y}$. Then $x_1 - x_2 \in \mathcal{Y} \subseteq \mathcal{N}(T)$, so we have

$$T(x_1) = T(x_2 + (x_1 - x_2)) = T(x_2) + T(x_1 - x_2) = T(x_2) + \mathbf{0} = T(x_2)$$

and $\bar{T}$ is well defined. The argument is reversible if (and only if) $\mathcal{Y} = \mathcal{N}(T)$: if $\bar{T}([x_1]) = \bar{T}([x_2])$, then $T(x_1) = T(x_2)$, $T(x_1 - x_2) = \mathbf{0}$, $x_1 - x_2 \in \mathcal{N}(T)$, so $[x_1] = [x_2]$. Thus $\bar{T}$ is one-to-one. 👍

III.11.3.10. Here is an alternate argument for the case $\mathcal{Y} = \mathcal{N}(T)$. Let $\mathcal{X}$ be a subspace of $\mathcal{V}$ complementary to $\mathcal{N}(T)$ (III.11.2.38.). Then $\tilde{T} := T|_{\mathcal{X}}$ is a linear transformation from $\mathcal{X}$ to $\mathcal{W}$. It is easily checked that $\tilde{T}$ is one-to-one (since $\mathcal{X} \cap \mathcal{N}(T) = \{\mathbf{0}\}$), and that $\mathcal{R}(\tilde{T}) = \mathcal{R}(T)$. $\bar{T}$ is then obtained by composing $\tilde{T}$ with the natural isomorphism from $\mathcal{V}/\mathcal{N}(T)$ onto $\mathcal{X}$ (III.11.2.47.).

Using III.11.2.48., we obtain:

III.11.3.11. COROLLARY. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over $\mathbb{F}$, and let $T : \mathcal{V} \to \mathcal{W}$ be a linear transformation. Then $dim(\mathcal{R}(T)) = codim(\mathcal{N}(T))$. In particular,

$$dim(\mathcal{N}(T)) + \dim(\mathcal{R}(T)) = \dim(\mathcal{V}) \ .$$

This corollary is often called the "dimension-counting theorem" for linear transformations. It has an important consequence for finite-dimensional vector spaces (using III.11.3.8.):

III.11.3.12. COROLLARY. Let $\mathcal{V}$ be a finite-dimensional vector space over $\mathbb{F}$, and let $T : \mathcal{V} \to \mathcal{V}$ be a linear transformation. Then T is injective if and only if it is surjective.

This can be false if $\mathcal{V}$ is infinite-dimensional ().

Idempotent Operators

As a good and important example of operators, we examine idempotent operators. These are the simplest operators other than scalar multiples of the identity. Many more complicated operators are built up from idempotents (cf. XX.14.1.6.).

III.11.3.13. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$. An operator $P \in \mathcal{L}(\mathcal{V})$ is *idempotent* if $P^2 = P$.

III.11.3.14. There are always at least two idempotent operators on any $\mathcal{V}$, the identity and the zero operator. Note that "idempotent" only makes sense for an operator from a vector space to itself.

The range and null space of an idempotent operator are simple to describe and are nicely behaved:

III.11.3.15. PROPOSITION. Let P be an idempotent operator on a vector space $\mathcal{V}$. Then
(i) $\mathcal{R}(P) = \{x \in \mathcal{V} : P(x) = x\}$.
(ii) $\mathcal{R}(P)$ and $\mathcal{N}(P)$ are complementary subspaces of $\mathcal{V}$.

PROOF: (i): If $P(x) = x$, then x is obviously in $\mathcal{R}(P)$. Conversely, if $x = P(y)$ for some $y \in \mathcal{V}$, then

$$P(x) = P(P(y)) = P^2(y) = P(y) = x \ .$$

(ii): If $x \in \mathcal{R}(P) \cap \mathcal{N}(P)$, then $x = P(x) = 0$. And if $x \in \mathcal{V}$, then $x = P(x) + (x - P(x))$, and $P(x) \in \mathcal{R}(P)$, $x - P(x) \in \mathcal{N}(P)$. 👍

III.11.3.16. If P is an idempotent operator on $\mathcal{V}$, then P is often called the *projection onto* $\mathcal{R}(P)$ *along* $\mathcal{N}(P)$. Examples of idempotent operators on $\mathbb{R}^2$ with one-dimensional range and null space show geometrically why this terminology is appropriate.

III.11.3.17. PROPOSITION. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and $\mathcal{X}$ and $\mathcal{Y}$ complementary subspaces of $\mathcal{V}$. Then there is a unique projection onto $\mathcal{X}$ along $\mathcal{Y}$.

PROOF: Any $v \in \mathcal{V}$ can be uniquely written as $x + y$ with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The unique projection onto $\mathcal{X}$ along $\mathcal{Y}$ is the operator P defined by $P(x + y) = x$ for $x \in \mathcal{X}$, $y \in \mathcal{Y}$. It is easily checked that P is well defined, linear, and idempotent, with $\mathcal{R}(P) = \mathcal{X}$, $\mathcal{N}(P) = \mathcal{Y}$, and this P is the only such possibility. 👍

III.11.3.18. Since any subspace $\mathcal{X}$ of a vector space $\mathcal{V}$ has a complementary subspace (III.11.2.38.), there is always a projection onto $\mathcal{X}$. But since complementary subspaces are highly nonunique in general, so are projections onto subspaces.

III.11.3.19. If P is an idempotent operator, then $I - P$ is also idempotent. In fact, if P is the projection onto $\mathcal{X}$ along $\mathcal{Y}$, then $I - P$ is the projection onto $\mathcal{Y}$ along $\mathcal{X}$.

Existence of Linear Transformations

III.11.3.20. PROPOSITION. Let V and W be vector spaces over $\mathbb{F}$, and let $B = \{v_j : j \in J\}$ be a basis for V. Then if f is any function from B to W, there is a unique linear transformation $T : V \to W$ such that $Tv_j = f(v_j)$ for all $j \in J$.

PROOF: If $x \in V$, then x can be uniquely written $x = \sum_{j \in J} \alpha_j v_j$ with $\alpha_j \in \mathbb{F}$ and only finitely many α_j nonzero. Then the only possible formula for T is

$$Tx = \sum_{j \in J} \alpha_j f(v_j)$$

(note that this is well defined since it is really a finite sum). One only needs to check that T is linear, which is routine. 👍

III.11.3.21. Thus, if V has a basis (automatic if V is finite-dimensional or the AC is assumed), then for any $W(\neq \{0\})$ there are many linear transformations from V to W.

Extension of Linear Transformations

Linear transformations can always be extended from subspaces:

III.11.3.22. PROPOSITION. Let V and W be vector spaces over $\mathbb{F}$, and Y a subspace of V. Any linear transformation from Y to W can be extended to a linear transformation from V to W.

PROOF: Let $T : Y \to W$ be a linear transformation. Let B be a basis for Y, and extend B to a basis $\tilde{B}$ for V. The transformation T gives a function from B to W. Extend this function in any way to $\tilde{B}$, e.g. sending the elements of $\tilde{B} \setminus B$ to 0. This function extends to a linear transformation extending T. 👍

III.11.3.23. As the construction makes clear, there is nothing unique about the extension. Even the extension which is zero on $\tilde{B} \setminus B$ is not canonical, since there are many choices for $\tilde{B}$.

III.11.4. The Matrix of a Linear Transformation

In this section, we work only with finite-dimensional vector spaces. There is then an intimate relationship between linear transformations and matrices, although the relationship requires a choice of bases and is therefore not "natural" or canonical in general.

III.11.4.24. First consider the case of linear transformations from $\mathbb{F}^n$ to $\mathbb{F}^m$. In this case the vector spaces have canonical bases $\{\boldsymbol{e}_1^{(n)}, \ldots, \boldsymbol{e}_n^{(n)}\}$ and $\{\boldsymbol{e}_1^{(m)}, \ldots, \boldsymbol{e}_n^{(m)}\}$ respectively. If $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$, set

$$T\boldsymbol{e}_k^{(n)} = \sum_{j=1}^{m} \alpha_{jk} \boldsymbol{e}_j^{(m)}$$

for $1 \leq j \leq n$. The scalars $\{\alpha_{jk}\}$ completely determine T, since if $\boldsymbol{x} \in \mathbb{F}^n$, then $\boldsymbol{x}$ can be uniquely written $\boldsymbol{x} = \lambda_1 \boldsymbol{e}_1^{(n)} + \cdots \lambda_n \boldsymbol{e}_n^{(n)}$, and by linearity

$$T\boldsymbol{x} = \sum_{j=1}^{m} \sum_{k=1}^{n} \alpha_{jk} \lambda_k \boldsymbol{e}_j^{(m)} \ .$$

Symbolically, if we write the elements of $\mathbb{F}^n$ and $\mathbb{F}^m$ as column vectors, we have

$$T\left(\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}\right) = \begin{bmatrix} \sum_{k=1}^{n} \alpha_{1k}\lambda_k \\ \vdots \\ \sum_{k=1}^{n} \alpha_{mk}\lambda_k \end{bmatrix} \ .$$

If we define an $m \times n$ matrix $[T]$ as $[\alpha_{jk}]$, we have

$$T\boldsymbol{x} = [T]\boldsymbol{x} \ .$$

The matrix $[T]$ is called the *matrix of* T (with respect to the standard bases). Conversely, if A is any $m \times n$ matrix over $\mathbb{F}$, define a function $T : \mathbb{F}^n \to \mathbb{F}^m$ by

$$T\boldsymbol{x} = A\boldsymbol{x}$$

for $x \in \mathbb{F}^n$. It follows from properties of matrix multiplication that T so defined is linear, and A is the matrix of T. Thus there is a natural one-one correspondence between $\mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$ and $M_{m,n}(\mathbb{F})$.

If $T \in \mathcal{L}(\mathbb{F}^n, \mathbb{F}^m)$, then the k'th column of $[T]$ is exactly the column vector $T\boldsymbol{e}_k^{(n)}$.

III.11.4.25. Now let V and W be finite-dimensional vector spaces over $\mathbb{F}$, with $dim(V) = n$, $dim(W) = m$. Fix (ordered) bases $X = \{x_1, \ldots, x_n\}$ for V and $Y = \{y_1, \ldots, y_m\}$ for W. Let $T \in \mathcal{L}(V, W)$. Then for each j, $1 \leq j \leq n$, we can write

$$Tx_k = \sum_{j=1}^{m} \alpha_{jk} y_j$$

for unique scalars α_{jk}. The scalars $\{\alpha_{jk}\}$ completely determine T, since if $v \in V$, then v can be uniquely written $v = \lambda_1 x_1 + \cdots \lambda_n x_n$, and by linearity

$$Tv = \sum_{k=1}^{m} \sum_{j=1}^{n} \alpha_{jk} \lambda_k y_j \ .$$

We can identify V and W with $\mathbb{F}^n$ and $\mathbb{F}^m$ respectively by

$$v = \sum_{k=1}^{n} \lambda_k x_k \leftrightarrow [v]_X := \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} \quad w = \sum_{j=1}^{m} \mu_j y_j \leftrightarrow [w]_Y := \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix}$$

(the column vector $[v]_X$ is the *coordinate vector* of v with respect to the basis X) and if we then define ${}_Y[T]_X = [\alpha_{jk}] \in M_{m,n}(\mathbb{F})$, we have

$$[Tv]_Y = {}_Y[T]_X [v]_X$$

for any $x \in V$. The matrix ${}_Y[T]_X$ is called the *matrix of* T *with respect to the bases* X and Y. Conversely, any $m \times n$ matrix over $\mathbb{F}$ defines an element of $\mathcal{L}(V, W)$. Thus, once bases for V and W are fixed, we get an identification of $\mathcal{L}(V, W)$ with $M_{m,n}(\mathbb{F})$. But this identification requires a choice of bases for V and W; if the bases are changed, the identification also changes ().

Composition

III.11.4.26. Let V_1, V_2, V_3 be finite-dimensional vector spaces over $\mathbb{F}$, with chosen bases $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_m\}$, $Z = \{z_1, \ldots, z_p\}$ respectively. Let $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, Z)$. Then the composition $S \circ T$ (usually written ST) is in $\mathcal{L}(V, Z)$. With respect to the given bases, these operators have matrices ${}_Y[T]_X$, ${}_Z[S]_Y$, and ${}_Z[ST]_X$ which are $m \times n$, $p \times m$, and $p \times n$ respectively. Identifying elements of V, W, Z with column vectors using the bases as above, we have

$${}_Z[ST]_X [v]_X = [(ST)v]_Z = [S(Tv)]_Z = {}_Z[S]_Y [Tv]_Y = {}_Z[S]_Y ({}_Y[T]_X [v]_X) = ({}_Z[S]_Y \, {}_Y[T]_X)[v]_X$$

for $v \in V$ by definition of composition and associativity of matrix multiplication. So we obtain:

III.11.4.27. PROPOSITION. In the above situation, ${}_Z[ST]_X = {}_Z[S]_Y\,{}_Y[T]_X$, i.e. the matrix of a composition of linear transformations is the product of the matrices of the transformations.

This justifies the notation ST for $S \circ T$. Of course, bases must be chosen, and the same basis must be used for the middle space for both transformations. The case $V = \mathbb{R}^n$, $W = \mathbb{R}^m$, $Z = \mathbb{R}^p$ is slightly easier to state using the standard bases. Note that the right side is a product of a $p \times m$ matrix and an $m \times n$ matrix, which is defined and $p \times n$, the size of the left side (this is of course not a proof of the formula, only an indication that the equation makes sense).

Change of Basis

III.11.4.28. A particularly important case is when $V = W$ and T is the identity map. Let V be a finite-dimensional vector space and $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ two bases for V. If $v \in V$, we can write

$$v = \sum_{k=1}^{n} \lambda_k x_k = \sum_{j=1}^{n} \mu_j y_j$$

or, in coordinates,

$$[v]_X = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} \quad [v]_Y = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$$

with $\lambda_k, \mu_j \in \mathbb{F}$. How are the λ_k and μ_j related?

III.11.4.29. For each j, write

$$x_k = \sum_{j=1}^{n} \alpha_{jk} y_j$$

then

$$v = \sum_{k=1}^{n} \lambda_k x_k = \sum_{k=1}^{n} \lambda_k \left(\sum_{j=1}^{n} \alpha_{jk} y_j \right) = \sum_{j=1}^{n} \left(\sum_{k=1}^{n} \alpha_{jk} \lambda_k \right) y_j$$

so, for each j,

$$\mu_j = \sum_{k=1}^{n} \alpha_{jk} \lambda_k \ .$$

Thus, if we let ${}_Y C_X = [\alpha_{jk}]$, then ${}_Y C_X$ is $n \times n$, and we have

$$[v]_Y = {}_Y C_X [v]_X$$

for any $v \in V$.

The matrix ${}_Y C_X$ is actually ${}_Y[I]_X$, the matrix of the identity transformation from V to V with respect to the bases X and Y, but the notation ${}_Y[I]_X$ may be confusing since this is not the identity $n \times n$ matrix unless $Y = X$, so we use ${}_Y C_X$ instead.

III.11.4.30. We have ${}_X C_X = I_n$, and if X, Y, Z are bases for V we have ${}_Z C_Y\,{}_Y C_X = {}_Z C_X$. Thus ${}_Y C_X$ is invertible for any X and Y, with inverse ${}_X C_Y$.

III.11.4.31. Definition. The matrix ${}_XC_Y = ({}_YC_X)^{-1}$ is called the *change-of-basis matrix* from the basis X to the basis Y.

III.11.4.32. It is conventional to take ${}_XC_Y$ and not ${}_YC_X$ as the change-of-basis matrix from X to Y: if we think of X as the "old" basis and Y the "new" basis, the columns of ${}_XC_Y$ give the coordinates of the new basis vectors in terms of the old ones. For example, if we "rescale" the basis X by a factor of α, i.e. $y_j = \alpha x_j$ for each j, then the change-of-basis matrix is αI_n.

III.11.4.33. If we have a linear transformation $T : V \to W$, and we have bases X, X' for V and Y, Y' for W, a straightforward calculation shows that

$$ {}_{Y'}[T]_{X'} = {}_{Y'}C_Y \, {}_Y[T]_X \, {}_XC_{X'} $$

and in particular, if X and Y are bases for V and $T \in \mathcal{L}(V)$, then

$$ {}_Y[T]_Y = {}_YC_X \, {}_X[T]_X \, {}_XC_Y = ({}_XC_Y)^{-1} \, {}_X[T]_X \, {}_XC_Y $$

i.e. when the basis for V is changed, the matrix of a linear operator on V with respect to the old basis is conjugated by the change-of-basis matrix to get the matrix with respect to the new basis.

III.11.5. Linear Functionals and Dual Spaces

Linear functionals are the key to the applications of linear algebra to vector analysis via differential forms, and the dual space is also one of the fundamental concepts of functional analysis.

III.11.5.1. Definition. Let V be a vector space over $\mathbb{F}$. A *linear functional* on V is a linear transformation from V to $\mathbb{F}$ (regarded as a vector space over itself). The set $\mathcal{L}(V, \mathbb{F})$ of linear functionals on V is called the *(algebraic) dual space of* V, denoted V^d.

Then V^d is a vector space over $\mathbb{F}$. The notation V^d is not standard: most authors use either the notation V' or V^* for the dual space of V. But in this book we will largely be considering topological vector spaces, and the common notation for the *topological* dual (the set of continuous linear functionals) is either V' or V^*, so we have adopted a notation for the algebraic dual different from either.

III.11.5.2. By III.11.3.22., if V is a vector space over $\mathbb{F}$ and Y is a subspace of V, every linear functional on Y extends to a linear functional on V. (In fact there are many extensions, and no canonical one,)

III.11.5.3. Proposition. Let V be a vector space over $\mathbb{F}$, and Y a proper subspace of V. If x is any vector in V not in Y, there is a linear functional ϕ on V with $V \subseteq \mathcal{N}(\phi)$ and $\phi(x) = 1$.

Proof: Let X be the subspace of V spanned by V and x. Then every vector in X can be uniquely written as $y + \alpha x$ for some $y \in Y$ and $\alpha \in \mathbb{F}$. Then the function $\psi : X \to \mathbb{F}$, $\psi(y + \alpha x) = \alpha$, is a linear functional with $\psi(x) = 1$ and $\psi(y) = 0$ for all $y \in Y$. Extend ψ to a linear functional ϕ on V.

The Dual Basis

III.11.5.4. If V is finite-dimensional, fix a basis $B = \{v_1, \ldots, v_n\}$ for V. By III.11.3.20., any function $f : B \to \mathbb{F}$ extends to a linear functional on V. For $1 \leq j \leq n$, set $f_j(v_k) = \delta_{jk}$. Then f_j extends to a linear functional on V, denoted v^j.

III.11.5.5. PROPOSITION. The set $\{v^1, \ldots, v^n\}$ is a basis for V^d, called the *dual basis* for B.

PROOF: If ϕ is any linear functional on V, set $\phi(v_j) = \alpha_j$ for $1 \leq j \leq n$. If $x \in V$, write $x = \lambda_1 v_1 + \cdots + \lambda_n v_n$. Then, by linearity,

$$\phi(x) = \sum_{j=1}^{n} \alpha_j \lambda_j = \sum_{j=1}^{n} \alpha_j v^j(x)$$

so $\phi = \sum_{j=1}^{n} \alpha_j v^j$ and the v_j span V^d. To show they are linearly independent, suppose $\sum_{j=1}^{n} \alpha_j v^j = 0$ (i.e. is the zero linear functional). Then, for each k,

$$0 = \sum_{j=1}^{n} \alpha_j v^j(v_k) = \alpha_k \ .$$



III.11.5.6. COROLLARY. Let V be a finite-dimensional vector space over $\mathbb{F}$. Then V^d is also finite-dimensional, and $dim(V^d) = dim(V)$.

III.11.5.7. Using the basis B for V and the basis $\{1\}$ for $\mathbb{F}$, V^d can be identified with the set of $1 \times n$ matrices (row n-vectors) over $\mathbb{F}$. This gives another proof of III.11.5.6.. Under this identification, the dual basis is the standard basis of $\mathbb{F}^n$ as row vectors. The pairing between V^d and V represented as column n-vectors is the matrix product or "dot product" (it is exactly the dot product if $\mathbb{F} = \mathbb{R}$, although not if $\mathbb{F} = \mathbb{C}$). Note, however, that although V and V^d are the same dimension and thus isomorphic, there is no *natural* isomorphism: the obvious isomorphism via the dual basis depends on a choice of basis for V. If $\mathbb{F} = \mathbb{R}$ and an inner product () is fixed on V, there is a natural isomorphism of V and V^d via the Riesz Representation Theorem (); in fact, an isomorphism of V and V^d is more or less the same thing as a choice of an inner product.



III.11.5.8. Nothing like this is true if V is infinite-dimensional. If $B = \{v_j : j \in J\}$ is a basis for V, then we can define the dual functionals v^j the same way, and they are linearly independent by the same argument, but they do not span V^d: any function f from B to $\mathbb{F}$ defines a linear functional ϕ on V, but if $f(v_j) \neq 0$ for infinitely many j, then ϕ is not a (finite) linear combination of the v^j.

The Second Dual

III.11.5.9. If V is a vector space over $\mathbb{F}$, set $V^{dd} = (V^d)^d$. V^{dd} is called the *second dual* of V.

There is a natural linear map from V into V^{dd}: if $v \in V$, define $\hat{v} : V^d \to \mathbb{F}$ by

$$\hat{v}(\phi) = \phi(v) \ .$$

If V has a basis (e.g. if V is finite-dimensional or the AC is assumed), it follows easily from the dual functional construction that this map is injective, and thus we may regard $V \subseteq V^{dd}$. If V is finite-dimensional, $dim(V^{dd}) = dim(V^d) = dim(V)$, so the map $v \mapsto \hat{v}$ from V to V^{dd} is an isomorphism. This can also be seen by noting that if $B = \{v_1, \ldots, v_n\}$ is a basis for V, and $B^d = \{v^1, \ldots, v^n\}$ is the dual basis for V^d, then the dual basis for B^d in V^{dd} is $\{\hat{v}_1, \ldots, \hat{v}_n\}$.

As a result, the situation is completely symmetric for V and V^d if V is finite-dimensional. The pairing $V^d \times V \to \mathbb{F}$, $(\phi, x) \mapsto \phi(x)$, is often written $\langle \phi, x \rangle$ or $\langle x, \phi \rangle$.

III.11.5.10. However, if V is infinite-dimensional, we never have (under AC) that the map from V to V^{dd} is surjective (it cannot even be proved to always be injective without some form of Choice). For if $\{v_j : j \in J\}$ is a basis for V and $\{v^j : j \in J\}$ are the dual functionals, then the v^j do not span V^d (III.11.5.8.), and if this set is expanded to a basis for V^d, the dual functionals to the additional basis elements are not in the image of V in V^{dd}.

Duals of Subspaces and Quotients

Let V be a vector space over $\mathbb{F}$ and Y a subspace. What is the relation between V^d, Y^d, and $(X/Y)^d$?

III.11.5.11. We define the *annihilator* $Y^\perp$ to be

$$Y^\perp = \{\phi \in V^d : \phi(y) = 0 \ \forall y \in Y\} \ .$$

Then $Y^\perp$ is a subspace of V^d. By III.11.5.3., $Y^\perp = \{0\}$ if and only if $Y = V$. At the other extreme, $\{0\}^\perp = V^d$.

III.11.5.12. Let ρ be the restriction map from V^d to Y^d. Then ρ is surjective by III.11.5.2., and its null space is exactly $Y^\perp$. Thus ρ drops to a vector space isomorphism $\bar{\rho}$ from $V^d/Y^\perp$ to Y^d.

III.11.5.13. Let $\pi : V \to V/Y$ be the quotient map. Every $\psi \in (X/Y)^d$ defines a linear functional $\psi \circ \pi$ on V, which vanishes on Y, i.e. $\psi \circ \pi \in Y^\perp$. The map $\psi \mapsto \psi \circ \pi$ is a linear map σ from $(X/Y)^*$ to $Y^\perp$. The map σ is clearly injective. Any $\phi \in Y^\perp$ has null space containing Y and thus drops to a linear functional $\bar{\phi}$ on X/Y, so σ is surjective. Thus $(V/Y)^d$ is naturally isomorphic to $Y^\perp$.

III.11.5.14. Thus the dual of a subspace of V is a quotient of V^*, and the dual of a quotient of V is a subspace of V^*. The correspondence between V and V^d is an important example of a *duality* (). In a duality, subobjects correspond to quotients and vice versa.

III.11.5.15. The duality between V and V^d is complete if V is finite-dimensional, but not if V is infinite-dimensional: not every subspace of V^d is $Y^\perp$ for a subspace Y of V.

Here is a technical consequence which will be useful, e.g. in ():

III.11.5.16. PROPOSITION. Let V be a vector space over $\mathbb{F}$, and $\psi_1, \ldots, \psi_n \in V^d$. If $\phi \in V^d$ and $\phi(x) = 0$ whenever $\psi_1(x) = \cdots = \psi_n(x) = 0$, then ϕ is a linear combination of $\{\psi_1, \ldots, \psi_n\}$.

PROOF: Passing to a subset if necessary, we may assume that $\{\psi_1, \ldots, \psi_n\}$ are linearly independent. Let Y be the intersection of the null spaces of the ψ_k. Define $T : V \to \mathbb{F}^n$ by

$$T(x) = (\phi_1(x), \ldots, \phi_n(x)) \ .$$

Then T is a linear transformation, and $\mathcal{N}(T)$ is exactly Y, so T drops to an injective linear transformation $\bar{T}$ from V/Y to $\mathbb{F}^n$, and therefore $dim(V/Y) \leq n$. Thus $dim((V/Y)^d) \leq n$ and, since $(V/Y)^d \cong Y^\perp$, $dim(Y^\perp) \leq n$. But $\{\psi_1, \ldots, \psi_n\}$ is a linearly independent set in $Y^\perp$, so $dim(Y^\perp) = n$ and $\{\psi_1, \ldots, \psi_n\}$ is a basis for $Y^\perp$. Since $\phi \in Y^\perp$, the result follows. 👍

As a corollary of the proof, we obtain:

III.11.5.17. COROLLARY. Let V be a vector space over $\mathbb{F}$, and $\{\psi_1, \ldots, \psi_n\}$ a linearly independent set in V^d. Then

(i) The intersection Y of the null spaces of the ψ_k has codimension n.

(ii) If $(\alpha_1, \ldots, \alpha_n)$ is any n-tuple of elements of $\mathbb{F}$, then there is an $x \in V$ with $\phi_k(x) = \alpha_k$ for all k, $1 \leq k \leq n$.

(iii) If W is a subspace of V of dimension greater than n, then $W \cap Y \neq \{0\}$.

PROOF: Using the notation of the proof of III.11.5.16., V/Y has dimension exactly n, so $\bar{T}$ is an injective linear transformation between two n-dimensional vector spaces and is therefore surjective. The restriction of T to W cannot be injective. 👍

III.11.6. Covariant vs. Contravariant

We digress to describe some inconsistent and confusing terminology in mathematics: use of the terms *covariant* and *contravariant*.

III.11.6.1. In many references on differential geometry, and in some on linear algebra, particularly older ones, elements of a vector space V are called *contravariant vectors* and elements of the dual space V^d are called *covariant vectors*, usually without any explanation of the reason. This terminology in this context was introduced by SYLVESTER in the nineteenth century. The apparent reason for the use of "contravariant" for vectors in V is that when one changes bases from X to Y in V, the new coordinates of a vector v are obtained from the old ones by multiplication by the *inverse* of the change-of-basis matrix ${}_XC_Y$ (III.11.4.31.). This is arbitrary and simply a result of the way the change-of-basis matrix is conventionally defined (granted, the convention does seem to be the natural one). On the other hand, in V^d a change of (dual) basis is effected by multiplication by the change-of-basis matrix itself, not its inverse (), so vectors in the dual space are "covariant." There are extensions of this use of the terms to vector bundles () and tensors (). Use of these terms in this context seems to be fading away (mercifully), although they are still common with tensors.

III.11.6.2. This usage is in unfortunate conflict with the usual use of *covariant* and *contravariant* in modern mathematics. A careful description of these modern terms uses category theory (II.11.): a functor between categories which preserves the direction of morphisms is called *covariant* and one which reverses them is called *contravariant*. Informally, a contravariant correspondence between objects turns arrows between them around. A good example is the dual space construction for vector spaces: the correspondence $V \mapsto V^d$ is contravariant since if $T : V \to W$ is a linear transformation, the dual transformation T^d goes from W^d to V^d.

III.11.6.3. The inconsistency of terminology is particularly acute with tensors. A tensor has a "contravariant" degree and a "covariant" degree (see () for details). Unfortunately, tensors are *covariant* with respect to the underlying vector space in the "contravariant" part and *contravariant* in the "covariant" part.

III.11.7. More on Matrices and Linear Equations

In this subsection, we describe the standard algorithms for solving systems of linear equations in linear algebra (matrix) terms. We will not go into detail about them; details can be found in any linear algebra text.

The basic problem we consider is solving an equation

$$A\boldsymbol{x} = \boldsymbol{b}$$

where A is an $m \times n$ matrix over a field $\mathbb{F}$ (fixed for the discussion), $\boldsymbol{b}$ is a column m-vector (element of $\mathbb{F}^m$), and $\boldsymbol{x}$ is an unknown column n-vector (element of $\mathbb{F}^n$).

Echelon Matrices

III.11.7.1. Definition. Let $\mathbb{F}$ be a field, and A an $m \times n$ matrix over $\mathbb{F}$. The matrix A is in *echelon form* if the first nonzero term in each row, if any, is 1 and all entries below and to the left of such a leading 1 are 0 (i.e. the leading 1 in each row is farther to the right than the leading 1 in the rows above). A is in *reduced echelon form* if also all the entries above each leading 1 are 0. Zero rows are permitted, but they must be at the bottom of the matrix (an $m \times n$ matrix in echelon form cannot have more than n nonzero rows).

III.11.7.2. Examples. The following are in reduced echelon form:

The identity matrix I_n and any zero matrix.

The matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The following are in echelon form but not reduced echelon form:

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

The following are not in echelon form:

$$\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

III.11.7.3. The importance of echelon matrices is that if we have an equation

$$A\boldsymbol{x} = \boldsymbol{b}$$

and A is in echelon form, the system can be easily solved *provided* that if A has any zero rows, the corresponding entries of $\boldsymbol{b}$ must be 0 (otherwise the system is *inconsistent* and has no solution). The solution algorithm is to start with the last nonzero row and solve for the leading variable, arbitrarily choosing values (say 0) for any later variables, then moving to the next row up, again arbitrarily choosing values for any later variables not already specified, and so on to the first row (if the leading term in the first row is not the first term, any previous variables do not appear in the equations and can again be chosen arbitrarily).

If A has fewer than n nonzero rows, and the system is consistent, there will be variables which can be chosen arbitrarily, and thus there will be more than one solution. In particular, if $n > m$ and the system is consistent there will be infinitely many solutions (if $\mathbb{F}$ is infinite). See III.11.7.13..

Elementary Row Operations

Every matrix can be reduced to echelon form, even reduced echelon form, by a sequence of elementary row operations.

III.11.7.4. The *elementary row operations* on an $m \times n$ matrix A over $\mathbb{F}$ are:

(i) Replace a row j of A by a scalar multiple of row j by a nonzero element α of $\mathbb{F}$.

(ii) Interchange two rows i and j of A.

(iii) Replace a row i of A by that row plus a scalar multiple (by an element α of $\mathbb{F}$) of another row j of A.

III.11.7.5. The algorithm for reducing A to echelon form by elementary row operations is:

(1) Ignore any initial columns of 0s of A. In the first column k_1 of A with a nonzero entry, interchange two rows if necessary to make the $(1, k_1)$ element nonzero.

(2) Multiply the first row by $a_{1k_1}^{-1}$ to make the $(1, k_1)$ element 1.

(3) For each $j > 1$, replace the j'th row by the j'th row plus $(-a_{jk_1})$ times the first row to make the (j, k_1) element 0.

(4) If all rows are finished, or if all remaining rows below are zero, terminate the process. Otherwise ignore the first row and apply (1)–(3) to the remaining matrix (the number k_2 of the first nonzero column in this matrix will satisfy $k_2 > k_1$).

(5) Continue until the process terminates.

One can even reduce to reduced echelon form by expanding (3) to also make the elements of previous rows above the initial 1 zero using (iii).

Note, however, that in general choices have to be made during the process (about which rows are brought to the top) so the echelon form matrix (even the reduced echelon form matrix) obtained is not necessarily unique.

III.11.7.6. Elementary row operations can be done by premultiplication of A by a certain $m \times m$ matrix:

Operation (i) is premultiplication of A by a diagonal matrix with $e_{jj} = \alpha$ and $e_{kk} = 1$ for $k \neq j$.

Operation (ii) is premultiplication of A by the matrix with $e_{ij} = e_{ji} = 1$, $e_{kk} = 1$ for $k \neq i, j$, and all other entries 0.

Operation (iii) is premultiplication of A by the matrix with $e_{ij} = \alpha$, $e_{kk} = 1$ for all k, and all other entries 0.

Matrices of one of these forms are called *elementary matrices*.

III.11.7.7. Thus for any $m \times n$ matrix A there is a finite set $E_1, \ldots, E_r$ of elementary matrices such that $E_r \cdots E_1 A$ is in echelon form (even reduced echelon form). We can convert the equation

$$A\boldsymbol{x} = \boldsymbol{b}$$

to the equation

$$(E_r \cdots E_1 A)\boldsymbol{x} = E_r \cdots E_1 \boldsymbol{b}$$

which can be solved as above (if it is consistent). Any solution to $A\boldsymbol{x} = \boldsymbol{b}$ is a solution to this new equation.

III.11.7.8. Every elementary matrix is invertible, with the inverse another elementary matrix of the same form (every elementary row operation can be "undone" by another elementary row operation of the same type). Thus if the process of III.11.7.7. is carried out to obtain a solution $\boldsymbol{x}$ with

$$(E_r \cdots E_1 A)\boldsymbol{x} = E_r \cdots E_1 \boldsymbol{b}$$

we have that $\boldsymbol{x}$ is also a solution to the original equation:

$$A\boldsymbol{x} = (E_1^{-1} \cdots E_r^{-1})(E_r \cdots E_1 A)\boldsymbol{x} = (E_1^{-1} \cdots E_r^{-1})(E_r \cdots E_1)\boldsymbol{b} = \boldsymbol{b} \ .$$

III.11.7.9. We thus have a good algorithm for solving (finding all solutions to) a system of linear equations. There is convenient notation for the process: we form the $m \times (n+1)$ matrix

$$\left[\begin{array}{c|c} A & \boldsymbol{b} \end{array}\right]$$

called the *augmented matrix*, apply elementary row operations to the augmented matrix to reduce it to echelon form, and solve the new system as in III.11.7.3..

To do this, when we reduce to echelon form, we cannot obtain a matrix with a row which is all 0s except for a nonzero entry in the augmented ($(n+1)$'st) column. Such a row is called an *inconsistency* and implies that the system has no solution. We say the system is *consistent* if no inconsistency arises in reducing to echelon form. (The reduction to echelon form can be done in more than one way in general, but if an inconsistency arises from one reduction, it will arise also in any other reduction.) It is worth stating the conclusion as a proposition:

III.11.7.10. PROPOSITION. Let

$$A\boldsymbol{x} = \boldsymbol{b}$$

be a system of linear equations over a field $\mathbb{F}$. Then the system has a solution if and only if it is consistent.

Homogeneous Systems and General Solutions

We can be more specific about the set of solutions to the equation $A\boldsymbol{x} = \boldsymbol{b}$.

III.11.7.11. A system of linear equations of the form

$$A\boldsymbol{x} = \boldsymbol{0}$$

is called *homogeneous*.

A homogeneous system of equations always has at least one solution, $\boldsymbol{0}$. And we have:

III.11.7.12. PROPOSITION. Let $A\boldsymbol{x} = \boldsymbol{0}$ be a homogeneous system of equations, with A $m \times n$. Then the set of solutions is a subspace of $\mathbb{F}^n$.

In fact, the solution space is just the null space of A (III.11.3.7.).

The relevance to general systems is:

III.11.7.13. PROPOSITION. Let $A\boldsymbol{x} = \boldsymbol{b}$ be a consistent system of linear equations. Let $\boldsymbol{x}_0 \in \mathbb{F}^n$ be a solution. Then the solutions are precisely all vectors of the form $\boldsymbol{x}_0 + \boldsymbol{y}$, where $\boldsymbol{y}$ is a solution to the corresponding homogeneous system $A\boldsymbol{x} = \boldsymbol{0}$. Thus the solution set is an affine subspace () of $\mathbb{F}^n$.

PROOF: If $\boldsymbol{y}$ is a solution to the homogeneous system, then

$$A(\boldsymbol{x}_0 + \boldsymbol{y}) = A\boldsymbol{x}_0 + A\boldsymbol{y} = \boldsymbol{b} + \boldsymbol{0} = \boldsymbol{b}$$

so $\boldsymbol{x}_0 + \boldsymbol{y}$ is a solution to $A\boldsymbol{x} = \boldsymbol{b}$. Conversely, if $\boldsymbol{x}_1$ is a solution to $A\boldsymbol{x} = \boldsymbol{b}$, then

$$A(\boldsymbol{x}_1 - \boldsymbol{x}_0) = A\boldsymbol{x}_1 - A\boldsymbol{x}_0 = \boldsymbol{b} - \boldsymbol{b} = \boldsymbol{0}$$

so $\boldsymbol{x}_1 - \boldsymbol{x}_0$ is a solution to the homogeneous system, and $\boldsymbol{x}_1 = \boldsymbol{x}_0 + (\boldsymbol{x}_1 - \boldsymbol{x}_0)$. 👍

We can determine the dimension of this affine subspace (III.11.7.17.).

Row and Column Rank of a Matrix

III.11.7.14. Let A be an $m \times n$ matrix over a field $\mathbb{F}$, and let B be a matrix in echelon form obtained from A by elementary row operations. The following are easy observations left to the reader:

(i) The rows of B are linear combinations of the rows of A, and since A can be obtained from B by other elementary row operations, conversely the rows of A are linear combinations of the rows of B. Thus the linear spans of the rows of A and the rows of B are identical. This subspace of $\mathbb{F}^n$ is called the *row space* of A (or B), and its dimension is the *row rank* of A (or B), denoted $rr(A)$. The row rank of A is the number of linearly independent rows of A (or B).

(ii) The nonzero rows of B are linearly independent; thus the row rank of A (or B) is the number of nonzero rows of B.

(iii) A and B have the same null space.

Note also that the columns of A span the range of A (), and thus the number of linearly independent columns of A is the dimension of the range of A, called the *column rank* of A, denoted $cr(A)$. By (), we have

$$cr(A) + dim(\mathcal{N}(A)) = n \ .$$

III.11.7.15. Proposition. Let A be an $m \times n$ matrix over a field $\mathbb{F}$. Then $dim(\mathcal{N}(A)) = n - rr(A)$.

Proof: If B is obtained from A by elementary row operations, then $\mathcal{N}(B) = \mathcal{N}(A)$ and $rr(B) = rr(A)$. Thus we may assume A is in echelon form. Set $r = rr(A)$, the number of nonzero rows in A. Applying the algorithm of III.11.7.3. to solve the homogeneous system $A\boldsymbol{x} = \mathbf{0}$ to find the vectors in the null space, we have $n - r$ free choice variables, and once these have been chosen the other r variables are determined. For each of the $n - r$ free variables, construct a vector by choosing that free variable to be 1 and the other free variables 0, and the r pivot variables what they must be for this choice. We thus obtain $n - r$ vectors in $\mathcal{N}(A)$ which clearly span $\mathcal{N}(A)$, and they are easily seen to be linearly independent. 👍

We state the obvious consequence as a theorem:

III.11.7.16. Theorem. Let A be an $m \times n$ matrix over a field $\mathbb{F}$. Then $rr(A) = cr(A) = n - \dim(\mathcal{N}(A))$.

If $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$, a simpler proof can be given using inner products ().

III.11.7.17. Corollary. Let A be an $m \times n$ matrix over a field $\mathbb{F}$, and $\boldsymbol{b} \in \mathbb{F}^m$. If the system $A\boldsymbol{x} = \boldsymbol{b}$ is consistent, the set of solutions is an affine subspace of $\mathbb{F}^n$ of dimension $n - rr(A) = n - cr(A)$.

Invertibility and Uniqueness of Solutions

III.11.7.18. We now restrict to the case where $m = n$, i.e. A is an $n \times n$ matrix over $\mathbb{F}$. If A is invertible, any equation $A\boldsymbol{x} = \boldsymbol{b}$ has a solution, namely $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. We can be more specific and give a list of conditions equivalent to invertibility. By III.11.3.12., the following are equivalent:

(i) A is invertible.

(ii) $\mathcal{N}(A) = \{\mathbf{0}\}$, i.e. the equation $A\boldsymbol{x} = \mathbf{0}$ has only the solution $\boldsymbol{x} = \mathbf{0}$.

(iii) The column rank of A is n.

By III.11.7.16., (iii) is equivalent to

(iv) The row rank of A is n.

The only $n \times n$ matrix in reduced echelon form with row rank n is the identity matrix. Thus (iv) is equivalent to

(v) A can be reduced to I_n by elementary row operations.

(vi) A is a finite product of elementary matrices.

(vii) The system $A\boldsymbol{x} = \boldsymbol{b}$ is consistent (i.e. has a solution) for every $\boldsymbol{b} \in \mathbb{F}^n$.

By III.11.7.13., (ii) is equivalent to the statement that if $A\boldsymbol{x} = \boldsymbol{b}$ is consistent, then it has a unique solution. Since (ii) $\iff$ (vii), we obtain the equivalence of

(viii) The system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for every $\boldsymbol{b} \in \mathbb{F}^n$.

Let $\{\boldsymbol{e}_k : 1 \leq k \leq n\}$ be the standard basis vectors in $\mathbb{F}^n$. If (viii) holds, then

(ix) The equation $A\boldsymbol{x} = \boldsymbol{e}_k$ has a solution for each k.

(x) The equation $A\boldsymbol{x} = \boldsymbol{e}_k$ has a unique solution for each k.

hold. Conversely, if (ix) holds, let $\boldsymbol{x}_k$ satisfy the equation $A\boldsymbol{x}_k = \boldsymbol{e}_k$ for each k. If $\boldsymbol{b}$ is any vector in $\mathbb{F}^n$, write $\boldsymbol{b} = \sum_{k=1}^n \alpha_k \boldsymbol{e}_k$. Then

$$A\left(\sum_{k=1}^n \alpha_k \boldsymbol{x}_k\right) = \sum_{k=1}^n \alpha_k A\boldsymbol{x}_k = \sum_{k=1}^n \alpha_k \boldsymbol{e}_k = \boldsymbol{b}$$

so (vii) holds. Thus (ix) and (x) are equivalent to the previous conditions.

III.11.7..1 Finding the Inverse of a Matrix

Condition (x) leads to a computationally efficient way to find the inverse of an invertible matrix.

III.11.7.1. Suppose A is an invertible $n \times n$ matrix over $\mathbb{F}$. By (x), the equation $A\boldsymbol{x} = \boldsymbol{e}_k$ has a unique solution $\boldsymbol{x}_k$ for each k. Note that $\boldsymbol{x}_k$ is just the k'th column of A^{-1} by III.11.4.24.. The vector $\boldsymbol{x}_k$ can be found by forming the augmented matrix

$$\left[\begin{array}{c|c} A & \boldsymbol{e}_k \end{array}\right]$$

and reducing it to reduced echelon form. The A part will reduce to I_n, and it is then easily seen that the augmented column must be $\boldsymbol{x}_k$.

III.11.7.2. We may efficiently do all the augmentations at once. Form the $n \times 2n$ matrix

$$\left[\begin{array}{c|c} A & I_n \end{array}\right]$$

and row reduce to reduced echelon form. We obtain

$$\left[\begin{array}{c|c} I_n & A^{-1} \end{array}\right]$$

so A^{-1} can be found by elementary row reductions. (This amounts to writing A as a product of elementary matrices, so A^{-1} is the product of the inverses, but is an efficient way to do this.)

If A is not invertible, if this process is tried the matrix on the left will reduce to a matrix with a row of 0s and the I_n will not, so an inconsistency will arise and there will be, as expected, no solution.

III.11.8. Affine Subspaces and Affine Functions

General Position

A useful variation of linear independence is general position.

III.11.8.1. Definition. Let $\mathcal{V}$ be a vector space over $\mathbb{F}$, and S a subset of $\mathcal{V}$. Then S is in *general position*, or the points in S are in general position, if, whenever $\{x_0, x_1, \ldots, x_m\}$ is a set of $m+1$ points in S with $m \leq dim(\mathcal{V})$, then $\{x_1 - x_0, \ldots, x_m - x_0\}$ is linearly independent.

This definition is applied most often when S is a finite subset of $\mathcal{V}$, but makes sense in general. It is most interesting when the size of S is at least the dimension of $\mathcal{V}$.

The proof of the next proposition is nearly trivial, and is left to the reader.

III.11.8.2. Proposition. Let $\mathcal{V}$ be an n-dimensional vector space over $\mathbb{F}$, and S a subset of $\mathcal{V}$ with r elements ($r = \infty$ is allowed). Set $m = \min(r, n+1)$. Then the following are equivalent:

(i) S is in general position.

(ii) Whenever $\{x_0, \ldots, x_{m-1}\}$ is a subset of S with m distinct elements, $\{x_1 - x_0, \ldots, x_{m-1} - x_0\}$ is linearly independent.

(iii) Whenever $\{x_0, \ldots, x_{m-1}\}$ is a subset of S with m distinct elements, the smallest affine subspace of $\mathcal{V}$ containing $\{x_0, \ldots, x_{m-1}\}$ has dimension $m-1$.

(iv) Whenever $\{x_0, \ldots, x_k\}$ is a subset of S with k distinct elements ($k \leq m-1$), the smallest affine subspace of $\mathcal{V}$ containing $\{x_0, \ldots, x_k\}$ has dimension k.

III.11.8.3. If $\mathbb{F}$ is a finite field and $\mathcal{V}$ is an n-dimensional vector space over $\mathbb{F}$, then there is a limit to the size of a set in general position. But if $\mathbb{F}$ is infinite, there can be infinite sets in general position. For example, if $\mathcal{V}$ is 1-dimensional, any set of vectors is in general position.

III.11.8.4. EXAMPLE. For a set S to be in general position, it is not enough that there be an $x_0 \in S$ such that, whenever $\{x_1, \ldots, x_{m-1}\}$ is a subset of S with $m-1$ distinct elements all distinct from x_0, then $\{x_1 - x_0, \ldots, x_{m-1} - x_0\}$ is linearly independent: consider the set

$$\{(0,0), (2,0), (0,2), (1,1)\} \subseteq \mathbb{R}^2 \ .$$

III.11.8.5. Any subset of a set in general position is in general position. A set of at most $n+1$ vectors in an n-dimensional vector space which is in general position is called *affinely independent*. If $\{x_0, \ldots, x_n\}$ is affinely independent, then any subset of size $k+1$ generates a k-dimensional affine subspace.

III.11.9. The Dot Product on $\mathbb{R}^n$ and Adjoints

We now discuss a piece of algebraic structure unique to $\mathbb{R}^n$ (there is a similar structure on $\mathbb{C}^n$).

One of the big algebraic things distinguishing $\mathbb{R}^n$, $n > 1$, from $\mathbb{R}$ is that there is no way to multiply (or divide) vectors. (There is a nonobvious way to define vector multiplication in $\mathbb{R}^n$ for $n = 2, 4$, or 8, cf. III.6.2.2. and III.6.3.2., but for no other $n > 1$.) However, there is a simple and very important way to "multiply" two vectors and obtain a scalar, called the dot product or inner product.

III.11.9.1. DEFINITION. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ be vectors in $\mathbb{R}^n$. The *dot product* of $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$\boldsymbol{x} \cdot \boldsymbol{y} = (x_1, \ldots, x_n) \cdot (y_1, \ldots, y_n) = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{k=1}^{n} x_k y_k \ .$$

Note that $\boldsymbol{x} \cdot \boldsymbol{y} \in \mathbb{R}$, i.e. the dot product defines a function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$.

III.11.9.2. The dot product $\boldsymbol{x} \cdot \boldsymbol{y}$ is sometimes written $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ (other notations are also occasionally used). The dot product is an example of an inner product, discussed in XX.11.1..

The next proposition lists the important algebraic properties of the dot product. The results are either obvious or proved by straightforward calculations, and the proofs are left to the reader.

III.11.9.3. PROPOSITION. Let $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ be vectors in $\mathbb{R}^n$. Then:

(i) $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}$.

(ii) $\boldsymbol{x} \cdot (\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{x} \cdot \boldsymbol{z}$.

(ii') $(\boldsymbol{x} + \boldsymbol{y}) \cdot \boldsymbol{z} = \boldsymbol{x} \cdot \boldsymbol{z} + \boldsymbol{y} \cdot \boldsymbol{z}$.

(iii) $(\alpha \boldsymbol{x}) \cdot \boldsymbol{y} = \alpha(\boldsymbol{x} \cdot \boldsymbol{y})$ for any $\alpha \in \mathbb{R}$. In particular, $\boldsymbol{0} \cdot \boldsymbol{y} = 0$.

(iii') $\boldsymbol{x} \cdot (\alpha \boldsymbol{y}) = \alpha(\boldsymbol{x} \cdot \boldsymbol{y})$ for any $\alpha \in \mathbb{R}$. In particular, $\boldsymbol{x} \cdot \boldsymbol{0} = 0$.

(iv) $\boldsymbol{x} \cdot \boldsymbol{x} = \|\boldsymbol{x}\|^2$. In particular, $\boldsymbol{x} \cdot \boldsymbol{x} \geq 0$, and $\boldsymbol{x} \cdot \boldsymbol{x} = 0$ implies $\boldsymbol{x} = \boldsymbol{0}$.

(v) $\boldsymbol{x} \cdot \boldsymbol{y} = \frac{1}{4}[\|\boldsymbol{x} + \boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2]$.

Property (i) is called *symmetry*; properties (ii)-(ii')-(iii)-(iii') are together called *bilinearity* (note that because of (i), (ii') and (iii') follow from (ii) and (iii) respectively). The last part of (iv) is called *positive definiteness*, and (v) is the *polarization identity*.

III.11.9.4. Corollary. Let T be a linear isometry of $\mathbb{R}^n$ into $\mathbb{R}^m$. If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, then

$$(T\boldsymbol{x}) \cdot (T\boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y} \ .$$

Proof: Apply polarization (III.11.9.3.(v)). 👍

There is a nice geometric interpretation of the dot product. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are nonzero vectors in $\mathbb{R}^n$, there is a well-defined angle θ between them if we take $0 \leq \theta \leq \pi$, measured in the two-dimensional subspace spanned by $\boldsymbol{x}$ and $\boldsymbol{y}$.

III.11.9.5. Proposition. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be nonzero vectors in $\mathbb{R}^n$, and θ the angle between them. Then

$$\boldsymbol{x} \cdot \boldsymbol{y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\| \cos\theta \ .$$

Proof: Since there is a linear isometry from $\mathbb{R}^2$ onto the two-dimensional subspace of $\mathbb{R}^n$ spanned by $\boldsymbol{x}$ and $\boldsymbol{y}$, by III.11.9.4. we may assume $n = 2$. And by III.11.9.3.(iii)-(iii') we may assume $\boldsymbol{x}$ and $\boldsymbol{y}$ are unit vectors.

There is a rotation (isometry) T of $\mathbb{R}^2$ sending $\boldsymbol{x}$ to $(1, 0)$. We then have $T\boldsymbol{y} = (\cos\theta, \pm\sin\theta)$, so

$$\boldsymbol{x} \cdot \boldsymbol{y} = (T\boldsymbol{x}) \cdot (T\boldsymbol{y}) = (1, 0) \cdot (\cos\theta, \pm\sin\theta) = \cos\theta \ .$$

👍

III.11.9.6. Corollary. [Cauchy's Inequality or CBS Inequality] Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be vectors in $\mathbb{R}^n$. Then

$$|\boldsymbol{x} \cdot \boldsymbol{y}| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\| \ .$$

Proof: If $\boldsymbol{x}$ and $\boldsymbol{y}$ are nonzero, let θ be the angle between them. Then

$$|\boldsymbol{x} \cdot \boldsymbol{y}| = \|\boldsymbol{x}\|\|\boldsymbol{y}\||\cos\theta| \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$$

since $|\cos\theta| \leq 1$. And if $\boldsymbol{x}$ or $\boldsymbol{y}$ is $\boldsymbol{0}$, both sides are 0. 👍

III.11.9.7. This inequality is commonly called the *Cauchy-Schwarz inequality* or even the *Schwarz inequality*. But it was first stated and proved in this form by CAUCHY in 1821, sixty years before SCHWARZ. For a more direct proof and extensions to general inner products, and a discussion of the history of the inequality and its names, see XX.11.2..

III.11.9.8. PROPOSITION. The dot product on $\mathbb{R}^n$ is jointly continuous: if $\boldsymbol{x}_k \to \boldsymbol{x}$ and $\boldsymbol{y}_k \to \boldsymbol{y}$, then $\boldsymbol{x}_k \cdot \boldsymbol{y}_k \to \boldsymbol{x} \cdot \boldsymbol{y}$.

PROOF: This follows easily from the CBS inequality, or just the fact that the formula is a polynomial function in the entries. 👍

III.11.10. Determinants

There is a number (scalar) called the *determinant* associated with any square matrix, which is of both theoretical and practical significance. In particular, a (square) matrix is invertible if and only if its determinant is nonzero.

Determinants have gotten somewhat of a bad rap. It is widely fashionable in linear algebra (cf. [Axl95]) to try to banish determinants from the subject. Indeed, elementary linear algebra can be efficiently and elegantly developed without reference to determinants (e.g. [Axl97]). Determinants are mostly excluded from numerical linear algebra. There is one good overriding reason for avoiding determinants wherever possible, especially in numerical work: *determinants are hard to compute*!

However, determinants cannot be entirely avoided, especially in applications to analysis. The reason is that the determinant has a natural and crucial geometric or measure-theoretic interpretation, and hence appears naturally in such analysis formulas as the change-of-variables formula for integrals (). (And modern numerical analysis techniques make determinants somewhat more efficient to compute than by the straightforward brute-force method.)

We will therefore review the basics of the theory of determinants from the analysis point of view. Note that the determinant (as we do it here) is strictly a finite-dimensional notion.

The Determinant of a Matrix

There are two approaches to defining the determinant of a matrix. From an abstract point of view, the "best" way is to prove the existence and uniqueness theorem (III.11.10.27.) from scratch and then derive a formula for the determinant from it. We will take a lower-tech approach, defining the determinant by a (fairly complicated) algebraic formula, and deriving its properties from the formula.

III.11.10.1. DEFINITION. Let $A = [a_{ij}]$ be an $n \times n$ matrix over $\mathbb{F}$. The *determinant* of A is the scalar

$$det(A) = \sum_{\sigma \in S_n} sgn(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

where S_n is the set of permutations () of $\{1, \ldots, n\}$ and $sgn(\sigma) = \pm 1$ is the sign () of the permutation σ.

There are $n!$ terms, each of which is a product of n entries of A, exactly one from each row and column, with all possible such products represented.

To see how this works, we write out the determinant explicitly for small n:

III.11.10.2. EXAMPLES. (i) $n = 1$: if $A = [a]$ is a 1×1 matrix (scalar), then $det(A) = a$ since there is only one permutation of $\{1\}$.

(ii) $n = 2$: Let $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$. There are two permutations of $\{1, 2\}$, the identity ι and the permutation σ, where $\sigma(1) = 2$ and $\sigma(2) = 1$, and $sgn(\iota) = 1$, $sgn(\sigma) = -1$. Thus we have

$$det(A) = sgn(\iota)a_{1\iota(1)}a_{2\iota(2)} + sgn(\sigma)a_{1\sigma(1)}a_{2\sigma(2)} = a_{11}a_{22} - a_{12}a_{21} \ .$$

This formula is usually remembered as:

$$det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc \ .$$

(iii) $n = 3$: We write this case out explicitly, but the formula more serves as a warning of how rapidly the determinant formula becomes complicated as n increases than as a useful formula. There are 6 permutations of $\{1, 2, 3\}$. Rather than working through all the details, we simply give the end result:

$$det\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right) = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} \ .$$

(iv) $n = 4$: This formula has 24 terms and exceeds the patience of the author to write.

We will obtain a somewhat simpler inductive method of computing determinants in (); but computing the determinant of even a general 10×10 matrix is a computational challenge on a computer (there are 10! =3,628,800 terms).

III.11.10.3. The notation $|A|$ is also frequently used for $det(A)$, especially when the matrix A is written out in an array. This notation can be somewhat confusing, since the determinant has nothing to do with absolute value (e.g. the determinant of a matrix with real entries can be negative). Besides, in analysis we often need to work with the absolute value of the determinant, possibly adding to the confusion. We will stick with the notation $det(A)$.



Elementary Properties of Determinants

The first observation is immediate from the definition of determinant:

III.11.10.4. PROPOSITION. For any n, the determinant of the identity matrix I_n is 1. More generally, the determinant of any diagonal matrix is the product of the diagonal elements.

In fact, it is only slightly harder to show that the determinant of any triangular matrix is the product of the diagonal elements.

Another immediate consequence of the definition is:

III.11.10.5. PROPOSITION. Let A be an $n \times n$ matrix over $\mathbb{F}$. If A has a zero row or zero column, then $det(A) = 0$.

This result has a generalization ().

The next result is also a simple consequence of the definition.

III.11.10.6. PROPOSITION. Let A be an $n \times n$ matrix over $\mathbb{F}$, and $r \in \mathbb{F}$. Let B be the matrix obtained from A by multiplying one row of A by r and leaving all other rows unchanged. Then

$$det(B) = r \cdot det(A) \ .$$

Note that the B in the proposition is *not* a scalar multiple of A. In fact, by applying the proposition successively to each row, we obtain that $det(rA) = r^n \cdot det(A)$.
PROOF: Comparing the formulas for $det(B)$ and $det(A)$, each term in the sum for $det(B)$ is obtained from the corresponding term for $det(A)$ by multiplying one of the factors by r and leaving the rest unchanged. Thus a factor of r can be taken out of the sum by the distributive property. 👍

The next result is more esoteric to state, but has a similarly simple proof:

III.11.10.7. PROPOSITION. Let $B = [b_{ij}]$ and $C = [c_{ij}]$ be $n \times n$ matrices which are identical except for some row, i.e. for which there is an i_0 such that $b_{ij} = c_{ij}$ for all $i \neq i_0$ and all j. Let $A = [a_{ij}]$ be the matrix with $a_{ij} = b_{ij} = c_{ij}$ for $i \neq i_0$ and all j, and $a_{i_0 j} = b_{i_0 j} + c_{i_0 j}$ for all j (i.e. the i_0'th row of A is the sum of the i_0'th rows of B and C and the other rows of A are identical to the other rows of B and C). Then

$$det(A) = det(B) + det(C) \ .$$

PROOF: We have

$$det(A) = \sum_{\sigma \in S_n} sgn(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)} = \sum_{\sigma \in S_n} sgn(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots (b_{i_0\sigma(i_0)} + c_{i_0\sigma(i_0)}) \cdots a_{n\sigma(n)}$$

$$= \sum_{\sigma \in S_n} sgn(\sigma) b_{1\sigma(1)} b_{2\sigma(2)} \cdots b_{n\sigma(n)} + \sum_{\sigma \in S_n} sgn(\sigma) c_{1\sigma(1)} c_{2\sigma(2)} \cdots c_{n\sigma(n)} = det(B) + det(C) \ .$$

👍

III.11.10.8. The previous two results are abbreviated by saying that the determinant is *multilinear* (or, more specifically, *n-linear*) in the rows of the matrix. See () for the general theory of multilinear functions.

Next we obtain a property which is simpler to state but a little trickier to prove:

III.11.10.9. PROPOSITION. Let A be an $n \times n$ matrix over $\mathbb{F}$, and B a matrix obtained by interchanging two rows of A. Then

$$det(B) = -det(A) \ .$$

PROOF: Suppose the i_1'th and i_2'th rows of A are interchanged. Let τ be the permutation of $\{1, \ldots, n\}$ which interchanges i_1 and i_2 and leaves everything else fixed. Then τ is a transposition, so $sgn(\tau) = -1$ and, more generally, $sgn(\sigma\tau) = -sgn(\sigma)$ for every $\sigma \in S_n$. By rearranging the order of factors in each term, we have

$$det(B) = \sum_{\sigma \in S_n} sgn(\sigma) a_{1\sigma\tau(1)} \cdots a_{n\sigma\tau(n)} = -\sum_{\sigma \in S_n} sgn(\sigma\tau) a_{1\sigma\tau(1)} \cdots a_{n\sigma\tau(n)} \ .$$

But as σ ranges over S_n, so does $\sigma\tau$; so the last sum is exactly $det(A)$. 👍

III.11.10.10. This property of the determinant is called the *alternating* property. It implies:

III.11.10.11. COROLLARY. Let A be an $n \times n$ matrix. If A has two identical rows, then $det(A) = 0$.

(This is actually not a corollary of III.11.10.9. if $\mathbb{F}$ has characteristic 2. But this result has a direct proof which works also in the characteristic 2 case; see Exercise III.11.12.1..)

Combining this with III.11.10.6., we obtain an extension:

III.11.10.12. COROLLARY. Let A be an $n \times n$ matrix over $\mathbb{F}$. If one row of A is a scalar multiple of another row, then $det(A) = 0$.

PROOF: If the i_2'th row of A is r times the i_1'th row, then by III.11.10.6., $det(A) = r \cdot det(B)$ for a matrix B with two identical rows. 👍

This will be further extended in ().

Now combining this with III.11.10.7., we obtain:

III.11.10.13. COROLLARY. Let A be an $n \times n$ matrix over $\mathbb{F}$, and let B be the matrix obtained from A by replacing one row by that row plus a scalar multiple of another row. Then $det(B) = det(A)$.

PROOF: Let C be the matrix obtained from A by replacing the i_1'th row by r times the i_2'th row for some $r \in R$. If $b_{i_1 j} = a_{i_1 j} + r a_{i_2 j}$ and $b_{ij} = a_{ij}$ for $i \neq i_1$, then by III.11.10.7. we have $det(B) = det(A) + det(C)$. But $det(C) = 0$ by III.11.10.12.. 👍

Another important property of determinants has a proof similar to the proof of III.11.10.9..

III.11.10.14. PROPOSITION. Let A be an $n \times n$ matrix over $\mathbb{F}$. Then $det(A^t) = det(A)$.

PROOF: We have

$$det(A^t) = \sum_{\sigma \in S_n} sgn(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} \cdots a_{\sigma(n)n} \ .$$

Since $sgn(\sigma^{-1}) = sgn(\sigma)$ for any σ, by rearranging the order of the factors in each term we get

$$det(A^t) = \sum_{\sigma \in S_n} sgn(\sigma^{-1}) a_{1\sigma^{-1}(1)} a_{2\sigma^{-1}(2)} \cdots a_{n\sigma^{-1}(n)} \ .$$

But as σ ranges over S_n, so does σ^{-1}, hence the last sum is just $det(A)$. 👍

As a result, the previous results on rows apply also to columns:

III.11.10.15. COROLLARY. Let A be an $n \times n$ matrix over $\mathbb{F}$.
(i) The determinant is multilinear in the columns of A. In particular, if B is obtained from A by multiplying a column of A by a scalar r, then $det(B) = r \cdot det(A)$.
(ii) If B is obtained from A by interchanging two columns, then $det(B) = -det(A)$.
(iii) If A has two identical columns, or if one column of A is a scalar multiple of another column, then $det(A) = 0$.
(iv) If B is obtained from A by replacing one column of A by that column plus a scalar times another column, then $det(B) = det(A)$.

There are many other properties of determinants discussed in abstract algebra books. We mention only a couple which are relevant for analysis. The proofs are quite obvious from the definition.

III.11.10.16. PROPOSITION. (i) Let A be an $n \times n$ matrix with entries in $\mathbb{Z}$. Then $det(A) \in \mathbb{Z}$.
(ii) Let $A = [a_{ij}]$ be an $n \times n$ matrix with entries in $\mathbb{C}$, and $\bar{A} = [\bar{a}_{ij}]$ be the complex conjugate matrix. Then $det(\bar{A}) = \overline{det(A)}$.

Combining (ii) with III.11.10.14., we obtain:

III.11.10.17. COROLLARY. Let A be an $n \times n$ matrix over $\mathbb{C}$, and A^* its adjoint matrix (). Then $det(A^*) = \overline{det(A)}$.

III.11.10.18. In fact, the theory of determinants works equally well over any *commutative* ring (some aspects require the ring to be unital): if R is a commutative ring, for any n there is a determinant function from $M_n(R)$ to R which is n-linear and alternating, defined by the same formula as in III.11.10.1.. If R and S are commutative rings and $\phi : R \to S$ is a ring-homomorphism, then there is an induced coordinatewise ring homomorphism ϕ_n from $M_n(R)$ to $M_n(S)$ for any n, and the determinant is *natural*, i.e. $det(\phi_n(A)) = \phi(det(A))$ for any $A \in M_n(R)$. See also III.11.10.53..

However, determinants simply do not work for matrices over noncommutative rings; there is no determinant-like function with reasonable properties on $M_n(R)$ for noncommutative R.

Expansion by Minors

The multilinear and alternating properties of the determinant lead to an inductive procedure for computing the determinant of an $n \times n$ matrix in terms of determinants of smaller submatrices. This procedure, called *expansion by minors* or *expansion by cofactors*, is the way most determinants (at least ones of modest size) are computed in practice.

We first define a set of submatrices of an $n \times n$ matrix:

III.11.10.19. Definition. Let A be an $n \times n$ matrix over $\mathbb{F}$, and $1 \le i, j \le n$. The (i, j) *minor* of A is the $(n-1) \times (n-1)$ matrix A^m_{ij} obtained from A by deleting the i'th row and j'th column. The (i, j) *cofactor* of A is $d_{ij}(A) = (-1)^{i+j} det(A^m_{ij})$.

The $(-1)^{i+j}$ in the definition of cofactor is conventional, and simplifies later formulas.

III.11.10.20. Theorem. Let A be an $n \times n$ matrix over $\mathbb{F}$. For any fixed i_0, $1 \le i_0 \le n$, we have

$$det(A) = \sum_{j=1}^{n} a_{i_0 j} \cdot d_{i_0 j}(A) = \sum_{j=1}^{n} (-1)^{i_0+j} a_{i_0 j} \cdot det(A^m_{i_0 j}) \ .$$

For any fixed j_0, $1 \le j_0 \le n$, we have

$$det(A) = \sum_{i=1}^{n} a_{i j_0} \cdot d_{i j_0}(A) = \sum_{i=1}^{n} (-1)^{i+j_0} a_{i j_0} \cdot det(A^m_{i j_0}) \ .$$

Thus the computation of an $n \times n$ determinant can be reduced to the computation of n determinants, each $(n-1) \times (n-1)$. Each of these can be calculated by computing $n-1$ determinants, each $(n-2) \times (n-2)$, etc. It does not matter which row or column is used at each stage; as a practical matter, if the matrix has some zero entries it usually saves work to choose the row or column with the most zeroes at each stage.

III.11.10.21. Example. Expanding along the first row ($i_0 = 1$), we have

$$det\left(\begin{bmatrix} 1 & 2 & -1 \\ 3 & -2 & 1 \\ 1 & 0 & -4 \end{bmatrix}\right) = (-1)^2(1) det\left(\begin{bmatrix} -2 & 1 \\ 0 & -4 \end{bmatrix}\right) + (-1)^3(2) det\left(\begin{bmatrix} 3 & 1 \\ 1 & -4 \end{bmatrix}\right) + (-1)^4(-1) det\left(\begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}\right)$$

$$= 8 - 2(-13) - 2 = 32 \ .$$

Expanding along the second column ($j_0 = 2$), we have

$$det\left(\begin{bmatrix} 1 & 2 & -1 \\ 3 & -2 & 1 \\ 1 & 0 & -4 \end{bmatrix}\right) = (-1)^3(2)det\left(\begin{bmatrix} 3 & 1 \\ 1 & -4 \end{bmatrix}\right)+(-1)^4(-2)det\left(\begin{bmatrix} 1 & -1 \\ 1 & -4 \end{bmatrix}\right)+(-1)^5(0)det\left(\begin{bmatrix} 1 & -1 \\ 3 & 1 \end{bmatrix}\right)$$

$$= -2(-13) - 2(-3) + 0 = 32 \ .$$

The second way is slightly simpler since one term is zero and only two 2×2 determinants must be computed.

Determinants and Elementary Matrices

There is another way to view the multilinear and alternating properties of the determinant. Each describes the behavior of the determinant under an elementary row operation. Each elementary row operation on a matrix A can be performed by multiplying on the left by an elementary matrix. We can easily compute the determinant of elementary matrices:

III.11.10.22. PROPOSITION. Determinants of elementary matrices are as follows:

(i) If $\Pi_{(ij)}$ is the matrix that interchanges i and j, then $det(\Pi_{(ij)}) = -1$.

(ii) If $M_i(r)$ is the diagonal matrix with (i, i) entry r and the rest of the diagonal entries 1, then $det(M_i(r)) = r$.

(iii) If $E_{ij}(r)$ $(i \neq j)$ is the matrix obtained from I_n by replacing the (i, j) element by r, then $det(E_{ij}(r)) = 1$.

This can be proved by applying III.11.10.9. for (i), III.11.10.6. (or just III.11.10.4.) for (ii), and III.11.10.13. for (iii) (or readily proved directly from the definition of determinant).

Comparing the effect of row operations on the determinant with III.11.10.22., we obtain:

III.11.10.23. COROLLARY. Let A be an $n \times n$ matrix over $\mathbb{F}$, and let E be an elementary $n \times n$ matrix over $\mathbb{F}$. Then $det(EA) = det(E) \cdot det(A)$.

Multiplicative Property of Determinants

The two most important properties of determinants for most applications are simple consequences of III.11.10.23. and ().

III.11.10.24. THEOREM. Let A be an $n \times n$ matrix over $\mathbb{F}$. Then A is invertible if and only if $det(A) \neq 0$.

PROOF: By (), there are finitely many elementary matrices $E_1, \ldots, E_m$ such that either $E_1 E_2 \cdots E_m A = I_n$ (if A is invertible) or $E_1 E_2 \cdots E_m A$ has a zero row (if A is singular). In the first case we have

$$A = E_m^{-1} E_{m-1}^{-1} \cdots E_1^{-1}$$

so A is a product of elementary matrices. By repeated applications of III.11.10.23., we obtain

$$det(A) = det(E_m^{-1}) \cdot det(E_{m-1}^{-1}) \cdots det(E_1^{-1}) \neq 0 \ .$$

In the second case, we have from III.11.10.23. that $det(E_1 \cdots E_m A)$ is a nonzero multiple of $det(A)$. But $E_1 \cdots E_m A$ has a zero row, so $det(E_1 \cdots E_m A) = 0$ and hence $det(A) = 0$. 👍

III.11.10.25. THEOREM. Let A and B be $n \times n$ matrices over $\mathbb{F}$. Then $det(AB) = det(A) \cdot det(B)$.

PROOF: If A is not invertible, then AB is not invertible () and $det(AB) = 0 = det(A) \cdot det(B)$. And if A is invertible, then A is a product of elementary matrices, so by repeated applications of III.11.10.23., we obtain that $det(AB) = det(A) \cdot det(B)$. 👍

Proving this result by direct computation from the definition of determinant is a mess even for $n = 2$, and becomes impossibly complicated for larger n. The result holds for matrices over general commutative rings, although it needs a different proof there (cf. III.11.10.28.).

III.11.10.26. In the case of $\mathbb{F} = \mathbb{R}$, the result follows immediately from the measure-scaling property of the determinant (). Note, however, that in establishing the properties of the determinant it is important to avoid circularity in arguments; our proof of the measure-scaling property () uses the multiplicativity of the determinant, so while the measure-scaling property is a good geometric *explanation* of multiplicativity, it is not a good way to prove it.

The Uniqueness Theorem

It turns out that the properties we have obtained uniquely characterize the determinant function:

III.11.10.27. THEOREM. Let $\mathbb{F}$ be a field.

(i) For any $n \in \mathbb{N}$ there is a unique function $\phi : M_n(\mathbb{F}) \to \mathbb{F}$ which is n-linear in the rows, alternating, and such that $\phi(I_n) = 1$. We have $\phi(A) = det(A)$ for all $A \in M_n(\mathbb{F})$.

(ii) If $\phi : M_n(\mathbb{F}) \to \mathbb{F}$ is n-linear in the rows, alternating, and $\phi(I_n) = 0$, then ϕ is identically zero.

See III.11.17.38. for the proof.

III.11.10.28. An alternate proof of III.11.10.25. can be based on the Uniqueness Theorem:
PROOF: Fix $B \in M_n(\mathbb{F})$. It is routine to check that the function $\phi : M_n(\mathbb{F}) \to \mathbb{F}$ given by $\phi(A) = det(AB)$ is n-linear and alternating. We have $\phi(I) = det(B)$. If $det(B) = 0$, then ϕ is identically zero by III.11.10.27.(ii); thus $det(AB) = 0 = det(A) \cdot det(B)$ for any A. And if $det(B) \neq 0$, set

$$\psi(A) = \frac{1}{det(B)}\phi(A) = \frac{1}{det(B)}det(AB) \ .$$

By III.11.10.27.(i), $\psi(A) = det(A)$ for all A, and thus $\phi(A) = det(AB) = det(A) \cdot det(B)$ for any A. ☝

Continuity and Differentiability of the Determinant

The Determinant of a Linear Transformation

III.11.10.29. If $\mathcal{V}$ is a finite-dimensional vector space over $\mathbb{F}$, and T is a linear transformation from $\mathcal{V}$ to $\mathcal{V}$, then T does not have a canonical matrix representation; to represent T by a matrix, a basis for $\mathcal{V}$ must be chosen (). Although the matrix of T depends on the choice of basis, there is not too much flexibility: if B and B' are bases, then the matrices of T with respect to B and B' are similar:

$$[T]_{B'} = P_{B,B'}^{-1}[T]_B P_{B,B'}$$

where $P_{B,B'}$ is the change-of-basis matrix (), and so we have

$$det([T]_{B'}) = det(P_{B,B'})^{-1}det([T]_B)det(P_{B,B'}) = det([T]_B)$$

(using commutativity of multiplication in $\mathbb{F}$), i.e. the determinant of $[T]_B$ does not depend on the choice of basis B. Thus the linear transformation T has a well-defined associated scalar, the determinant of the matrix representation of T with respect to any basis.

III.11.10.30. DEFINITION. Let T be a linear transformation from a finite-dimensional vector space $\mathcal{V}$ over $\mathbb{F}$ to itself. Set

$$det(T) = det([T]_B)$$

where B is any basis for $\mathcal{V}$ over $\mathbb{F}$.

III.11.10.31. Note that if S, T are linear transformations from $\mathcal{V}$ to $\mathcal{V}$, then

$$det(S \circ T) = det(S)det(T) \ .$$

III.11.10.32. The determinant of a linear transformation does depend on the choice of base field $\mathbb{F}$, however. If $\mathcal{V}$ is a vector space over $\mathbb{C}$ and $T : \mathcal{V} \to \mathcal{V}$ a linear transformation, then $det(T)$ is not the same if $\mathcal{V}$ is regarded as a vector space over $\mathbb{R}$ as it is over $\mathbb{C}$ in general. (See Exercise () for the relationship.)

Geometric Interpretation of the Determinant

Although the theory of determinants is purely algebraic and works equally well over any field (in fact, over any commutative ring), there is a geometric interpretation of the determinant which is exclusive to the case where the field $\mathbb{F}$ is $\mathbb{R}$ (and, to a limited extent, to the case $\mathbb{F} = \mathbb{C}$). This interpretation underlies the importance of the determinant in analysis.

III.11.10.33. Volume Principle of Determinants: Let A be an $n \times n$ matrix with real entries. Regard A as giving a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ (with respect to the standard basis). Then T multiplies n-dimensional "volume" by a constant factor $|det(A)|$, i.e. if E is a subset of $\mathbb{R}^n$ with a well-defined n-dimensional "volume" $V(E)$ (technically a Jordan region () or, more generally, a Lebesgue measurable subset), then $T(E)$ also has a well-defined n-dimensional volume and $V(T(E)) = |det(A)| \cdot V(E)$.

III.11.10.34. Actually, we should first define what we mean by n-dimensional "volume" of a subset of $\mathbb{R}^n$. This is not as easy as it sounds: the general problem is essentially the basic subject matter of measure theory. It would take us too far afield to try to treat this question here, so we will just assume that "volume" is defined for sufficiently well-behaved subsets of $\mathbb{R}^n$, and satisfies

(i) If R is a "rectangle" $[a_1, b_1] \times \cdots \times [a_n, b_n]$, then $V(R) = (b_1 - a_1) \cdots (b_n - a_n)$.

(ii) If $E \subseteq F$, then $V(E) \leq V(F)$.

(iii) If $E_1, \ldots, E_m$ are nonoverlapping regions ("nonoverlapping" means the intersection of any two is contained in the boundary of both), then

$$V(E_1 \cup \cdots \cup V_m) = V(E_1) + \cdots + V(E_m).$$

(iv) Congruent regions (including allowing reflections) have the same volume.

(There is actually a subtle technical problem with (iii), since the boundary of a region can have positive measure, but this will not cause difficulty if we consider only regions bounded by finitely many smooth hypersurfaces. Actually we will only need to work with regions bounded by finitely many hyperplanes.)

III.11.10.35. We first look at the case $n = 2$, which is the only one we will examine in detail. Here 2-dimensional "volume" is "area." Suppose

$$A = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

is a 2×2 matrix. The $T(1, 0) = (a, b)$ and $T(0, 1) = (c, d)$. The image of the unit square $[0, 1] \times [0, 1]$ is the parallelogram with vertices $(0, 0)$, (a, b), (c, d), and $(a + c, b + d)$. We claim the area of this parallelogram is $|det(A)| = |ad - bc|$.

The argument reduces to several special cases. First consider the case where $a > 0$, $b \geq 0$, $d > 0$, $c \geq 0$, and either $c = 0$ or $\frac{d}{c} > \frac{b}{a}$ (Figure III.1).

The area of the total rectangle is $(a + c)(b + d)$. The blue triangles fit together into a rectangle of length a and height b, so the blue area is ab. Similarly, the green area is cd. The two red rectangles each have area bc. Thus the area of the parallelogram is

$$(a + c)(b + d) - ab - cd - 2bc = ab + ad + bc + cd - ab - cd - 2bc = ad - bc \ .$$

The case where $\frac{d}{c} < \frac{b}{a}$ can be handled by interchanging columns of A, which multiplies the determinant by -1.

III.11.10.36. Now look at the case $c < 0$, but $a, b, d \geq 0$ (Figure III.2).

The area of the large rectangle is $(a - c)(b + d)$. The blue triangles fit together into a rectangle of area ab, and the green triangles have total area $-cd$. So the area of the parallelogram is

$$(a - c)(b + d) - ab + cd = ab + ad - bc - cd - ab + cd = ad - bc \ .$$

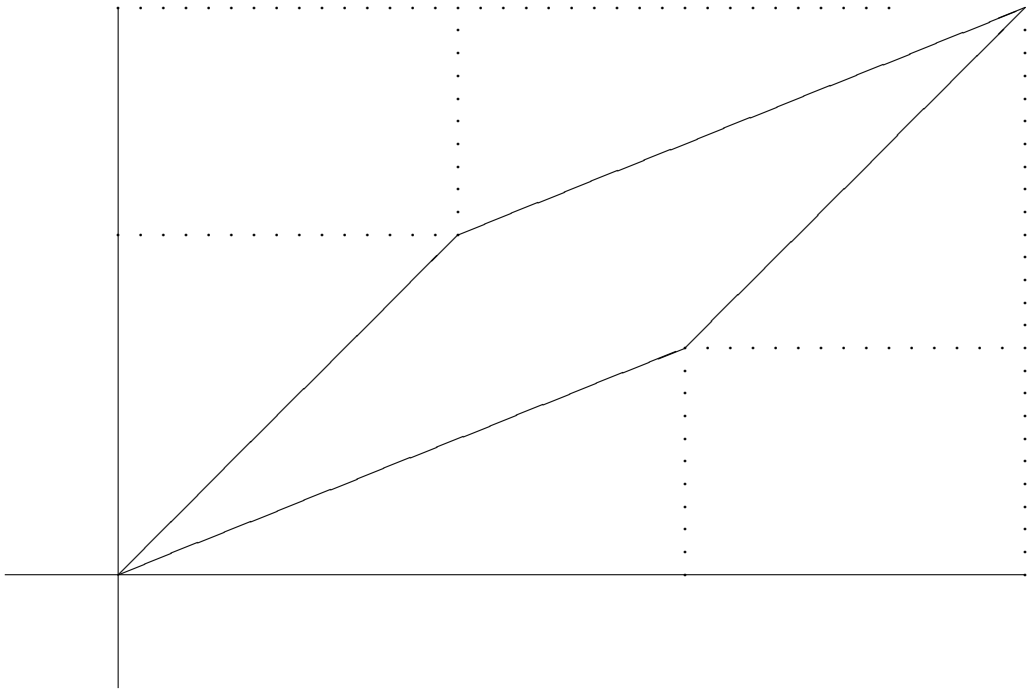


Figure III.1: The Image of the Unit Square

III.11.10.37. The other cases can be handled similarly, or by reducing to these cases by interchanging rows or columns and/or multiplying one or both rows and/or columns by -1, none of which change the absolute value of the determinant.

III.11.10.38. If instead of the unit square, we begin with the rectangle with vertices $(0,0)$, $(r,0)$, $(0,s)$, and (r,s), a virtually identical argument shows that the image parallelogram has area $|det(A)|rs$. The image of any other coordinate rectangle (rectangle with sides parallel to the axes) of length rs is congruent to this parallelogram, hence has the same area by property (iv). Thus T multiplies areas of all coordinate rectangles by the constant factor $|det(A)|$.

III.11.10.39. For the case $n = 3$, the image of the unit cube under the transformation T is the parallelepiped with vertices at the origin, the column vectors of A, and their sums. A more complicated argument along the same lines as the dimension 2 case shows that the volume of this parallelepiped is exactly $|det(A)|$. The argument generalizes to the images of rectangular solids with edges parallel to the axes (*coordinate rectangles*).

III.11.10.40. The case $n \geq 4$ is difficult to visualize and quite complicated to give a geometric argument for. But some special cases can be easily shown for any n:

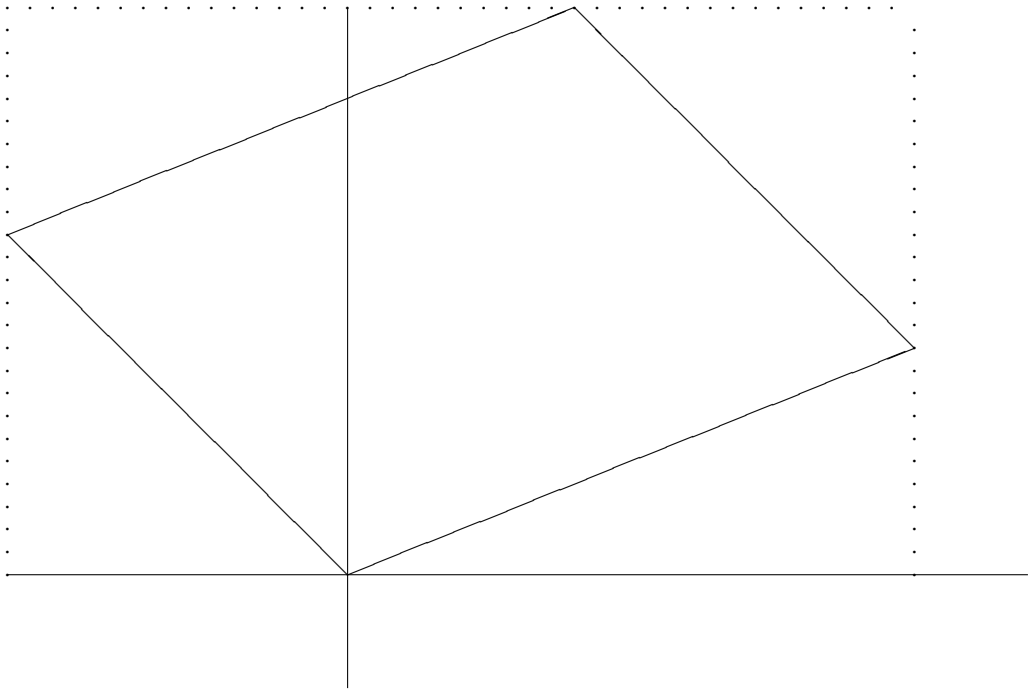


Figure III.2: The Image of the Unit Square

(i) If A is a diagonal matrix, the image of any coordinate rectangle R is another coordinate rectangle, whose n-dimensional volume is $V(R)$ times the absolute value of the product of the diagonal elements of A, which is $|det(A)|$.

(ii) If A is a permutation matrix (), and R is a coordinate rectangle, then $T(R)$ is a coordinate rectangle congruent to R, so $V(T(R)) = V(R)$; and $det(A) = \pm 1$, so $|det(A)| = 1$ and the measure-scaling property holds.

(iii) If A is the elementary matrix with r in the (i, j) place (), then the image of the unit cube is a prism whose base is a rhombus of side 1; thus $V(T(R)) = V(R)$, and $det(A) = 1$ ().

By the multiplicative property of determinants and the fact that any invertible square matrix is a product of elementary matrices, the measure-scaling property holds for coordinate rectangles for general invertible A. For singular A, see ().

III.11.10.41. To handle a general (Jordan) region E, cover E by a fine grid so that the inside and outside estimates for $V(E)$ () from this grid are close together. The image $T(E)$ is then trapped between finite unions of parallelotopes, both of which have volume close to $|det(A)| \cdot V(E)$ by property (iii) of volume. The approximations can be made as closely as desired. By property (ii), $V(T(E)) = |det(A)| \cdot V(E)$. A similar argument () gives the result for Lebesgue measurable E.

III.11.10.42. If A is an $n \times n$ matrix over $\mathbb{C}$, then A can be thought of as giving a linear transformation T from $\mathbb{C}^n \cong \mathbb{R}^{2n}$ to itself, and it can be shown that the factor that T multiplies $2n$-dimensional "volume" by is the square of $|det(A)|$.

The Inverse of a Matrix

There is actually an explicit algebraic formula for the inverse of a (square) matrix, which we develop here.

III.11.10.43. DEFINITION. Let A be an $n \times n$ matrix over $\mathbb{F}$. The *classical adjoint* of A is the matrix

$$adj(A) = [d_{ji}(A)] = [d_{ij}(A)]^t$$

where $d_{ij}(A)$ is the (i, j) cofactor of A (III.11.10.19.).

The classical adjoint has nothing to do with the (ordinary) adjoint of a matrix over $\mathbb{C}$ as defined in (). Note the transpose in the definition of $adj(A)$.

The importance of the classical adjoint is:

III.11.10.44. THEOREM. Let A be an $n \times n$ matrix over $\mathbb{F}$. Then

$$A \cdot adj(A) = adj(A) \cdot A = (det(A))I_n \ .$$

As a corollary, we get an algebraic formula for the inverse of a matrix:

III.11.10.45. COROLLARY. Let A be an $n \times n$ matrix over $\mathbb{F}$. If A is invertible (i.e. $det(A) \neq 0$), then

$$A^{-1} = \frac{1}{det(A)} adj(A) \ .$$

III.11.10.46. The algebraic formula for the inverse via the classical adjoint is not computationally efficient to use; it is pretty cumbersome even for $n = 4$. It is much more computationally efficient to compute the inverse by row reduction (Gaussian elimination) (). But the formula at least gives a very simple way to write down the inverse of a 2×2 matrix:

III.11.10.47. COROLLARY. Let $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ be a 2×2 matrix over $\mathbb{F}$. Then A is invertible if and only if $ad - bc \neq 0$, and in this case

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} \frac{d}{ad-bc} & -\frac{b}{ad-bc} \\ -\frac{c}{ad-bc} & \frac{a}{ad-bc} \end{bmatrix} \ .$$

This formula may be easily checked by direct computation.

The algebraic formula for the inverse does have an important theoretical consequence:

III.11.10.48. Corollary. The entries of the inverse of an invertible (square) matrix A over $\mathbb{F}$ are rational functions of the entries in A.

If $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$, this can be nicely rephrased.

III.11.10.49. Definition. Let $\mathbb{F}$ be a field. Write $GL_n(\mathbb{F})$ for the set of invertible $n \times n$ matrices over $\mathbb{F}$. $GL_n(\mathbb{F})$ is called the *general linear group* of size n over $\mathbb{F}$.

If $A, B \in GL_n(\mathbb{F})$, then AB and A^{-1} are also in $GL_n(\mathbb{F})$; thus $GL_n(\mathbb{F})$ is a group () under matrix multiplication. By III.11.10.24., $GL_n(\mathbb{F}) = \{A \in M_n(\mathbb{F}) : det(A) \neq 0\}$.

III.11.10.50. Theorem. Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$. Then $GL_n(\mathbb{F})$ is an open set in $M_n(\mathbb{F}) \cong \mathbb{F}^{n^2}$, and $A \mapsto A^{-1}$ is an analytic (hence C^∞) function from $GL_n(\mathbb{F})$ to $GL_n(\mathbb{F}) \subseteq M_n(\mathbb{F})$.

Proof: Since $det : M_n(\mathbb{F}) \to \mathbb{F}$ is continuous (), $GL_n(\mathbb{F})$ is the inverse image of the open set $\mathbb{F} \setminus \{0\}$ under a continuous map, hence open. The other statement follows from III.11.10.48. since rational functions are analytic on their domains. 👍

III.11.10.51. Corollary. Let $A = [a_{ij}]$ be an invertible $n \times n$ matrix over $\mathbb{R}$ or $\mathbb{C}$. Then there is an $\epsilon > 0$ (depending on A) such that, whenever $B = [b_{ij}]$ is an $n \times n$ matrix over $\mathbb{R}$ or $\mathbb{C}$ with $|a_{ij} - b_{ij}| < \epsilon$ for all i and j, then B is also invertible.

The formula for the inverse of a matrix has other consequences, for example:

III.11.10.52. Corollary. Let A be an $n \times n$ matrix with integer entries. Then A has an inverse with integer entries (i.e. A is invertible in $M_n(\mathbb{Z})$) if and only if $det(A) = \pm 1$.

Proof: If A has an inverse A^{-1} with integer entries, then $det(A)$ and $det(A^{-1}) = \frac{1}{det(A)}$ are integers, hence $det(A) = \pm 1$. Conversely, if $det(A) = \pm 1$, III.11.10.45. (applied with $\mathbb{F} = \mathbb{R}$) shows that A is invertible and that A^{-1} has integer entries. 👍

This result generalizes to unital commutative rings:

III.11.10.53. Theorem. Let R be a commutative unital ring. If $A \in M_n(R)$, then the following are equivalent:

(i) A is invertible in $M_n(R)$.

(ii) A is left invertible in $M_n(R)$.

(iii) $det(A)$ is a unit (invertible element) of R.

III.11.10.54. Corollary. Let R be a commutative unital ring. If $A \in M_n(R)$, then A is left invertible if and only if A is right invertible.

This can fail if R is noncommutative, even for $n = 1$ ().

Here is another consequence. If R is a unital subring of a larger unital ring R', an element of R can in general be invertible in the larger ring R' without being invertible in R (e.g. $R = \mathbb{Z}$, $R' = \mathbb{R}$). But if $\mathbb{F}$ is a field contained in a larger field $\mathbb{K}$, and we regard $M_n(\mathbb{F})$ as a subring of $M_n(\mathbb{K})$, this cannot happen:

III.11.10.55. Corollary. Let $\mathbb{F} \subseteq \mathbb{K}$ be fields. If $A \in M_n(\mathbb{F})$, then A is invertible in $M_n(\mathbb{K})$ if and only if A is invertible in $M_n(\mathbb{F})$, i.e. if A is invertible in $M_n(\mathbb{K})$, then the entries of A^{-1} are in $\mathbb{F}$.

This applies for example if $\mathbb{F} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$, or if $\mathbb{F} = \mathbb{Q}$ and $\mathbb{K} = \mathbb{R}$. Similarly, if $A \in M_n(\mathbb{Z}) \subseteq M_n(\mathbb{C})$, and A is invertible in $M_n(\mathbb{C})$, then $A^{-1} \in M_n(\mathbb{Q})$; and if $det(A) = \pm 1$, then $A^{-1} \in M_n(\mathbb{Z})$ (and conversely).

Cramer's Rule

III.11.11. Eigenvalues and Eigenvectors

One of the most important concepts in linear algebra is the theory of eigenvalues and eigenvectors.

III.11.11.1. Definition. Let V be a vector space over $\mathbb{F}$, and T a linear operator on V. A nonzero vector $x \in V$ is an *eigenvector of T with eigenvalue $\lambda \in \mathbb{F}$* if $Tx = \lambda x$.

An element $\lambda \in \mathbb{F}$ is an *eigenvalue* of T if there exists an eigenvector of T with eigenvalue λ.

Knowing the eigenvalues and eigenvectors of an operator T are often the key to understanding T. For example, if V is finite-dimensional and V has a basis consisting of eigenvectors of T, then the matrix of T with respect to this basis is a diagonal matrix and it is easy to understand T (this often happens, but not always). If V is finite-dimensional, it turns out there are only finitely many eigenvalues for any T, which can be found by an algebraic calculation (III.11.11.6.). If V is infinite-dimensional, the situation is far more complicated, sometimes impossibly complicated.

III.11.11.2. The term *eigenvalue* was apparently coined and first used by A. S. Eddington in 1927. Many mathematicians have regarded this hybrid German-English word as a language abomination, and used alternatives; some of the alternatives commonly used are *characteristic value*, *characteristic root*, and *latent root*. Use of alternatives has become increasingly rare in recent years, and *eigenvalue* is now firmly established and quite universal. Similarly, terms like *characteristic vector* have been used in place of *eigenvector*, but increasingly rarely.

III.11.11.3. Examples. (i) If T is the identity operator I on V, every nonzero vector in V is an eigenvector with eigenvalue 1.

(ii) Any nonzero vector in the null space of T is an eigenvector with eigenvalue 0; conversely, an eigenvector of T with eigenvalue 0 is in the null space of T. Thus 0 is an eigenvalue of T if and only if $\mathcal{N}(T) \neq \{0\}$, i.e. if and only if T is not one-to-one.

(iii) An operator need not have any eigenvalues or eigenvectors. Let T be the operator on $\mathbb{R}^2$ with matrix $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Geometrically, T is a 90-degree clockwise rotation. No nonzero vector is sent to a scalar multiple of itself.

(iv) Let s be the vector space of all sequences of complex numbers, with coordinatewise addition and scalar multiplication. Let T be the *shift* on s:

$$T((z_1, z_2, \dots)) = (0, z_1, z_2, \dots) .$$

If $(z_1, z_2, \dots)$ is an eigenvector for T with eigenvalue λ, then $z_{k+1} = \lambda z_k$ for all k and $z_1 = 0$, so $z_k = 0$ for all k by induction, a contradiction. Thus T has no eigenvalues.

(v) Let s be as in (iii), and S the *backward shift*:

$$S((z_1, z_2, \dots)) = (z_2, z_3, \dots) .$$

Then $(1, 0, 0, \dots)$ is an eigenvector for S with eigenvalue 0, and if $\lambda \neq 0$, then $(\lambda, \lambda^2, \lambda^3, \dots)$ is an eigenvector for S with eigenvalue λ. So every $\lambda \in \mathbb{C}$ is an eigenvalue for S.

III.11.11.4. An eigenvector must be nonzero by definition; thus the corresponding eigenvalue is unique. Any nonzero linear combination of eigenvectors for T for the *same* eigenvalue λ is also an eigenvector for T with eigenvalue λ. Thus, if λ is an eigenvalue of T, the set of eigenvectors of T with eigenvalue λ is the set of nonzero vectors in a subspace of V, called the *eigenspace* of T with eigenvalue λ. For example, if the null space of T is nontrivial, it is the eigenspace for T with eigenvalue 0.

III.11.11.5. The observation in III.11.11.3.(ii) is the key to finding eigenvalues and eigenvectors. Note that $Tx = \lambda x$ if and only if $(T - \lambda I)x = 0$. Thus the eigenvectors of T with eigenvalue λ are exactly the nonzero vectors (if any) in $\mathcal{N}(T - \lambda I)$, so the eigenvalues of T are exactly the scalars λ such that $T - \lambda I$ is not one-to-one.

III.11.11.6. If V is finite-dimensional, the eigenvalues can be found by an algebraic procedure: λ is an eigenvalue of T if and only if $det(T - \lambda I) = 0$. (Recall that the determinant of a linear operator is well defined ().) In fact, the function $p(x) = det(T - xI)$ is a polynomial of degree $dim(V)$ with coefficients in $\mathbb{F}$, called the *characteristic polynomial* of T, and the roots of this polynomial are exactly the eigenvalues of T. The eigenvectors can then be described as the (nonzero) null spaces of the corresponding operators, which can be found by solving a system of (homogeneous) linear equations. In particular, if V has dimension n, then there are at most n eigenvalues (there can easily be fewer, e.g. in III.11.11.3.(i): 1 is the only eigenvalue of I).

III.11.11.7. EXAMPLE. Let T be the operator on $\mathbb{R}^2$ from III.11.11.3.(iii). Then its characteristic polynomial is $p(x) = x^2 + 1$. Thus we see again algebraically that T has no eigenvalues (as an operator on $\mathbb{R}^2$).

However, if S is the operator on $\mathbb{C}^2$ with the same matrix, then S does have eigenvalues $\pm i$. The vector $(i, 1)$ is an eigenvector with eigenvalue i, and $(1, i)$ is an eigenvector with eigenvalue $-i$.

III.11.11.8. Thus, if $\mathbb{F}$ is algebraically closed (e.g. $\mathbb{F} = \mathbb{C}$), and V is finite-dimensional, then every operator on V has at least one eigenvalue. (This can be false if V is infinite-dimensional, cf. III.11.11.3.(iv).) In fact, some parts of linear algebra are technically simpler when working over an algebraically closed field. This is one reason we usually work over $\mathbb{C}$ in operator theory in functional analysis (but not the only one!)

III.11.11.9. We often talk about the eigenvalues and eigenvectors of a (square) matrix, really meaning the eigenvalues and eigenvectors of the corresponding operator. But when doing this, the base field must be specified and kept in mind, as Example III.11.11.7. shows!

III.11.11.10. Example. Let T be the operator on either $\mathbb{R}^2$ or $\mathbb{C}^2$ with matrix $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$. Then the characteristic polynomial is $p(x) = x^2$, so the only eigenvalue is 0. But the eigenspace (null space) is only one-dimensional. So even if $\mathbb{F}$ is algebraically closed and V is finite-dimensional, there may not be enough eigenvectors of T to span V; in particular, this operator is not diagonalizable (this matrix is not similar to a diagonal matrix).

III.11.11.11. If x is an eigenvector for T with eigenvalue λ, then x is also an eigenvector for T^2 with eigenvalue λ^2; thus if λ is an eigenvalue for T, then λ^2 is an eigenvalue for T^2. (T^2 may have more eigenvalues, cf. III.11.11.3.(iii).) Similarly, if x is an eigenvector for T with eigenvalue λ, then x is also an eigenvector for T^n with eigenvalue λ^n. More generally, if $p(x)$ is a polynomial with coefficients in $\mathbb{F}$, we may form $p(T)$; if x is an eigenvector for T with eigenvalue λ, then x is also an eigenvector for $p(T)$ with eigenvalue $p(\lambda)$. In particular, if T is *nilpotent*, i.e. $T^n = 0$ for some n, then the only eigenvalue of T is 0 (0 is an eigenvalue since T cannot be one-to-one).

III.11.11.12. When I was in college, I had a summer job doing data analysis for a company run by two psychologists. One of them wanted me to do a factor analysis of some arrays of data. I didn't know what a factor analysis was, so he gave me a book describing the process, written for psychologists. As I gradually deciphered the mathematics in the book, I realized it was nothing but finding the eigenvectors of a (symmetric) matrix corresponding to the eigenvalues in decreasing order. When I reported this to my employer, his response was: "What's an eigenvector?"

III.11.12. Exercises

III.11.12.1. Let A be an $n \times n$ matrix over $\mathbb{F}$. Show directly from the definition that if A has two identical rows, then the terms in the expansion of $det(A)$ cancel in pairs, and hence $det(A) = 0$. [If τ is the transposition interchanging the two rows, for any σ consider the terms for σ and $\sigma\tau$.]

III.11.12.2. Say that a $2m \times 2n$ matrix A over $\mathbb{R}$ is *complex* if, when A is partitioned into 2×2 blocks, each subblock is of the form

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$$

where the a and b are real numbers which vary from block to block (cf. III.6.3.1.).

(a) Show that a sum or product of complex matrices, if defined, is complex. [Use III.11.1.22..]

(b) If a $2n \times 2n$ matrix over $\mathbb{R}$ is complex and invertible (in $M_{2n}(\mathbb{R})$), show that its inverse is also complex.

(c) If A is the complex matrix

$$\begin{bmatrix} a_{11} & -b_{11} & a_{12} & -b_{12} & \cdots & a_{1n} & -b_{1n} \\ b_{11} & a_{11} & b_{12} & a_{12} & \cdots & b_{1n} & a_{1n} \\ a_{21} & -b_{21} & a_{22} & -b_{22} & \cdots & a_{2n} & -b_{2n} \\ b_{21} & a_{21} & b_{22} & a_{22} & \cdots & b_{2n} & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & -b_{m1} & a_{m2} & -b_{m2} & \cdots & a_{mn} & -b_{mn} \\ b_{m1} & a_{m1} & b_{m2} & a_{m2} & \cdots & b_{mn} & a_{mn} \end{bmatrix}$$

identify A with the $m \times n$ matrix

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{11} & c_{12} & \cdots & c_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{bmatrix}$$

with complex entries, where $c_{jk} = a_{jk} + ib_{jk}$ for $1 \leq j \leq m$, $1 \leq k \leq n$. Show that this correspondence preserves addition, multiplication, and inversion.

(d) Interpret this correspondence as follows. Let $\mathcal{S} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ be the standard (ordered) basis for $\mathbb{C}^n$ as a vector space over $\mathbb{C}$. Then the set $\mathcal{B} = \{\boldsymbol{e}_1, i\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n, i\boldsymbol{e}_n\}$ is an ordered basis for $\mathbb{C}^n$ as a real vector space. If T is a real-linear transformation from $\mathbb{C}^n$ to $\mathbb{C}^m$, then the matrix for T with respect to the $\mathcal{B}$-bases is complex if and only if T is complex-linear, in which case its matrix (over $\mathbb{C}$) with respect to the standard bases is the corresponding matrix in $M_{mn}(\mathbb{C})$.

(e) More generally, show that if R is a ring, and R' is a subring of $M_k(R)$, then the $m \times n$ matrices over R' can be identified with the set of $km \times kn$ matrices over R whose $k \times k$ block entries are in R', and the identification preserves matrix addition, multiplication, and (if R' is closed under inverses in $M_k(R)$) inversion. In particular, if $\mathbb{F}$ is a field and $\mathbb{F}'$ is a finite extension field of degree k, then $\mathbb{F}'$ can be embedded as a (unital) subring of $M_k(\mathbb{F})$, so this construction applies.

(f) Interpret this problem as a special case of the tensor product construction ().

III.11.12.3. Let $\mathbb{F}$ and $\mathbb{K}$ be fields, with $\mathbb{F} \subseteq \mathbb{K}$. Let A be an $m \times n$ matrix over $\mathbb{F}$, and $\boldsymbol{b} \in \mathbb{F}^m$. Then the system $A\boldsymbol{x} = \boldsymbol{b}$ may be thought of as a system of linear equations over $\mathbb{F}$ or over $\mathbb{K}$.

(a) Show that the system $A\boldsymbol{x} = \boldsymbol{b}$ has a solution in $\mathbb{F}^n$ if and only if it has a solution in $\mathbb{K}^n$. [Note that the system is consistent over $\mathbb{F}$ if and only if it is consistent over $\mathbb{K}$, and apply III.11.7.10..] The dimension (over $\mathbb{F}$) of the affine subspace of solutions in $\mathbb{F}^n$ is the same as the dimension (over $\mathbb{K}$) of the affine subspace of the solutions in $\mathbb{K}^n$; in fact the space of solutions over $\mathbb{K}$ is the affine (over $\mathbb{K}$) subspace generated ("spanned") by the space of solutions over $\mathbb{F}$.

(b) Show that the result of (a) generalizes III.11.10.55., i.e. give an alternate proof of III.11.10.55. using (a).

(c) Show that (a) is equivalent to the statement that $\mathbb{K}$, regarded as a module over $\mathbb{F}$, is flat ().

III.11.13. Multilinear Algebra

Throughout this section, $\mathbb{F}$ will denote a fixed field, and "vector space" will mean "vector space over $\mathbb{F}$." There is an important generalization of linear maps:

III.11.13.1. DEFINITION. (i) Let $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{Z}$ be vector spaces (over $\mathbb{F}$). A function $F : \mathcal{V} \times \mathcal{W} \to \mathcal{Z}$ is *bilinear* if, for every $v, v_1, v_2 \in \mathcal{V}$, $w, w_1, w_2 \in \mathcal{W}$, $\alpha \in \mathbb{F}$, we have

$$F(v_1 + v_2, w) = F(v_1, w) + F(v_2, w)$$

$$F(v, w_1 + w_2) = F(v, w_1) + F(v, w_2)$$

$$F(\alpha v, w) = F(v, \alpha w) = \alpha F(v, w) \ .$$

(ii) Let $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_m, \mathcal{Z}$ be vector spaces. A function $F : \mathcal{V}_1 \times \cdots \times \mathcal{V}_m \to \mathcal{Z}$ is *multilinear* if, for every $x_k, y_k \in \mathcal{V}_k$ $(1 \leq k \leq m)$ and $\alpha \in \mathbb{F}$, we have

$$F(x_1, \ldots, x_{k-1}, x_k + y_k, x_{k+1}, \ldots, x_m) = F(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_m) + F(x_1, \ldots, x_{k-1}, y_k, x_{k+1}, \ldots, x_m)$$

$$F(\alpha x_1, x_2, \ldots, x_m) = F(x_1, \alpha x_2, \ldots, x_m) = \cdots = F(x_1, x_2, \ldots, \alpha x_m) = \alpha F(x_1, x_2, \ldots, x_m) \ .$$

In other words, a function is bilinear (multilinear) if it is linear in each coordinate separately when the other coordinates are fixed. We should technically say "$\mathbb{F}$-bilinear" and "$\mathbb{F}$-multilinear." (Of course, "bilinear" is just "multilinear" with $m = 2$, but it is convenient and standard to have a distinct name in this case. The term "m-linear" is also sometimes used when there are m factors.)

III.11.14. Tensor Products

In this section, we again fix a field $\mathbb{F}$, and we consider only vector spaces over $\mathbb{F}$. The theory of tensor products of vector spaces is very much simpler and better behaved than tensor products of more general modules over rings. We will concentrate on tensor products of finite-dimensional vector spaces, where things are even simpler, since this is the case of primary interest in real analysis.

Concrete Construction

We first give an explicit concrete construction of the tensor product, and then show it is uniquely determined up to natural isomorphism by a universal property.

III.11.14.1. Let $V_1, \ldots, V_m$ be vector spaces over $\mathbb{F}$. We first form the free vector space X over the set $V_1 \times \cdots \times V_m$, i.e. the collection of all formal sums

$$\sum \alpha_{(v_1,\ldots,v_m)}(v_1, \ldots, v_m)$$

where the sum is over all m-tuples in $V_1 \times \cdots \times V_m$, the $\alpha_{(v_1,\ldots,v_m)}$ are in $\mathbb{F}$, and only finitely many of the $\alpha_{(v_1,\ldots,v_m)}$ are nonzero in each sum. Addition and scalar multiplication are defined in the obvious way:

$$\Big[\sum \alpha_{(v_1,\ldots,v_m)}(v_1, \ldots, v_m)\Big] + \Big[\sum \beta_{(v_1,\ldots,v_m)}(v_1, \ldots, v_m)\Big] = \sum [\alpha_{(v_1,\ldots,v_m)} + \beta_{(v_1,\ldots,v_m)}](v_1, \ldots, v_m)$$

$$\alpha \sum \beta_{(v_1,\ldots,v_m)}(v_1,\ldots,v_m) = \sum \alpha\beta_{(v_1,\ldots,v_m)}(v_1,\ldots,v_m) \ .$$

Note that X is a huge infinite-dimensional vector space over $\mathbb{F}$ in general: there is one dimension for each m-tuple in $V_1 \times \cdots \times V_m$, i.e. the sums where exactly one $\alpha_{(v_1,\ldots,v_m)}$ is nonzero form a basis for X (we just write $(v_1,\ldots,v_m)$ for the sum with one term $1(v_1,\ldots,v_m)$).

III.11.14.2. Now let N be the subspace of X spanned by all sums of the form

$$(v_1,\ldots,v_{k-1},v_k+w_k,v_{k+1},\ldots,v_m)-(v_1,\ldots,v_{k-1},v_k,\ldots,v_m)-(v_1,\ldots,v_{k-1},w_k,v_{k+1},\ldots,v_m)$$

$$(v_1,\ldots,v_{k-1},\alpha v_k,v_{k+1},\ldots,v_m)-\alpha(v_1,\ldots,v_{k-1},v_k,v_{k+1},\ldots,v_m)$$

for $1 \le k \le m$, $v_k, w_k \in V_k$, $\alpha \in \mathbb{F}$. Define

$$V_1 \otimes_{\mathbb{F}} V_2 \otimes_{\mathbb{F}} \cdots \otimes_{\mathbb{F}} V_m = X/N \ .$$

$V_1 \otimes_{\mathbb{F}} V_2 \otimes_{\mathbb{F}} \cdots \otimes_{\mathbb{F}} V_m$, usually just written $V_1 \otimes V_2 \otimes \cdots \otimes V_m$, is called the *tensor product of* $V_1,\ldots,V_m$ (over $\mathbb{F}$). It is of course also a vector space over $\mathbb{F}$.

III.11.14.3. If $(v_1,\ldots,v_m)$ is one of the basis elements of X, write $v_1 \otimes \cdots \otimes v_m$ for its image in $V_1 \otimes \cdots \otimes V_m$. Such an element of $V_1 \otimes \cdots \otimes V_m$ is called an *elementary tensor*; every element of $V_1 \otimes \cdots \otimes V_m$ is a linear combination of elementary tensors. But not all elementary tensors are linearly independent, i.e. the expression of an element of $V_1 \otimes \cdots \otimes V_m$ as a linear combination of elementary tensors is far from unique in general. Distinct elementary tensors can even be equal. For the definition of the quotient can be rewritten to say that

$$v_1 \otimes \cdots \otimes v_{k-1} \otimes (v_k+w_k) \otimes v_{k+1} \otimes \cdots \otimes v_m = v_1 \otimes \cdots \otimes v_{k-1} \otimes v_k \otimes v_{k+1} \otimes \cdots \otimes v_m + v_1 \otimes \cdots \otimes v_{k-1} \otimes w_k \otimes v_{k+1} \otimes \cdots \otimes v_m$$

$$v_1 \otimes \cdots \otimes v_{k-1} \otimes (\alpha v_k) \otimes v_{k+1} \otimes \cdots \otimes v_m = \alpha[v_1 \otimes \cdots \otimes v_{k-1} \otimes v_k \otimes v_{k+1} \otimes \cdots \otimes v_m]$$

for $1 \le k \le m$, $v_k, w_k \in V_k$, $\alpha \in \mathbb{F}$. (Because of the last relation any element of $V_1 \otimes \cdots \otimes V_m$ is actually a finite *sum* of elementary tensors, although in a highly nonunique way in general.) An element of the tensor product which can be written as an elementary tensor is called a *decomposable element*. In general, "most" elements of a tensor product are not decomposable.

Because of this nonuniqueness, one must be careful about defining linear maps out of a tensor product by specifying their value on elementary tensors; it must be shown that such a map is well defined (i.e. agrees on equal sums of elementary tensors). This is normally done using the universal property of III.11.14.5..

III.11.14.4. Although X and N are generally very large infinite-dimensional vector spaces, the tensor product is not nearly as large: if the V_k are finite-dimensional, the tensor product is also finite-dimensional (III.11.14.14.).

An Abstract Characterization

III.11.14.5. DEFINITION. Let $V_1,\ldots,V_m$ be vector spaces over $\mathbb{F}$. A *tensor product* of $\{V_1,\ldots,V_m\}$ over $\mathbb{F}$ is a vector space Y over $\mathbb{F}$ and an m-linear map $M : V_1 \times \cdots \times V_m \to Y$ with the universal property that if W is any vector space over $\mathbb{F}$ and $T : V_1 \times \cdots \times V_m \to W$ is an m-linear map, then there is a unique linear map $S : Y \to W$ with $T = S \circ M$.

III.11.14.6. By the usual argument (), any two tensor products of $\{V_1, \dots, V_m\}$ over $\mathbb{F}$ are naturally isomorphic.

III.11.14.7. PROPOSITION. Let $V_1, \dots, V_m$ be vector spaces over $\mathbb{F}$. Then $V_1 \otimes \cdots \otimes V_m$ is a tensor product of $\{V_1, \dots, V_m\}$ over $\mathbb{F}$ in the sense of III.11.14.5..

PROOF: The map $M : V_1 \times \cdots \times V_m \to V_1 \otimes \cdots \otimes V_m$ is the obvious one:

$$M(v_1, \dots, v_m) = v_1 \otimes \cdots \otimes v_m$$

and the relations of III.11.14.3. show that this map is m-linear. If W is a vector space over $\mathbb{F}$ and $T : V_1 \times \cdots \times V_m \to W$ is an m-linear map, then T gives a function from the standard basis of X to W, which extends uniquely to a linear map $R : X \to W$. The m-linearity of T implies that N is contained in the null space of R, so R drops to a linear map $S : X/N = V_1 \otimes \cdots \otimes V_m \to W$, and obviously $T = S \circ M$. S is uniquely determined on elementary tensors, hence on their span, which is all of $V_1 \otimes \cdots \otimes V_m$. 👍

III.11.14.8. Thus, if $V_1, \dots, V_m, W$ are vector spaces over $\mathbb{F}$, there is a natural one-one correspondence between the set $\mathcal{L}^m(V_1, \dots, V_m; W)$ of m-linear maps from $V_1 \times \cdots \times V_m$ to W and the set $\mathcal{L}(V_1 \otimes \cdots \otimes V_m, W)$. Both sets are naturally vector spaces over $\mathbb{F}$, and the correspondence is an isomorphism. In particular, there is a natural isomorphism between $\mathcal{L}^m(V_1, \dots, V_m) := \mathcal{L}^m(V_1, \dots, V_m; \mathbb{F})$ and $\mathcal{L}(V_1 \otimes \cdots \otimes V_m, \mathbb{F}) = (V_1 \otimes \cdots \otimes V_m)^*$.

Associativity and Commutativity

III.11.14.9. PROPOSITION. Let V_1, V_2, V_3 be vector spaces over $\mathbb{F}$. Then the identifications

$$(v_1 \otimes v_2) \otimes v_3 \leftrightarrow v_1 \otimes (v_2 \otimes v_3) \leftrightarrow v_1 \otimes v_2 \otimes v_3$$

($v_k \in V_k$) extend to isomorphisms

$$(V_1 \otimes V_2) \otimes V_3 \cong V_1 \otimes (V_2 \otimes V_3) \cong V_1 \otimes V_2 \otimes V_3 \ .$$

PROOF: This may seem trivial, but it is not (quite): it must be shown that the correspondences give well-defined maps (cf. the caution in III.11.14.3.), which can be done in a straightforward way using the universal property. For example, the bilinear map $(v_1, v_2) \mapsto v_1 \otimes v_2$ gives a map $(v_1, v_2, v_3) \mapsto (v_1 \otimes v_2, v_3)$ from $V_1 \times V_2 \times V_3$ to $(V_1 \otimes V_2) \times V_3$, and the composition of this map with the bilinear map from $(V_1 \otimes V_2) \times V_3$ to $(V_1 \otimes V_2) \otimes V_3$ defined by $(v_1 \otimes v_2, v_3) \mapsto (v_1 \otimes v_2) \otimes v_3$ is easily checked to be 3-linear, hence gives a well-defined map $v_1 \otimes v_2 \otimes v_3 \mapsto (v_1 \otimes v_2) \otimes v_3$. The other maps are similar. 👍

III.11.14.10. PROPOSITION. Let V_1 and V_2 be vector spaces over $\mathbb{F}$. Then the correspondence

$$v_1 \otimes v_2 \leftrightarrow v_2 \otimes v_1$$

($v_k \in V_k$) extends to an isomorphism

$$V_1 \otimes V_2 \cong V_2 \otimes V_1 \ .$$

PROOF: We must again show the maps are well defined. The map $(v_1, v_2) \mapsto v_2 \otimes v_1$ from $V_1 \times V_2$ to $V_2 \otimes V_1$ is bilinear, hence gives a linear map $v_1 \otimes v_2 \mapsto v_2 \otimes v_1$ from $V_1 \otimes V_2$ to $V_2 \otimes V_1$. Symmetrically, we get a map $v_2 \otimes v_1 \mapsto v_1 \otimes v_2$ from $V_2 \otimes V_1$ to $V_1 \otimes V_2$. The two maps are inverses, hence isomorphisms. 👍

III.11.14.11. As a result we may somewhat cavalierly (and by slight abuse of notation) insert or remove parentheses and/or reorder the factors in any multiple tensor product, and we will do this without comment. (We technically also need III.11.14.25. to extend III.11.14.9. and III.11.14.10. to multiple factors.)

Bases and Dimension of Tensor Products

III.11.14.12. Let $V_1, \ldots, V_m$ be vector spaces over $\mathbb{F}$. For $1 \leq k \leq m$ let $\{x_j^{(k)}\}$ be a basis for V_k. Elementary tensors $v_1 \otimes \cdots \otimes v_m$ where each v_k is one of the basis vectors $x_j^{(k)}$ are called *basic elementary tensors* (with respect to the chosen bases). An argument similar to the proof of III.11.14.7. shows:

III.11.14.13. PROPOSITION. The set of basic elementary tensors forms a basis for $V_1 \otimes \cdots \otimes V_m$.

PROOF: We write the proof only for $m = 2$; the case of general m is essentially identical but the notation becomes oppressive (the reader should mentally go through the general argument). Thus we change notation and let V and W be vector spaces over $\mathbb{F}$ with bases $\{x_j : j \in J\}$ and $\{y_k : k \in K\}$ respectively; we must show that $B = \{x_j \otimes y_k : j \in J, k \in K\}$ is a basis for $V \otimes W$. If $v \in V$, $w \in W$, then $v = \sum_{j\in J} \alpha_j x_j$ and $w = \sum_{k\in K} \beta_k y_k$ with only finitely many α_j and β_k nonzero. Then

$$v \otimes w = \left[\sum_{j\in J} \alpha_j x_j\right] \otimes \left[\sum_{k\in K} \beta_k y_k\right] = \sum_{j\in J}\sum_{k\in K} \alpha_j\beta_k[x_j \otimes y_k]$$

so $v \otimes w$ is in the span of B. Since every element of $V \otimes W$ is a sum of elementary tensors, B spans $V \otimes W$. To show that B is linearly independent, let Y be the free vector space with basis B, i.e. the set of finite formal linear combinations of the $x_j \otimes y_k$. There is an obvious surjective linear map $R : Y \to V \otimes W$. On the other hand, there is a bilinear map $T : V \times W \to Y$ defined by

$$T\left(\sum_{j\in J} \alpha_j x_j, \sum_{k\in K} \beta_k y_k\right) = \sum_{j\in J}\sum_{k\in K} \alpha_j\beta_k[x_j \otimes y_k]$$

(this is well defined since the x_j and the y_k are linearly independent) which by the universal property is the form $S \circ M$ for $S : V \otimes W \to Y$. It is clear that $S \circ R$ is the identity on Y, hence R is injective, i.e. an isomorphism, which preserves linear independence. 👍

III.11.14.14. Corollary. Let $V_1, \ldots, V_m$ be finite-dimensional vector spaces over $\mathbb{F}$, with $dim(V_k) = d_k$. then $V_1 \otimes \cdots \otimes V_m$ is also finite-dimensional, and

$$dim(V_1 \otimes \cdots \otimes V_m) = d_1 d_2 \cdots d_m \ .$$

Linear Transformations and Tensor Products

III.11.14.15. Proposition. Let V_1, V_2, W be vector spaces over $\mathbb{F}$. Then there are natural isomorphisms

$$\mathcal{L}(V_1 \otimes V_2, W) \cong \mathcal{L}(V_1, \mathcal{L}(V_2, W)) \cong \mathcal{L}(V_2, \mathcal{L}(V_1, W)) \ .$$

All can be naturally identified with the set $\mathcal{L}^2(V_1, V_2; W)$ of bilinear maps from $V_1 \times V_2$ to W.

Proof: The identification of $\mathcal{L}(V_1 \otimes V_2, W)$ with the set of bilinear maps from $V_1 \times V_2$ to W is via the universal property. For the other isomorphisms, let $T \in \mathcal{L}(V_1 \otimes V_2, W)$. For $v_1 \in V_1$, define $T_{v_1} : V_2 \to W$ by $T_{v_1}(v_2) = T(v_1 \otimes v_2)$. It is easily checked that $T_{v_1} \in \mathcal{L}(V_2, W)$, and that $\hat{T} : V_1 \to \mathcal{L}(V_2, W)$ defined by $\hat{T}(v_1) = T_{v_1}$ is linear. In the other direction, if $S \in \mathcal{L}(V_1, \mathcal{L}(V_2, W))$, let $R : V_1 \times V_2 \to W$ be defined by $R(v_1, v_2) = [S(v_1)](v_2)$. It is easily checked that R is bilinear, so defines a linear map $\check{S} : V_1 \otimes V_2 \to W$. The maps $T \mapsto \hat{T}$ and $S \mapsto \check{S}$ are inverses, hence isomorphisms (it is easily checked that these maps are linear). The isomorphism with $\mathcal{L}(V_2, \mathcal{L}(V_1, W))$ follows from $V_1 \otimes V_2 \cong V_2 \otimes V_1$. 👍

III.11.14.16. There are more complicated similar correspondences for multiple tensor products; statements are left to the reader. "Natural" means, among other things, that the correspondences respect composition in an appropriate sense.

III.11.14.17. Now let $V_1, \ldots, V_m, W_1, \ldots, W_m$ be vector spaces over $\mathbb{F}$, and $T_k \in \mathcal{L}(V_k, W_k)$ for each k. Define a map $S : V_1 \times \cdots \times V_m \to W_1 \otimes \cdots \otimes W_m$ by

$$S(v_1, \ldots, v_m) = T_1 v_1 \otimes \cdots \otimes T_m v_m \ .$$

It is easily checked that S is m-linear, hence defines a linear map $T : V_1 \otimes \cdots \otimes V_m \to W_1 \otimes \cdots \otimes W_m$. Write (at least temporarily) $T = T_1 \odot \cdots \odot T_m$, where we use $\odot$ instead of $\otimes$ to avoid ambiguity.

III.11.14.18. The map $(T_1, \ldots, T_m) \mapsto T_1 \odot \cdots \odot T_m$ from $\mathcal{L}(V_1, W_1) \times \cdots \times \mathcal{L}(V_m, W_m)$ to $\mathcal{L}(V_1 \otimes \cdots \otimes V_m, W_1 \otimes \cdots \otimes W_m)$ is m-linear, hence defines a linear map

$$\Phi : \mathcal{L}(V_1, W_1) \otimes \cdots \otimes \mathcal{L}(V_m, W_m) \to \mathcal{L}(V_1 \otimes \cdots \otimes V_m, W_1 \otimes \cdots \otimes W_m)$$

$$\Phi(T_1 \otimes \cdots \otimes T_m) = T_1 \odot \cdots \odot T_m$$

(this is the reason for using the symbol $\odot$).

III.11.14.19. Question. Is Φ an isomorphism?

III.11.14.20. We will show that the answer is yes if the vector spaces are finite-dimensional, but (at least under AC) Φ will not be surjective if the vector spaces are infinite-dimensional. Injectivity of Φ should hold in general, at least under AC. In the case where Φ is injective, we may dispense with the symbol $\odot$ and just use $\otimes$ unambiguously.

III.11.14.21. THEOREM. (AC) The map Φ is an isomorphism for V_k finite-dimensional and W_k arbitrary.

PROOF: We write the proof for $m = 2$ to ease notation; the general case is essentially identical (or can be obtained by iteration). Fix bases $\{x_{i1} : i \in I\}$, $\{x_{j2} : j \in J\}$, $\{y_{k1} : k \in K\}$, $\{y_{\ell 2} : \ell \in L\}$ for V_1, V_2, W_1, and W_2 respectively, for index sets I, J, K, L with I and J finite. A linear transformation T from V_1 to W_1 is completely determined by the finite set of vectors $\{Tx_{i1} : i \in I\}$, and these finitely many vectors are all linear combinations of $\{y_{k1} : k \in K_T\}$ for some finite subset K_T of K (depending on T), and similarly for any element of $\mathcal{L}(V_2, W_2\}$. For each $i1$ and $k1$ there is a rank-one linear transformation E_{ik1} from V_1 to W_1 with $E_{ik1}x_{i1} = y_{k1}$ and $E_{ik1}x_{i'1} = 0$ for $i' \neq i$, and any $T_1 \in \mathcal{L}(V_1, W_1)$ is a linear combination of the E_{ik1} (this is where finite-dimensionality of V_1 is used). It is clear that the E_{ik1} are linearly independent in $\mathcal{L}(V_1, W_1)$, so they form a basis. Similarly, there are rank-one elements $E_{j\ell 2}$ of $\mathcal{L}(V_2, W_2)$ forming a basis. So, by III.11.14.13., $\{E_{ik1} \otimes E_{j\ell 2}\}$ is a basis for $\mathcal{L}(V_1, W_1) \otimes \mathcal{L}(V_2, W_2\}$.

Set $v_{ij} = x_{i1} \otimes x_{j2}$ and $w_{k\ell} = y_{k1} \otimes y_{\ell 2}$ for each i, j, k, ℓ. Then $\{v_{ij}\}$ is a basis for $V_1 \otimes V_2$ and $\{w_{k\ell}\}$ is a basis for $W_1 \otimes W_2$, also by III.11.14.13.. By the same argument as above, the rank-one operators $E_{ijk\ell}$ sending v_{ij} to w_{kl} form a basis for $\mathcal{L}(V_1 \otimes V_2, W_1 \otimes W_2)$ since $V_1 \otimes V_2$ is finite-dimensional. But $E_{ijk\ell} = E_{ik1} \odot E_{j\ell 2} = \Phi(E_{ik1} \otimes E_{j\ell 2})$. Thus Φ sends a basis for $\mathcal{L}(V_1, W_1) \otimes \mathcal{L}(V_2, W_2)$ to a basis for $\mathcal{L}(V_1 \otimes V_2, W_1 \otimes W_2)$ and is therefore an isomorphism. 👍

III.11.14.22. The proof uses the AC because bases must be chosen for the W_k. But the AC is not needed if the W_k are also finite-dimensional (or, more generally, known to have bases).

III.11.14.23. Here is an important special case. If V is a vector space, denote by V^* the set of linear functionals on V, i.e. $V^* = \mathcal{L}(V, \mathbb{F})$. If V is finite-dimensional, then V^* is also finite-dimensional and $dim(V^*) = dim(V)$; and $(V^*)^*$ is naturally isomorphic to V (this fails if V is infinite-dimensional, at least under AC). Then, as a special case of III.11.14.21. we obtain: if $V_1, \ldots, V_m$ are finite-dimensional vector spaces over $\mathbb{F}$, then Φ gives an isomorphism

$$V_1^* \otimes \cdots \otimes V_m^* \cong (V_1 \otimes \cdots \otimes V_m)^* \ .$$

III.11.14.24. Tensor product of operators preserves multiplication: if $V_1, \ldots, V_m, W_1, \ldots, W_m, Z_1, \ldots, Z_m$ are vector spaces over $\mathbb{F}$, $T_k \in \mathcal{L}(V_k, W_k)$, $S_k \in \mathcal{L}(W_k, Z_k)$ for each k, then

$$(S_1 \odot \cdots \odot S_m)(T_1 \odot \cdots \odot T_m) = S_1T_1 \odot \cdots \odot S_mT_m$$

as elements of $\mathcal{L}(V_1 \otimes \cdots \otimes V_m, Z_1 \otimes \cdots \otimes Z_m)$.

From this we get the following "obvious" fact:

III.11.14.25. PROPOSITION. Let V_1, V_2, W_1, W_2 be vector spaces over $\mathbb{F}$. If $T_1 : V_1 \to W_1$ and $T_2 : V_2 \to W_2$ are isomorphisms, then $T_1 \odot T_2 : V_1 \otimes V_2 \to W_1 \otimes W_2$ is an isomorphism with inverse $T_1^{-1} \odot T_2^{-1}$.

We also similarly have:

III.11.14.26. PROPOSITION. Let V_1, V_2, W_1, W_2 be vector spaces over $\mathbb{F}$. If $T_1 : V_1 \to W_1$ and $T_2 : V_2 \to W_2$ are surjective, then $T_1 \odot T_2 : V_1 \otimes V_2 \to W_1 \otimes W_2$ is also surjective.

PROOF: Let $w_1 \in W_1$ and $w_2 \in W_2$. Then there are $v_1 \in V_1$ and $v_2 \in V_2$ such that $T_1 v_1 = w_1$ and $T_2 v_2 = w_2$. Then $(T_1 \odot T_2)(v_1 \otimes v_2) = w_1 \otimes w_2$ Thus every elementary tensor in $W_1 \otimes W_2$ is the image of an elementary tensor in $V_1 \otimes V_2$. 👍

Direct Sums and Subspaces

III.11.14.27. Let V_1, V_2, W be vector spaces over $\mathbb{F}$. The map $(v_1 \oplus v_2, w) \mapsto (v_1 \otimes w) \oplus (v_2 \otimes w)$ is a bilinear map from $(V_1 \oplus V_2) \times W$ to $(V_1 \otimes W) \oplus (V_2 \otimes W)$, hence gives a linear map T from $(V_1 \oplus V_2) \otimes W$ to $(V_1 \otimes W) \oplus (V_2 \otimes W)$, with

$$T((v_1 \oplus v_2) \otimes w) = (v_1 \otimes w) \oplus (v_2 \otimes w) \ .$$

In the other direction, the inclusion maps $R_1 : V_1 \to V_1 \oplus V_2$ and $R_2 : V_2 \to V_1 \oplus V_2$, $R_1(v_1) = v_1 \oplus 0$, $R_2(v_2) = 0 \oplus V_2$ induce maps $R_1 \odot 1 : V_1 \otimes W \to (V_1 \oplus V_2) \otimes W$ and $R_2 \odot 1 : V_2 \otimes W \to (V_1 \oplus V_2) \otimes W$, and hence we can define $S : (V_1 \otimes W) \oplus (V_2 \otimes W) \to (V_1 \oplus V_2) \otimes W$ by

$$S((v_1 \otimes w_1) \oplus (v_2 \otimes w_2)) = (R_1 \odot 1)(v_1 \otimes w_1) + (R_2 \odot 1)(v_2 \otimes w_2) \ .$$

It is clear that S and T are inverses, so we obtain:

III.11.14.28. PROPOSITION. The maps S and T give natural isomorphisms

$$(V_1 \oplus V_2) \otimes W \cong (V_1 \otimes W) \oplus (V_2 \otimes W) \ .$$

By iteration, the same holds for finite direct sums in each factor of a multiple tensor product. Using III.11.19.4., the result extends to arbitrary direct sums.

Flatness

We have a result which may seem as "obvious" as III.11.14.25. (and which is overlooked in many references); but it is not obvious after all and requires proof. If the AC is assumed (in the form of III.11.2.38.), or in the finite-dimensional case where the AC is not needed for III.11.2.38., it is an almost immediate consequence of III.11.14.28., but a proof without the AC takes more work, cf. III.11.19.5..

III.11.14.29. Corollary. Let V and W be vector spaces over $\mathbb{F}$, and let X be a subspace of V. Let T be the inclusion map from X to V. Then $T \odot 1 : X \otimes W \to V \otimes W$ is injective.

Proof: (AC) Let Y be a subspace of V complementary to X; then $V \cong X \oplus Y$. Then $T \odot 1$ is the composition of the following maps:

$$X \otimes W \to (X \otimes W) \oplus (Y \otimes W) \to (X \oplus Y) \otimes W \to V \otimes W$$

where the first map sends $x \otimes w$ to $(x \otimes w) \oplus 0$, the second map is the map S of III.11.14.28., and the third map is from III.11.14.25.. Since all the maps are injective, so is their composition. 👍

We are thus justified in regarding $X \otimes W$ as a subspace of $V \otimes W$.

III.11.14.30. If W is a vector space such that the map $T \odot 1$ is injective for every V and X, then W is called *flat* (III.11.15.20.). The result is then that every vector space is flat.

III.11.15. Modules

Modules are "vector spaces over rings." If the ring is noncommutative (and, to a degree, even in the commutative case), we must distinguish between left modules and right modules. The definitions are essentially the same, and some of the theory is identical, but module theory is much more complicated (and interesting!) than the theory of vector spaces. We give only the rudiments here.

III.11.15.1. DEFINITION. Let R be a ring. A *left R-module* is a set M with two binary operations $+ : M \times M \to M$ (written $+(x, y) = x + y$) and $\cdot : R \times M \to M$ (written $\cdot(\alpha, x) = \alpha \cdot x$ or usually just αx) satisfying:

(A1) The binary operation $+$ is *associative*: $x + (y + z) = (x + y) + z$ for all $x, y, z \in M$.

(A2) The binary operation $+$ is *commutative*: $x + y = y + x$ for all $x, y \in M$.

(A3) There is an element $0 \in M$ such that $0 + x = x$ for all $x \in M$.

(A4) For every $x \in M$ there is an element $(-x) \in M$ with $x + (-x) = 0$.

(M1) The binary operation $\cdot$ is *associative*: $\alpha(\beta x) = (\alpha\beta)x$ for all $\alpha, \beta \in R$, $x \in M$.

(M2) The binary operation $\cdot$ is *distributive*: $\alpha(x + y) = \alpha x + \alpha y$ and $(\alpha + \beta)x = \alpha x + \beta x$ for all $\alpha, \beta \in R$, $x, y \in M$.

If R is unital, we also require

(M3) $1_R \cdot x = x$ for all $x \in M$.

If R is unital, a left R-module is not always required to satisfy (M3) by some authors; one satisfying (M3) is called a *unital left R-module*. We will only consider unital modules. This is not a serious restriction, cf. III.11.19.7..

III.11.15.2. DEFINITION. Let R be a ring. A *right R-module* is a set M with two binary operations $+ : M \times M \to M$ (written $+(x, y) = x + y$) and $\cdot : M \times R \to M$ (written $\cdot(x, \alpha) = x \cdot \alpha$ or usually just $x\alpha$) satisfying:

(A1) The binary operation $+$ is *associative*: $x + (y + z) = (x + y) + z$ for all $x, y, z \in M$.

(A2) The binary operation $+$ is *commutative*: $x + y = y + x$ for all $x, y \in M$.

(A3) There is an element $0 \in M$ such that $0 + x = x$ for all $x \in M$.

(A4) For every $x \in M$ there is an element $(-x) \in M$ with $x + (-x) = 0$.

(M1) The binary operation $\cdot$ is *associative*: $(x\alpha)\beta = x(\alpha\beta)$ for all $\alpha, \beta \in R$, $x \in M$.

(M2) The binary operation $\cdot$ is *distributive*: $(x + y)\alpha = x\alpha + y\alpha$ and $x(\alpha + \beta) = x\alpha + x\beta$ for all $\alpha, \beta \in R$, $x, y \in M$.

If R is unital, we also require

(M3) $x \cdot 1_R = x$ for all $x \in M$.

III.11.15.3. EXAMPLES. (i) If R is a field, a left R-module is just a vector space over R.

(ii) For any ring R, a (left or right) R-module is an abelian group under addition if the R-action is ignored. In particular, a (left or right) $\mathbb{Z}$-module is an abelian group under addition. Conversely, if G is an abelian group, with operation written additively, G has a unique $\mathbb{Z}$-action making it into a left $\mathbb{Z}$-module: $nx = x + \cdots + x$ (n x's) for $x \in G$ if $n \in \mathbb{N}$, $0x = 0$, $(-n)x = -(nx)$ if $n \in \mathbb{N}$. Thus left $\mathbb{Z}$-modules are just abelian groups. Similarly, there is a right $\mathbb{Z}$-action on an abelian group G defined in the same way, with $xn = nx$, so right $\mathbb{Z}$-modules, left $\mathbb{Z}$-modules, and abelian groups are really the same thing (just notationally different).

(iii) If R is any ring, R is both a left R-module under left multiplication and a right R-module under right multiplication. More generally, a left ideal in R is a left R-module, and a right ideal is a right R-module.

(iv) If R is a ring, then R^n is a left R-module under coordinatewise addition and left scalar multiplication, and similarly a right R-module under right coordinatewise scalar multiplication. A module isomorphic to one of these is a *free* left (or right) R-module. More generally, an arbitrary direct sum of copies of R is a free R-module.

III.11.15.4. There are several standard operations on modules, done just as with vector spaces: direct sums, submodules, quotient modules.

There is an analog of linear transformations:

III.11.15.5. DEFINITION. Let R be a ring, and M and N left R-modules. A function $T : M \to N$ is an *R-module morphism*, or *R-linear*, if

(i) $T(x + y) = Tx + Ty$ for all $x, y \in M$.

(ii) $T(\alpha x) = \alpha(Tx)$ for all $\alpha \in R$, $x \in M$.

The set of R-linear functions from M to N is denoted $Hom_R(M, N)$ or $\mathcal{L}_R(M, N)$, or just $Hom(M, N)$ or $\mathcal{L}(M, N)$ if R is understood.

Module morphisms between right R-modules are defined similarly.

III.11.15.6. Module morphisms behave like linear transformations: if $T \in \mathcal{L}_R(M, N)$, then

$T(0) = 0$.

If K is a submodule of M, then $T(K)$ is a submodule of N. In particular, the range of T is a submodule of N.

If K is a submodule of N, then $T^{-1}(K)$ is a submodule of M. In particular, the *kernel* (or *null space*) $Ker(T) = \{x \in M : Tx = 0\} = T^{-1}(\{0\})$ is a submodule of M.

Compositions of R-linear maps are R-linear.

These are proved just as for vector spaces.

For fixed R, left R-modules and R-linear maps form an abelian category () (as do right R-modules).

III.11.15.7. The biggest difference between modules and vector spaces, which causes most of the difficulties (and richness) in module theory, is that there is no analog of III.11.2.28.. Linear independence, span, and bases can be defined in left R-modules just as in vector spaces, but are much rarer: most modules do not have a basis, i.e. most modules are not free modules. For example, quotient modules or left ideals are rarely free modules. The fundamental fact about vector spaces is that every finite-dimensional vector space, and, under AC, every vector space (i.e. module over a field), is a free module, and every submodule is a direct summand (i.e. every short exact sequence of vector spaces splits). "Most" short exact sequences of modules over rings which are not fields do not split (e.g. III.11.15.19.).

Bimodules

III.11.15.8. DEFINITION. Let R and S be rings. An (S, R)-*bimodule* is a set M with a left action of S making it a left S-module and a right action of R making it a right R-module, with the compatibility (associative) property that $(sx)r = s(xr)$ for all $x \in M$, $r \in R$, $s \in S$.

We usually consider only the case where R and S are unital, in which case we require that $1_S x = x 1_R = x$ for all $x \in M$.

One of the most important situations is when $S = R$. We call an (R, R)-bimodule an R-*bimodule*.

III.11.15.9. EXAMPLES. (i) Let R be any ring, not necessarily commutative. Then R becomes an R-bimodule under the actions $r \cdot x = rx$ and $x \cdot r = xr$. The compatibility condition is just associativity of multiplication in R.
More generally, if J is a two-sided ideal in R, then J becomes an R-bimodule with the same actions.

(ii) Let R be a ring. The set $M_{m,n}(R)$ of $m \times n$ matrices over R with its usual addition is a left $M_m(R)$-module under left multiplication and a right $M_n(R)$-module under right multiplication. The actions are compatible by the associativity of matrix multiplication, so $M_{m,n}(R)$ is a $(M_m(R), M_n(R))$-bimodule.

(iii) Similarly, let V and W be vector spaces over a field $\mathbb{F}$. Then $\mathcal{L}(V, W)$ has a *right* action of $\mathcal{L}(V)$ given by $T \cdot S = T \circ S$ for $T \in \mathcal{L}(V, W)$, $S \in \mathcal{L}(V)$, and a *left* action of $\mathcal{L}(W)$ given by $S \cdot T = S \circ T$ for $T \in \mathcal{L}(V, W)$, $S \in \mathcal{L}(W)$. (Carefully note the order: left and right for actions can be confusing.) The actions are compatible by associativity of function composition, so $\mathcal{L}(V, W)$ is a $(\mathcal{L}(W), \mathcal{L}(V))$-bimodule.

III.11.15.10. If R is commutative, there is a canonical extension of a left action to a compatible right action: if M is a left R-module, define $x \cdot r = rx$ for $x \in M$, $r \in R$. It is easy to check that this makes M into an R-bimodule. This bimodule is called the *standard extension* of the left module to a bimodule. Similarly, any right R-module has a standard extension to an R-bimodule; the standard extension (or standard bimodule) has identical left and right actions, i.e. $rx = xr$ for all x and r. So we will consider any R-module to be an R-bimodule via the standard extension. In particular, any abelian group is a $\mathbb{Z}$-bimodule, and any vector space over a field $\mathbb{F}$ is an $\mathbb{F}$-bimodule.

Commutativity of R is essential here: the construction simply does not work for a noncommutative R, and in fact there is no way to make a general left R-module (e.g. a left ideal in R) into an R-bimodule, at least in the unital case (there is a stupid nonunital way to do it, setting $x \cdot r = 0$ for all x and r). The formula $x \cdot r = rx$ makes M a right module over the opposite ring R^{op} of R (R with multiplication order reversed), but does not make M into an (R, R^{op})-bimodule since the compatibility condition is not satisfied for R noncommutative.

III.11.15.11. EXAMPLE. Not every R-bimodule for R commutative is a standard R-bimodule, i.e. the left and right actions can be different, even if R is a field. Make $\mathbb{C}$ into a left $\mathbb{C}$-module (vector space) by $\alpha \cdot z = \alpha z$, and into a right $\mathbb{C}$-module by $z \cdot \alpha = \bar{\alpha} z$. It is easily checked that this makes $\mathbb{C}$ into a nonstandard $\mathbb{C}$-bimodule.

Tensor Products of Modules

Tensor products can be defined in considerable generality analogously to the definition of tensor products of vector spaces. Tensor products of *free* modules behave in much the same way as tensor products of vector spaces, but the theory of tensor products of general modules is fraught with considerable subtleties which are beyond the scope of this book. We only briefly describe the general construction.

III.11.15.12. For notational simplicity, let us concentrate on the tensor product of two modules. Let R be a ring, not necessarily commutative; to avoid pathologies we will assume that R is unital. Let M be a right R-module and N a left R-module. Form the free abelian group $\mathcal{G}$ with generators $\{(x, y) : x \in M, y \in N\}$, and let $\mathcal{H}$ be the subgroup of (finite) sums of elements of the form

$$(x_1 + x_2, y) - (x_1, y) - (x_2, y)$$

$$(x, y_1 + y_2) - (x, y_1) - (x, y_2)$$

$$(xr, y) - (x, ry)$$

for $x, x_1, x_2 \in M$, $y, y_1, y_2 \in N$, $r \in R$ (the set of such sums is indeed a subgroup). Set $M \otimes_R N = \mathcal{G}/\mathcal{H}$; then $M \otimes_R N$ is an abelian group. Write $x \otimes y$ for the image of (x, y) in $M \otimes_R N$; then every element of $M \otimes_R N$ is a finite sum of such elementary tensors (in a highly nonunique way in general).

III.11.15.13. If M is an (S, R)-bimodule, where S is another (unital) ring, then we can define $s(x \otimes y) = (sx) \otimes y$ for elementary tensors, for $s \in S$, $x \in M$, $y \in N$; it is an easy exercise to see that this extends to a well-defined left S-module structure on $M \otimes_R N$.

Similarly, if N is an (R, T)-bimodule for a (unital) ring T, then $M \otimes_R N$ has a natural structure as a right T-module with $(x \otimes y)t = x \otimes (yt)$ for $x \in M$, $y \in N$, $t \in T$.

If both things occur, i.e. M is an (S, R)-bimodule and N an (R, T)-bimodule, then it is easily checked that $M \otimes_R N$ is an (S, T)-bimodule.

III.11.15.14. More generally, if $R_0, R_1, \ldots, R_m$ are (unital) rings, and for $1 \leq k \leq m$ M_k is an (R_{k-1}, R_k)-bimodule, we can form

$$M_1 \otimes_{R_1} M_2 \otimes_{R_2} \cdots \otimes_{R_{m-1}} M_m$$

which is an (R_0, R_m)-bimodule.

III.11.15.15. The most common situation is where R is a (unital) commutative ring and M and N are R-modules (hence R-bimodules ()). Then $M \otimes_R N$ is naturally an R-(bi)module. If $M_1, \ldots, M_m$ are R-modules, then $M_1 \otimes_R M_2 \otimes_R \cdots \otimes_R M_m$ is a well-defined R-(bi)module. If R is a field, this is nothing but the vector space tensor product as defined in ().

Exact Sequences and Flatness

In this subsection, we will only consider modules over a fixed *commutative* unital ring R.

III.11.15.16. Recall that an *exact sequence* (say, of R-modules) is a finite or infinite chain of R-modules with R-module homomorphisms from each to the next, such that the image of each is exactly the kernel of the next. Most important for our purposes are *short exact sequences*, of the form

$$0 \longrightarrow N \xrightarrow{\iota} M \xrightarrow{\pi} Q \longrightarrow 0$$

which is exact at N precisely when ι is injective and exact at Q precisely when π is surjective, and exact at M precisely when the kernel of π is N, so that $Q \cong M/N$.

Here we examine the question: does tensoring with a fixed R-module preserve (short) exact sequences, i.e. if

$$0 \longrightarrow N \xrightarrow{\iota} M \xrightarrow{\pi} Q \longrightarrow 0$$

is a short exact sequence of R-modules and F is another R-module, is the sequence

$$0 \longrightarrow N \otimes_R F \xrightarrow{\iota\odot 1} M \otimes_R F \xrightarrow{\pi\odot 1} Q \otimes_R F \longrightarrow 0$$

necessarily exact?

The general answer is "no, not quite, but almost":

III.11.15.17. Proposition. Let R be a commutative unital ring, and N, M, Q, F R-modules. If

$$0 \longrightarrow N \xrightarrow{\iota} M \xrightarrow{\pi} Q \longrightarrow 0$$

is a short exact sequence, then

$$N \otimes_R F \xrightarrow{\iota\odot 1} M \otimes_R F \xrightarrow{\pi\odot 1} Q \otimes_R F \longrightarrow 0$$

is also exact (we say tensoring with F over R is *right exact*).

Proof: It is obvious that the sequence is exact at $Q \otimes_R F$, since every elementary tensor in $Q \otimes_R F$ is the image of an elementary tensor in $M \otimes_R F$ (cf. III.11.14.26.). So we must prove the sequence is exact at $M \otimes_R F$.

Let $K = (\iota \odot 1)(N \otimes_R F)$. Then K is a submodule of $M \otimes_R F$, and we must show that $K = Ker(\pi \odot 1)$. We clearly have $K \subseteq Ker(\pi \odot 1)$ since the image of every elementary tensor in $N \otimes_R F$ is in the kernel of $\pi \odot 1$. Form the quotient module $(M \otimes_R F)/K$, and let $\rho : M \otimes_R F \to (M \otimes_R F)/K$ be the quotient map. The map $\pi \odot 1$ factors through ρ since $K \subseteq Ker(\pi \odot 1)$, i.e. there is $T : (M \otimes_R F)/K \to Q \otimes_R F$ with $\pi \odot 1 = T \circ \rho$. On the other hand, there is a map Φ from $Q \times F$ to $(M \otimes_R F)/K$ defined as follows: if $(q, f) \in Q \times F$, choose $x \in M$ with $\pi(x) = q$, and set $\Phi(q, f) = \rho(x \otimes f)$. Φ is well defined since if $\pi(x_1) = \pi(x_2) = q$, then $x_1 - x_2 \in N$, so $(x_1 - x_2) \otimes f \in K$. It is clear that Φ is bilinear, so defines a linear map $S : Q \otimes_R F \to (M \otimes_R F)/K$. S is surjective since the image of every elementary tensor in $M \otimes_R F$ is in the image, and $T \circ S$ is the identity on $Q \otimes_R F$, so T is injective and $K = Ker(\pi \odot 1)$. 👍

III.11.15.18. Note that the conclusion does not have a short exact sequence since there is no 0 at the beginning, i.e. it is not asserted that the map $\iota \odot 1$ is injective; in fact it is not in general.

III.11.15.19. Example. Let $R = \mathbb{Z}$, so we just have abelian groups. Consider the exact sequence

$$0 \longrightarrow 2\mathbb{Z} \xrightarrow{\iota} \mathbb{Z} \xrightarrow{\pi} \mathbb{Z}_2 \longrightarrow 0$$

of abelian groups. Take $F = \mathbb{Z}_2$. We have $\mathbb{Z} \otimes_{\mathbb{Z}} Z_2 \cong \mathbb{Z}_2 \otimes_{\mathbb{Z}} \mathbb{Z}_2 \cong \mathbb{Z}_2$, and the map $\iota \odot 1$ is the zero map in this case. The exact sequence in the conclusion is

$$\mathbb{Z}_2 \xrightarrow{0} Z_2 \xrightarrow{1} \mathbb{Z}_2 \longrightarrow 0 \quad .$$

III.11.15.20. Definition. An R-module F is *flat* if, for any short exact sequence

$$0 \longrightarrow N \xrightarrow{\iota} M \xrightarrow{\pi} Q \longrightarrow 0$$

of R-modules, the tensored sequence

$$0 \longrightarrow N \otimes_R F \xrightarrow{\iota \odot 1} M \otimes_R F \xrightarrow{\pi \odot 1} Q \otimes_R F \longrightarrow 0$$

is exact, i.e. the map $\iota \odot 1$ is injective.

III.11.15.21. Some facts about flat modules:

(i) Any free module is flat (cf. III.11.19.3.).

(ii) A direct summand of a flat module is flat (cf. ()). In particular, a projective module (III.11.19.9.) is flat.

(iii) A direct limit of flat modules is flat (cf. III.11.19.4.).

(iv) A $\mathbb{Z}$-module (abelian group) is flat if and only if it is torsion-free. (This may be the origin of the term "flat," meaning "no torsion.")

(v) Any module over a field (vector space) is a flat module (III.11.14.29., III.11.19.5.).

III.11.16. The Tensor Algebra of a Vector Space

In this section, $\mathbb{F}$ will be a fixed field.

Tensor Powers of a Vector Space

III.11.16.1. Definition. Let V be a vector space over $\mathbb{F}$, and $k \in \mathbb{N}$. The *k'th tensor power* of V is $V \otimes V \otimes \cdots \otimes V$ (k factors). This vector space is variously denoted $\mathcal{T}^k(V)$, $\bigotimes^k V$, or $V^{\otimes k}$. We also define $\mathcal{T}^0(V)$ to be $\mathbb{F}$.

III.11.16.2. If V is n-dimensional, with basis $\{x_1, \ldots, x_n\}$, then for $k \in \mathbb{N}$ $\mathcal{T}^k(V)$ is also finite-dimensional, with dimension n^k, and

$$\{x_{j_1} \otimes x_{j_2} \otimes \cdots \otimes x_{j_k} : 1 \leq j_1, j_2, \ldots, j_k \leq n\}$$

is a basis (III.11.14.13.). $\mathcal{T}^0(V)$ also fits the dimension pattern: $dim(\mathcal{T}^0(V)) = n^0$.

III.11.16.3. If V and W are vector spaces over $\mathbb{F}$, linear transformations from $\mathcal{T}^k(V)$ to W correspond canonically to k-linear maps from $V \times \cdots \times V$ to W, the set of which is denoted $\mathcal{L}^k(V, W)$; thus $\mathcal{L}(\mathcal{T}^k(V), W) \cong \mathcal{L}^k(V, W)$. If V is finite-dimensional, then by III.11.14.21.

$$\mathcal{L}(\mathcal{T}^k(V), W) \cong \mathcal{L}^k(V, W) \cong \mathcal{T}^k(\mathcal{L}(V, W)) \ .$$

In particular, taking $W = \mathbb{F}$,

$$(\mathcal{T}^k(V))^* \cong \mathcal{L}^k(V, \mathbb{F}) \cong \mathcal{T}^k(V^*) \ .$$

Caution: Some authors use the notation $\mathcal{T}^k(V)$ to denote what we call $\mathcal{T}^k(V^*)$, i.e. $\mathcal{L}^k(V, \mathbb{F})$; for these authors our $\mathcal{T}^k(V)$ would be written $\mathcal{T}^k(V^*)$ if V is finite-dimensional. Some authors write $\otimes_k(V)$ for $\mathcal{T}^k(V)$ and $\otimes^k(V)$ for $\mathcal{T}^k(V^*)$. The subscript/superscript notation has some advantages, but we will not use it in this context (it is universally used in homological algebra, where we *will* use it).



The Tensor Algebra of a Vector Space

III.11.16.4. There is a "multiplication" $\mathcal{T}^k(V) \times \mathcal{T}^\ell(V) \to \mathcal{T}^{k+\ell}(V)$ defined on elementary tensors by

$$(v_1 \otimes v_2 \otimes \cdots \otimes v_k)(w_1 \otimes w_2 \otimes \cdots \otimes w_\ell) = v_1 \otimes v_2 \otimes \cdots \otimes v_k \otimes w_1 \otimes w_2 \otimes \cdots \otimes w_\ell$$

and extended linearly. This map can be shown to be well defined by the usual multilinearity argument (it is essentially a version of III.11.14.9.). The "multiplication" is associative in the obvious sense, but far from commutative.

III.11.16.5. We can convert this into a true multiplication. Define

$$\mathcal{T}(V) = \bigoplus_{k=0}^{\infty} \mathcal{T}^k(V)$$

with coordinatewise addition. There is a multiplication on $\mathcal{T}(V)$ defined on elementary tensors as above and extended linearly. This multiplication is associative, distributes over addition, and is compatible with scalar multiplication (which is just multiplication by elements of $\mathcal{T}^0(V) = \mathbb{F}$). Thus $\mathcal{T}(V)$ is an algebra over $\mathbb{F}$ (in particular a ring). $\mathcal{T}(V)$ is unital: the unit is $1 \in \mathbb{F} = \mathcal{T}^0(V)$. $\mathcal{T}(V)$ is infinite-dimensional (if $V \neq \{0\}$), and highly noncommutative if $dim(V) > 1$.

III.11.16.6. Definition. $\mathcal{T}(V)$ is called the *tensor algebra* of V.

Graded Algebras

III.11.16.7. DEFINITION. A *graded algebra* over a field $\mathbb{F}$ is a unital algebra A with a decomposition

$$A = \bigoplus_{k=0}^{\infty} A_k$$

of A into vector subspaces (some A_k may be $\{0\}$, e.g. the direct sum may be finite), with $A_0 = \mathbb{F}$, such that if $x \in A_k$ and $y \in A_\ell$, then $xy \in A_{k+\ell}$ and $1 \in A_0 = \mathbb{F}$ is the unit of A. The elements of A_k are called the *elements of A of degree k*.

If $A = \oplus A_k$ and $B = \oplus B_k$ are graded algebras over $\mathbb{F}$, a unital algebra homomorphism $\phi : A \to B$ is a *graded algebra-homomorphism* if $\phi(A_k) \subseteq B_k$ for all k.

III.11.16.8. The prototype example of a graded algebra is $\mathcal{T}(V)$ for a vector space V: the elements of $\mathcal{T}^k(V)$ are degree k. This algebra is generated as an algebra by 1 and the elements of $\mathcal{T}^1(V) = V$.

In fact, $\mathcal{T}(V)$ is the "free graded algebra" on V, with the following universal property:

III.11.16.9. THEOREM. Let V be a vector space over $\mathbb{F}$. If A is any graded algebra over $\mathbb{F}$ and T is any linear map from V to A_1, then T extends uniquely to a unital graded algebra-homomorphism from $\mathcal{T}(V)$ to A.

As a special case, the tensor algebra construction is functorial:

III.11.16.10. COROLLARY. Let V and W be vector spaces over $\mathbb{F}$. Any linear map $T : V \to W$ extends uniquely to a unital graded algebra-homomorphism $\otimes T$ from $\mathcal{T}(V)$ to $\mathcal{T}(W)$ (write $\otimes^k T$ for the map from $\mathcal{T}^k(V)$ to $\mathcal{T}^k(W)$). On elementary tensors we have

$$[\otimes^k T](v_1 \otimes \cdots \otimes v_k) = (Tv_1) \otimes \cdots \otimes (Tv_k) \ .$$

An immediate consequence of the uniqueness is that the tensor algebra construction is natural:

III.11.16.11. COROLLARY. Let V, W, and Z be vector spaces over $\mathbb{F}$. If $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, Z)$, then $\otimes(S \circ T) = (\otimes S) \circ (\otimes T)$.

PROOF: Both sides are extensions of $S \circ T$. 👍

Tensor powers of an operator preserve many properties:

III.11.16.12. PROPOSITION. Let V and W be vector spaces over $\mathbb{F}$, and $T : V \to W$ a linear transformation. Then

(i) If T is injective, so is $\otimes^k T$ for each k.

(ii) If T is surjective, so is $\otimes^k T$ for each k..

(i) If $V = W$ and T is idempotent, so is $\otimes^k T$ for each k.

PROOF: (i) There is an $S \in \mathcal{L}(W, V)$ with $S \circ T = I_V$, so for each k $(\otimes^k S) \circ (\otimes^k T = I_{\mathcal{T}^k(V)}$ and $\otimes^k T$ is injective.

(ii) Almost identical to (i), but with $T \circ S = I_W$.

(iii) is obvious from III.11.16.11.. 👍

The AC is needed for (i) and (ii), but not if V and W are finite-dimensional.

III.11.17. The Exterior Algebra of a Vector Space

The literature on exterior algebra is somewhat of a mess, with inconsistent notations and definitions. We will try to find the best path through the subject, but what we do will inevitably be in conflict with some other references; we will try to highlight the inconsistencies. ✋

In this section, $\mathbb{F}$ will be a fixed field of characteristic 0 (there are technical difficulties which will become apparent if $\mathbb{F}$ has characteristic $p > 0$, especially if $p = 2$). Applications to analysis almost exclusively use the case $\mathbb{F} = \mathbb{R}$, so the reader can just take $\mathbb{F} = \mathbb{R}$ if desired. In fact, we will have to reduce to $\mathbb{F} = \mathbb{R}$ in some places, e.g. ().

The k'th exterior power of a vector space can be equivalently defined as either a subspace or a quotient of the k'th tensor power. It is useful to have both pictures for different purposes, although the quotient picture seems to be more useful since it corresponds to a universal property similar to the universal property for tensor products.

Exterior Powers of a Vector Space (Quotient Construction)

III.11.17.1. DEFINITION. Let V be a vector space over $\mathbb{F}$. An elementary tensor $v_1 \otimes \cdots \otimes v_k$ in $\mathcal{T}^k(V)$ is *redundant* (or *repetitive*) if $v_i = v_j$ for some $i \neq j$. Let $\mathcal{R}^k(V)$ be the linear span of the redundant elementary tensors in $\mathcal{T}^k(V)$, and set $\bigwedge^k(V) = \mathcal{T}^k(V)/\mathcal{R}^k(V)$. $\bigwedge^k(V)$ is called the *k'th exterior power* of V. If $k = 1$, we set $\mathcal{R}^k(V) = 0$ (redundancy cannot occur in a 1-tuple!) and $\bigwedge^1(V) = \mathcal{T}^1(V) = V$.

If $v_1, \ldots, v_k \in V$, denote by $v_1 \wedge \cdots \wedge v_k$ the image of $v_1 \otimes \cdots \otimes v_k$ in $\bigwedge^k(V)$.

Thus every element of $\bigwedge^k(V)$ is a finite sum of elementary wedges, in a highly nonunique way in general. There is, of course, the same collapsing as in the tensor product, and much more:

III.11.17.2. PROPOSITION. Let V be a vector space over $\mathbb{F}$, and $v_1, \ldots, v_k \in V$. For any i and j, $1 \leq i < j \leq k$,

$$v_1 \wedge \cdots \wedge v_{i-1} \wedge v_j \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_i \wedge v_{j+1} \wedge \cdots \wedge v_k = -v_1 \wedge \cdots \wedge v_{i-1} \wedge v_i \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_j \wedge v_{j+1} \wedge \cdots \wedge v_k$$

(in other words, interchanging two elements in an elementary wedge changes the sign.)

PROOF: We have

$$v_1 \otimes \cdots \wedge v_{i-1} \otimes (v_i + v_j) \otimes v_{i+1} \otimes \cdots \otimes v_{j-1} \otimes (v_i + v_j) \otimes v_{j+1} \otimes \cdots \otimes v_k \in \mathcal{R}^k(V)$$

$$0 = v_1 \wedge \cdots \wedge v_{i-1} \wedge (v_i + v_j) \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge (v_i + v_j) \wedge v_{j+1} \wedge \cdots \wedge v_k$$

$$= v_1 \wedge \cdots \wedge v_{i-1} \wedge v_i \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_i \wedge v_{j+1} \wedge \cdots \wedge v_k + v_1 \wedge \cdots \wedge v_{i-1} \wedge v_j \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_j \wedge v_{j+1} \wedge \cdots \wedge v_k$$

$$+ v_1 \wedge \cdots \wedge v_{i-1} \wedge v_i \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_j \wedge v_{j+1} \wedge \cdots \wedge v_k + v_1 \wedge \cdots \wedge v_{i-1} \wedge v_j \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_i \wedge v_{j+1} \wedge \cdots \wedge v_k$$

$$= v_1 \wedge \cdots \wedge v_{i-1} \wedge v_i \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_j \wedge v_{j+1} \wedge \cdots \wedge v_k + v_1 \wedge \cdots \wedge v_{i-1} \wedge v_j \wedge v_{i+1} \wedge \cdots \wedge v_{j-1} \wedge v_i \wedge v_{j+1} \wedge \cdots \wedge v_k$$

since the first two terms are 0, so this sum is also 0. 👍

(If $\mathbb{F}$ had characteristic 2, what would this result say?)

III.11.17.3. Recall that a permutation σ of $\{1, \ldots, k\}$ has a *sign* $\operatorname{sgn}(\sigma) = \pm 1$ depending whether σ is a product of an even or odd number of transpositions. Denote by S_k the group of permutations of $\{1, \ldots, k\}$. Then S_k is a group with $k!$ elements. By repeated applications of III.11.17.2., we obtain:

III.11.17.4. COROLLARY. Let V be a vector space over $\mathbb{F}$, and $v_1, \ldots, v_k \in V$. Let $\sigma \in S_k$. Then

$$v_{\sigma(1)} \wedge \cdots \wedge v_{\sigma(k)} = \operatorname{sgn}(\sigma)[v_1 \wedge \cdots \wedge v_k]$$

(in other words, permuting elements in an elementary wedge preserves or changes the sign depending whether the permutation is even or odd.)

The Universal Property

III.11.17.5. DEFINITION. Let V and W be vector spaces over $\mathbb{F}$. A k-linear function $T : V \times \cdots \times V \to W$ is *alternating* if $T(v_1, \ldots, v_k) = 0$ whenever $v_i = v_j$ for some $i \neq j$. Denote by $\mathcal{L}^k_{Alt}(V, W)$ the set of alternating k-linear maps from V to W. (We declare any 1-linear map from V to W to be alternating, i.e. $\mathcal{L}^1_{Alt}(V, W) = \mathcal{L}^1(V, W) = \mathcal{L}(V, W)$.)

The notation $\mathcal{L}^k_a(V, W)$ is used for $\mathcal{L}^k_{Alt}(V, W)$ by some authors, but this is in slight conflict with the notation of ().

Alternating k-linear functions have similar properties to the ones in III.11.17.2. and III.11.17.4., proved the same way:

III.11.17.6. PROPOSITION. Let V and W be vector spaces over $\mathbb{F}$, and $T \in \mathcal{L}^k_{Alt}(V,W)$. Then

$$T(v_{\sigma(1)},\dots,v_{\sigma(k)}) = \operatorname{sgn}(\sigma)T(v_1,\dots,v_k)$$

for any $(v_1,\dots,v_k) \in V^k$ and $\sigma \in S_k$. In particular,

$$T(v_1,\dots,v_{i-1},v_j,v_{i+1},\dots,v_{j-1},v_i,v_{j+1},\dots,v_k) = -T(v_1,\dots,v_{i-1},v_i,v_{i+1},\dots,v_{j-1},v_j,v_{j+1},\dots,v_k)$$

for $(v_1,\dots,v_k) \in V^k$ and $1 \le i < j \le n$.

III.11.17.7. PROPOSITION. Let V be a vector space over $\mathbb{F}$, and $k \in \mathbb{N}$. There is an alternating k-linear function $N : V \times \cdots \times V \to \bigwedge^k(V)$ defined by

$$N(v_1,\dots,v_k) = v_1 \wedge \cdots \wedge v_k$$

which is universal in the sense that if W is any vector space over $\mathbb{F}$ and $T \in \mathcal{L}^k_{Alt}(V,W)$, there is a unique linear function $S : \bigwedge^k(V) \to W$ such that $T = S \circ N$.

PROOF: Let $\rho : \mathcal{T}^k(V) \to \bigwedge^k(V)$ be the quotient map. Then $N = \rho \circ M$, so N is well defined and k-linear, and it is easily checked that N is alternating. Now let W be a vector space and $T \in \mathcal{L}^k_{Alt}(V,W)$. By the universal property of tensor products, there is an $\tilde{S} : \mathcal{T}^k(V) \to W$ with $T = \tilde{S} \circ M$. Since T is alternating, $\tilde{S}(\mathcal{R}^k(V)) = 0$, so $\tilde{S}$ factors through $\mathcal{T}^k(V)/\mathcal{R}^k(V)$, i.e. there is an $S : \bigwedge^k(V) \to W$ with $\tilde{S} = S \circ \rho$, so $T = S \circ \rho \circ M = S \circ N$. Uniqueness of S follows from uniqueness of $\tilde{S}$. 

Thus there is a natural isomorphism between $\mathcal{L}(\bigwedge^k(V),W)$ and $\mathcal{L}^k_{Alt}(V,W)$ for any W; in particular, $(\bigwedge^k(V))^* \cong \bigwedge^k(V^*)$ if V is finite-dimensional (Exercise III.11.19.11.).

Caution: Some authors use the notation $\bigwedge^k(V)$ to denote what we call $\bigwedge^k(V^*)$, i.e. $\mathcal{L}^k_{Alt}(V,\mathbb{F})$; for these authors our $\bigwedge^k(V)$ would be written $\bigwedge^k(V^*)$ if V is finite-dimensional. 

Bases

We now assume V is finite-dimensional, the most important case for applications in vector analysis.

III.11.17.8. DEFINITION. Let V be a finite-dimensional vector space over $\mathbb{F}$. Fix an *ordered* basis $B = \{x_1,\dots,x_n\}$ of V. If $k \in \mathbb{N}$, a *basic elementary wedge* (k-wedge) in $\bigwedge^k(V)$ (with respect to the ordered basis B) is a wedge $x_{j_1} \wedge x_{j_2} \wedge \cdots \wedge x_{j_k}$ where $1 \le j_1 < j_2 < \cdots < j_k \le n$.

III.11.17.9. PROPOSITION. Let V be a finite-dimensional vector space over $\mathbb{F}$. Fix an ordered basis $B = \{x_1,\dots,x_n\}$ of V. If $k \in \mathbb{N}$, the basic elementary wedges with respect to B form a basis for $\bigwedge^k(V)$.

PROOF: Let $x_{i_1} \otimes \cdots \otimes x_{i_k}$ in $\mathcal{T}^k(V)$ be a basic elementary tensor. If $i_r = i_s$ for some $r \ne s$, then $x_{i_1} \wedge \cdots \wedge x_{i_k} = 0$. Otherwise,

$$x_{i_1} \wedge \cdots \wedge x_{i_k} = \pm x_{j_1} \wedge \cdots \wedge x_{j_k}$$

for a unique basic elementary wedge $x_{j_1} \wedge \cdots \wedge x_{j_k}$, where $(j_i, \ldots, j_k)$ is $(i_1, \ldots, i_k)$ listed in increasing order. Thus the basic elementary wedges span $\bigwedge^k(V)$. To show they are linearly independent, let Y be the free vector space with basis consisting of the basic elementary wedges. There is an obvious linear map $T : Y \to \bigwedge^k(V)$. On the other hand, the map sending a basic elementary tensor in $\mathcal{T}^k(V)$ to $\pm$ the appropriate basic elementary wedge as above defines a linear map from $\mathcal{T}^k(V)$ to Y. The corresponding k-linear map from V^k to Y is easily checked to be alternating (using the argument of III.11.17.2.), hence defines a linear map $S : \bigwedge^k(V) \to Y$ which is the inverse of T. 👍

III.11.17.10. Corollary. Let V be an n-dimensional vector space over $\mathbb{F}$, and $k \in \mathbb{N}$. Then
(i) If $k > n$, then $\bigwedge^k(V) = \{0\}$.
(ii) If $1 \leq k \leq n$, then $dim\left(\bigwedge^k(V)\right) = \binom{n}{k} = \dfrac{n!}{k!(n-k)!}$.

Proof: If $k > n$, then every basic elementary tensor is in $\mathcal{R}^k(V)$ by the pigeonhole principle. If $1 \leq k \leq n$, there is a 1-1 correspondence between k-element subsets of $\{1, \ldots, n\}$ and basic elementary k-wedges since such a set has a unique induced ordering. 👍

III.11.17.11. We also define $\mathcal{R}^0(V)$ to be $\{0\}$ and $\bigwedge^0(V) = \mathbb{F}$ (recall that $\mathcal{T}^0(V) = \mathbb{F}$). These definitions are consistent, and we have $dim\left(\bigwedge^0(V)\right) = \binom{n}{0} = 1$.

Exterior Powers of a Vector Space (Subspace Construction)

We continue to work only in the finite-dimensional case, although some of this works more generally.

III.11.17.12. Let V be a finite-dimensional vector space over $\mathbb{F}$, and let $k \in \mathbb{N}$. We define a linear map A_k from $\mathcal{T}^k(V)$ to $\mathcal{T}^k(V)$ by setting

$$A_k(v_1 \otimes \cdots \otimes v_k) = \frac{1}{k!} \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma)[v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(k)}]$$

(this is a place where there would be a technical difficulty if $\mathbb{F}$ were of characteristic $p > 0$ since then $k! = 0$ in $\mathbb{F}$ for $k \geq p$.) The map A_k is well defined by the usual argument. The map A_k is called the *alternating map*, sometimes denoted *Alt* or Alt_k.

III.11.17.13. Proposition. Let V be a finite-dimensional vector space over $\mathbb{F}$, and let $k \in \mathbb{N}$ and A_k be the alternating map. Then
(i) A_k is an idempotent map on $\mathcal{T}^k(V)$.
(ii) The null space $\mathcal{N}(A_k)$ is $\mathcal{R}^k(V)$.

Proof: The proof of (i) is a straightforward calculation left to the reader. For (ii), if $v_1 \otimes \cdots \otimes v_k$ is a redundant tensor, then $A_k(v_1 \otimes \cdots \otimes v_k) = 0$ since the terms in the sum cancel in pairs. Thus $\mathcal{R}^k(V) \subseteq \mathcal{N}(A_k)$,

so A_k drops to a well-defined map from $\bigwedge^k(V)$ to $\mathcal{T}^k(V)$. The images of the basic elementary wedges are linearly independent since they are linear combinations of pairwise disjoint sets of basic elementary tensors, so this map is injective, i.e. $\mathcal{N}(A_k) = \mathcal{R}^k(V)$. 👍

Using III.11.3.15., we obtain:

III.11.17.14. Corollary. Let V be a finite-dimensional vector space over $\mathbb{F}$, and let $k \in \mathbb{N}$ and A_k be the alternating map. Let $\mathcal{A}^k(V)$ be the range of A_k. Then $\mathcal{A}^k(V)$ and $\mathcal{R}^k(V)$ are complementary subspaces of $\mathcal{T}^k(V)$, and A_k drops to an isomorphism from $\bigwedge^k(V)$ to $\mathcal{A}^k(V)$.

Another way of expressing this is: if $v_1, \dots, v_k \in V$, then

$$v_1 \wedge v_2 \wedge \cdots \wedge v_k = \frac{1}{k!} \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma)[v_{\sigma(1)} \wedge v_{\sigma(2)} \wedge \cdots \wedge v_{\sigma(k)}]$$

in $\bigwedge^k(V)$. The expression on the right is the *alternating form* of the elementary wedge, corresponding to the element

$$A_k(v_1 \otimes v_2 \otimes \cdots \otimes v_k) = \frac{1}{k!} \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma)[v_{\sigma(1)} \otimes v_{\sigma(2)} \otimes \cdots \otimes v_{\sigma(k)}]$$

of $\mathcal{A}^k(V)$.

III.11.17.15. We may define a new basis for $\bigwedge^k(V)$ corresponding to an ordered basis $\{x_1, \dots, x_n\}$ of V: if $1 \le j_1 < j_2 < \cdots < j_k \le n$, set

$$x_{(j_1,\dots,j_k)} = \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma)[x_{j_{\sigma(1)}} \wedge \cdots \wedge x_{j_{\sigma(k)}}]$$

so that $x_{(j_1,\dots,j_k)} = k!(x_{j_1} \wedge \cdots \wedge x_{j_k})$. The $x_{(j_1,\dots,j_k)}$ are called the *standard alternating basis* for $\bigwedge^k(V)$. We can also regard $x_{(j_1,\dots,j_k)}$ as an element of $\mathcal{A}^k(V)$:

$$x_{(j_1,\dots,j_k)} = \sum_{\sigma \in S_k} \operatorname{sgn}(\sigma)[x_{j_{\sigma(1)}} \otimes \cdots \otimes x_{j_{\sigma(k)}}] \ .$$

We may also similarly define $x_{(j_1,\dots,j_k)}$ for $\{j_1, \dots, j_k\}$ distinct but not necessarily in increasing order: $x_{(j_1,\dots,j_k)} = \operatorname{sgn}(\sigma) x_{(\sigma(j_1),\dots,\sigma(j_k))}$, where σ permutes $\{j_1, \dots, j_k\}$ to increasing order. Thus the standard alternating basis depends only weakly on the ordering of the basis $\{x_1, \dots, x_n\}$ (if $\mathbb{F} = \mathbb{R}$, it depends only on the orientation defined by the basis ()).

III.11.17.16. Thus we may equally work with $\mathcal{A}^k(V)$ or $\bigwedge^k(V)$. Some authors define and work only with one of them, usually using the notation $\bigwedge^k(V)$ (or $\bigwedge_k(V)$) for whichever one (if $\bigwedge_k(V)$ is used, $\bigwedge^k(V)$ generally denotes what we call $\bigwedge^k(V^*)$). However, as pointed out in [GH19], it is useful to have both and to distinguish them with different names. Elements of $\mathcal{A}^k(V)$ ($k > 1$) are a little hard to work with since there are no elementary tensors (except 0) in $\mathcal{A}^k(V)$, and thus $\bigwedge^k(V)$ is more convenient for most purposes.

The Exterior or Wedge Product

There is inconsistency in the literature about the definition of the exterior or wedge product; the definitions in various references differ by constant factors. We will describe the two most common normalizations.

III.11.17.17. The most straightforward way to define the wedge product is from the tensor product: if V is a vector space over $\mathbb{F}$ and $\{v_1, \dots, v_k\}$ and $\{w_1, \dots, w_\ell\}$ are sets of vectors in V, we define

$$(v_1 \wedge \cdots \wedge v_k) \wedge (w_1 \wedge \cdots w_\ell) = v_1 \wedge \cdots \wedge v_k \wedge w_1 \wedge \cdots \wedge w_\ell$$

and extend by linearity.

III.11.17.18. But we must check that this is well defined as a map from $\bigwedge^k(V) \times \bigwedge^\ell(V)$ to $\bigwedge^{k+\ell}(V)$. This can be done in two ways. One is via the standard argument that this gives an alternating $(k+\ell)$-linear map from $V^k \times V^\ell$ to $\bigwedge^{k+\ell}(V)$, which is straightforward (details are left to the reader).

A slicker way is as follows. Note that the tensor product of a redundant elementary tensor with any elementary tensor, in either order, is redundant. It follows that $\otimes : \mathcal{T}^k(V) \times \mathcal{T}^\ell(V) \to \mathcal{T}^{k+\ell}(V)$ sends $\mathcal{R}^k(V) \times \mathcal{T}^\ell(V)$ and $\mathcal{T}^k(V) \times \mathcal{R}^\ell(V)$ into $\mathcal{R}^{k+\ell}(V)$, and hence drops to a well-defined map from $\bigwedge^k(V) \times \bigwedge^\ell(V)$ to $\bigwedge^{k+\ell}(V)$.

III.11.17.19. If $\omega = v_1 \otimes \cdots \otimes v_k$ is an elementary tensor in $\mathcal{T}^k(V)$, then $\omega \otimes \omega$ is a redundant tensor. Thus, if $\omega = v_1 \wedge \cdots \wedge v_k$ is an elementary wedge in $\bigwedge^k(V)$, then $\omega \wedge \omega = 0$.

However, this is not necessarily true for elements of $\bigwedge^k(V)$ which are not elementary wedges. Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \boldsymbol{e}_4\}$ be the standard ordered basis for $\mathbb{R}^4$, and let $\omega = \boldsymbol{e}_1 \wedge \boldsymbol{e}_2 + \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 \in \bigwedge^2(\mathbb{R}^4)$. Then

$$\omega \wedge \omega = \boldsymbol{e}_1 \wedge \boldsymbol{e}_2 \wedge \boldsymbol{e}_1 \wedge \boldsymbol{e}_2 + \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 \wedge \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 + \boldsymbol{e}_1 \wedge \boldsymbol{e}_2 \wedge \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 + \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 \wedge \boldsymbol{e}_1 \wedge \boldsymbol{e}_2$$

$$= 2\boldsymbol{e}_1 \wedge \boldsymbol{e}_2 \wedge \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 \neq 0$$

since the first two terms are redundant, hence 0, and $\boldsymbol{e}_1 \wedge \boldsymbol{e}_2 \wedge \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 = \boldsymbol{e}_3 \wedge \boldsymbol{e}_4 \wedge \boldsymbol{e}_1 \wedge \boldsymbol{e}_2$ (the rearrangement is an even permutation).

III.11.17.20. The wedge product is not commutative, but is much closer to being commutative than the tensor product. If $\omega = v_1 \wedge \cdots \wedge v_k$ and $\eta = w_1 \wedge \cdots \wedge w_\ell$ are elementary wedges in $\bigwedge^k(V)$ and $\bigwedge^\ell(V)$ respectively, then

$$\omega \wedge \eta = v_1 \wedge \cdots \wedge v_k \wedge w_1 \wedge \cdots \wedge w_\ell$$

$$\eta \wedge \omega = w_1 \wedge \cdots \wedge w_\ell \wedge v_1 \wedge \cdots \wedge v_k$$

in $\bigwedge^{k+\ell}(V)$ are reorderings, so

$$\eta \wedge \omega = \operatorname{sgn}(\sigma)\omega \wedge \eta = \pm\omega \wedge \eta$$

where σ is the permutation of $\{1, \dots, k+\ell\}$ interchanging the first k and the last ℓ numbers. To compute $\operatorname{sgn}(\sigma)$, each number from $k+1$ to $k+\ell$ must be moved past each of the numbers from 1 to k, which can be done with k transpositions. Thus σ can be written as a product of $k\ell$ transpositions, so $\operatorname{sgn}(\sigma) = (-1)^{k\ell}$, i.e.

$$\eta \wedge \omega = (-1)^{k\ell}\omega \wedge \eta$$

(such a multiplication is called *graded commutative*; some authors use the term *anticommutative*, but this is not strictly correct terminology). The same formula applies if ω is just a sum of k-wedges, i.e. $\omega \in \bigwedge^k(V)$, and $\eta \in \bigwedge^\ell(V)$, by applying the result to each term.

III.11.17.21. In particular, if $\omega \in \bigwedge^k(V)$ with k *odd*, then

$$\omega \wedge \omega = (-1)^{k^2}\omega \wedge \omega = -\omega \wedge \omega$$

so $\omega \wedge \omega = 0$. But if k is even, then $\omega \wedge \omega \neq 0$ in general (if ω is not an elementary wedge). But note also that if $\omega \in \bigwedge^k(V)$ and $\eta \in \bigwedge^\ell(V)$ and $k + l > dim(V)$, then $\omega \wedge \eta \in \bigwedge^{k+\ell}(V) = \{0\}$, so $\omega \wedge \eta = 0$.

III.11.17.22. We have defined one of the standard versions of the wedge product, the one expressed most simply and nicely with respect to the basis of basic elementary wedges. To motivate the other standard normalization, we compute the wedge product we have defined for the standard alternating basis vectors: if $\{i_1, \dots, i_k\} \cap \{j_1, \dots, j_\ell\} \neq \emptyset$, then $x_{(i_1,\dots,i_k)} \wedge x_{(j_1,\dots,j_\ell)} = 0$, but if $\{i_1, \dots, i_k\}$ and $\{j_1, \dots, j_\ell\}$ are disjoint, we have

$$x_{(i_1,\dots,i_k)} \wedge x_{(j_1,\dots,j_\ell)} = k!(x_{i_1} \wedge \cdots \wedge x_{i_k}) \wedge \ell!(x_{j_1} \wedge \cdots \wedge x_{j_\ell}) = \frac{k!\ell!}{(k+\ell)!} x_{(i_1,\dots,i_k,j_1,\dots,j_\ell)} \ .$$

To eliminate the somewhat ugly constant, we normalize differently:

III.11.17.23. Definition. Let V be a vector space over $\mathbb{F}$. If $\omega \in \bigwedge^k(V)$ and $\eta \in \bigwedge^\ell(V)$, define

$$\omega \curlywedge \eta := \frac{(k+\ell)!}{k!\ell!}\omega \wedge \eta = \binom{k+\ell}{k}\omega \wedge \eta = \binom{k+\ell}{\ell}\omega \wedge \eta \ .$$

III.11.17.24. It is obvious that $\curlywedge$ is well defined, since for fixed k and ℓ it is just a constant multiple of $\wedge$. A straightforward calculation (Exercise III.11.19.13.) shows that $\curlywedge$ is associative. The product $\curlywedge$ is not so nice when written using the basic elementary wedges as basis, but is much nicer when the standard alternating basis is used: if $\{i_1, \dots, i_k\}$ and $\{j_1, \dots, j_\ell\}$ are disjoint, we have

$$x_{(i_1,\dots,i_k)} \curlywedge x_{(j_1,\dots,j_\ell)} = x_{(i_1,\dots,i_k,j_1,\dots,j_\ell)} \ .$$

III.11.17.25. The notation $\wedge$ for the wedge product is consistent with the use of $\wedge$ in the definition of elementary wedges. We can get consistent notation for $\curlywedge$ also: if $v, w \in \bigwedge^1(V) = V$, then

$$v \curlywedge w = \frac{2!}{1!1!}v \wedge w = 2(v \wedge w) = v \wedge w - w \wedge v$$

and, if we define

$$v_1 \curlywedge \cdots \curlywedge v_k = k!(v_1 \wedge \cdots \wedge v_k) = \sum_{\sigma \in S_k} \mathrm{sgn}(\sigma)[v_{\sigma(1)} \wedge \cdots \wedge v_{\sigma(k)}]$$

for $v_1, \dots, v_k \in V$, then we have

$$(v_1 \curlywedge \cdots \curlywedge v_k) \curlywedge (w_1 \curlywedge \cdots \curlywedge w_\ell) = v_1 \curlywedge \cdots \curlywedge v_k \curlywedge w_1 \curlywedge \cdots \curlywedge w_\ell$$

for $v_1, \dots, v_k, w_1, \dots, w_\ell \in V$. With this notation, if $\{x_1, \dots, x_n\}$ is an ordered basis for V, for $1 \leq j_1 < j_2 < \cdots < j_k \leq n$ we have

$$x_{(j_1,\dots,j_k)} = x_{j_1} \curlywedge \cdots \curlywedge x_{j_k} \ .$$

The Exterior or Grassmann Algebra of a Vector Space

III.11.17.26. DEFINITION. Let V be a vector space over $\mathbb{F}$. Define

$$\bigwedge(V) = \bigoplus_{k=0}^{\infty} \bigwedge^k(V)$$

with coordinatewise addition and wedge product. Then $\bigwedge(V)$ is a graded associative algebra called the *exterior algebra* or *Grassmann algebra* of V.

Some authors define the exterior or Grassmann algebra of V to be $\bigwedge(V^*)$.

III.11.17.27. One could use $\curlywedge$ in place of $\wedge$. The map from $\bigwedge(V)$ to $\bigwedge(V)$ which multiplies elements of $\bigwedge^k(V)$ by $k!$ intertwines $\curlywedge$ and $\wedge$, so $(\bigwedge(V), \curlywedge)$ is isomorphic as a graded algebra to $(\bigwedge(V), \wedge)$.

III.11.17.28. Another way to define the exterior algebra of V is to note that, setting $\mathcal{R}^0(V) = \{0\}$ and

$$\mathcal{R}(V) = \bigoplus_{k=0}^{\infty} \mathcal{R}^k(V)$$

we have from III.11.17.18. that $\mathcal{R}(V)$ is a two-sided ideal in $\mathcal{T}(V)$ and we may define

$$\bigwedge(V) = \mathcal{T}(V)/\mathcal{R}(V)$$

as a graded algebra. ($\mathcal{R}(V)$ is the ideal of $\mathcal{T}(V)$ generated by $\{v \otimes v : v \in \mathcal{T}^1(V)\} \subseteq \mathcal{T}^2(V)$.)

III.11.17.29. If V is finite-dimensional with $dim(V) = n$, then since $\bigwedge^k(V) = \{0\}$ for $k > n$ and $dim\left(\bigwedge^k(V)\right) = \binom{n}{k}$ for $0 \leq k \leq n$, we have that $\bigwedge(V)$ is finite-dimensional and

$$dim\left(\bigwedge(V)\right) = \sum_{k=0}^{n} \binom{n}{k} = 2^n \ .$$

Universal Property of the Exterior Algebra

III.11.17.30. DEFINITION. A graded algebra $A = \oplus_k A_k$ is *graded commutative* if, whenever $x \in A_k$, $y \in A_\ell$, then $xy = (-1)^{k\ell} yx$.

This terminology can be slightly confusing. Note that a graded commutative (graded) algebra is not generally a commutative graded algebre (i.e. a graded algebra which is a commutative algebra). In fact, the multiplication in a graded commutative algebra is not commutative except in degenerate cases.

III.11.17.31. The prototype example of a graded commutative graded algebra is $\bigwedge(V)$ for a vector space V. This algebra is generated as an algebra by 1 and the elements of $\bigwedge^1(V) = V$.

III.11.17.32. THEOREM. Let V be a vector space over $\mathbb{F}$. If A is any graded commutative graded algebra over $\mathbb{F}$ and T is any linear map from V to A_1, then T extends uniquely to a unital graded algebra-homomorphism from $\bigwedge(V)$ to A.

This is an almost immediate consequence of III.11.16.9.. T extends to a (unique) unital graded algebra homomorphism $\phi : \mathcal{T}(V) \to A$. If $v \in \mathcal{T}^1(V) = V$, then $\phi(v \otimes v) = \phi(v)^2 = 0$ since A is graded commutative. Thus the ideal $\mathcal{R}(V)$ is contained in the kernel of ϕ and ϕ drops to a unital graded homomorphism from $\mathcal{T}(V)/\mathcal{R}(V) = \bigwedge(V)$ to A.

As with the tensor algebra construction, we conclude that the exterior algebra construction is functorial and natural:

III.11.17.33. COROLLARY. Let V and W be vector spaces over $\mathbb{F}$. Any linear map $T : V \to W$ extends uniquely to a unital graded algebra-homomorphism $\bigwedge T$ from $\bigwedge(V)$ to $\bigwedge(W)$ (write $\bigwedge^k T$ for the map from $\bigwedge^k(V)$ to $\bigwedge^k(W)$). On elementary tensors we have

$$[\bigwedge^k T](v_1 \wedge \cdots \wedge v_k) = (Tv_1) \wedge \cdots \wedge (Tv_k) \ .$$

III.11.17.34. COROLLARY. Let V, W, and Z be vector spaces over $\mathbb{F}$. If $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, Z)$, then $\bigwedge(S \circ T) = (\bigwedge S) \circ (\bigwedge T)$.

Exterior powers of an operator also preserve many properties:

III.11.17.35. PROPOSITION. Let V and W be vector spaces over $\mathbb{F}$, and $T : V \to W$ a linear transformation. Then

(i) If T is injective, so is $\bigwedge^k T$ for each k.

(ii) If T is surjective, so is $\bigwedge^k T$ for each k..

(i) If $V = W$ and T is idempotent, so is $\bigwedge^k T$ for each k.

The proof is almost identical to the proof of III.11.16.12..

Determinants

III.11.17.36. Let V be an n-dimensional vector space over $\mathbb{F}$. Then $\bigwedge^n(V)$ is one-dimensional. If $V = \mathbb{F}^n$, so that V has a canonical ordered basis $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$, then $\bigwedge^n(V)$ is spanned by $\boldsymbol{e}_1 \wedge \cdots \wedge \boldsymbol{e}_n$, and hence can be canonically identified with $\mathbb{F}$ by calling this element 1.

If we pass to the dual, let $\{\boldsymbol{e}^1, \ldots, \boldsymbol{e}^n\}$ be the dual basis of $V^* = (\mathbb{F}^n)^*$; then $\bigwedge^n(V^*) \cong (\bigwedge^n(V))^*$ is one-dimensional, spanned by $\boldsymbol{e}^1 \wedge \cdots \wedge \boldsymbol{e}^n$, which can be identified with $1 \in \mathbb{F}$. Identifying $(\bigwedge^n(V))^*$ with the set of alternating n-linear functions from V to $\mathbb{F}$, we obtain:

III.11.17.37. THEOREM. Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Let $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ be the standard basis of $\mathbb{F}^n$. Then there is a unique alternating n-linear function ϕ from $\mathbb{F}^n \times \cdots \times \mathbb{F}^n$ (n factors) to $\mathbb{F}$ with $\phi(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) = 1$.

There is a direct proof of this result which works also in the characteristic p case, and even for commutative unital rings (III.11.19.14.).

Since the determinant on $M_n(\mathbb{F})$ is an alternating n-linear function on the rows of the matrix with $det(I_n) = 1$, we obtain:

III.11.17.38. COROLLARY. Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Then the determinant is the unique function from $M_n(\mathbb{F})$ to $\mathbb{F}$ which is n-linear and alternating on the rows, taking the value 1 on I_n ["rows" can be replaced by "columns"].

The Symmetric Algebra of a Vector Space

There is a complementary construction to the exterior algebra construction, the symmetric algebre construction. Since we will not use this construction in the sequel, we give only a bare outline.

III.11.17.39. Fix a vector space V over $\mathbb{F}$ and $k \in \mathbb{N}$. Define a map $Sym_k : \mathcal{T}^k(V) \to \mathcal{T}^k(V)$ by

$$Sym_k(v_1 \otimes \cdots \otimes v_k) = \frac{1}{k!} \sum_{\sigma \in S_k} v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(k)}$$

extended by linearity. Then Sym_k is an idempotent linear map on $\mathcal{T}^k(V)$; let $\mathcal{S}^k(V)$ be the range. If $k \geq 2$ we have $\mathcal{A}^k(V) \subseteq \mathcal{N}(Sym_k)$ and $\mathcal{S}^k(V) \subseteq \mathcal{R}^k(V)$ (both containments are proper unless $k = 2$). Note that Sym_1 is the identity map on $\mathcal{T}^1(V) = V$, so $\mathcal{S}^1(V) = V$. Define $\mathcal{S}^0(V) = \mathbb{F}$.

Unlike for $\bigwedge^k(V)$, $\mathcal{S}^k(V) \neq \{0\}$ for all k (if $V \neq \{0\}$), since Sym_k is never the zero operator: any tensor of the form $v \otimes \cdots \otimes v$ is in $\mathcal{S}^k(V)$. In fact, if $dim(V) = n$, then $dim\left(\mathcal{S}^k(V)\right) = \binom{n+k-1}{k}$ for all k.

III.11.17.40. Set $\mathcal{S}(V) = \bigoplus_{k=0}^{\infty} \mathcal{S}^k(V)$. Just as with $\bigwedge(V)$, there is a product on $\mathcal{S}(V)$ making it a unital graded algebra. This multiplication is commutative. $\mathcal{S}(V)$ is called the *symmetric algebra* of V.

III.11.17.41. $\mathcal{S}(V)$ can also be defined as a quotient: if V is a vector space over $\mathbb{F}$, let $\mathcal{J}(V)$ be the ideal of $\mathcal{T}(V)$ generated by $\{(v \otimes w) - (w \otimes v) : v, w \in \mathcal{T}^1(V) = V\} \subseteq \mathcal{T}^2(V)$; then $\mathcal{S}(V) \cong \mathcal{T}(V)/\mathcal{J}(V)$.

As with the tensor and exterior algebras, the symmetric algebra has a universal property, which implies the symmetric algebra construction is functorial and natural:

III.11.17.42. THEOREM. Let V be a vector space over $\mathbb{F}$. If A is any commutative graded algebra over $\mathbb{F}$ and T is any linear map from V to A_1, then T extends uniquely to a unital graded algebra-homomorphism from $\mathcal{S}(V)$ to A.

III.11.17.43. Corollary. Let V and W be vector spaces over $\mathbb{F}$. Any linear map $T : V \to W$ extends uniquely to a unital graded algebra-homomorphism ΞT from $\mathcal{S}(V)$ to $\mathcal{S}(W)$.

III.11.17.44. Corollary. Let V, W, and Z be vector spaces over $\mathbb{F}$. If $T \in \mathcal{L}(V, W)$ and $S \in \mathcal{L}(W, Z)$, then $\Xi(S \circ T) = (\Xi S) \circ (\Xi T)$.

III.11.18. Mixed Tensors and General Tensor Algebra

Tensor algebra is traditionally primarily an exercise in index manipulation. The theory behind it justifies reduction of many aspects and applications to mechanical symbol manipulation (this can be said about other parts of mathematics too, but it is especially evident here).

Summation Convention

III.11.18.1. Many references in this subject use the *summation convention*, introduced by Einstein. Although I have never liked this convention, readers of the literature need to be aware of it and understand it (which is not difficult). It does streamline formulas and sometimes makes them appear more natural, but also causes problems.

III.11.18.2. When a summation expression like

$$\sum_{k=1}^{n} x_k y_k$$

has a repeated index which is summed over, the summation is omitted notationally, writing simply

$$x_k y_k \ .$$

The convention is most often used when both subscripts and superscripts are employed: an index which appears twice, once as a subscript and once as a superscript, is regarded as a dummy index which is summed over.

III.11.18.3. As a simple example illustrating the notation and how it can make formulas appear simpler and more natural, let A be an $m \times n$ matrix and B an $n \times p$ matrix, and $C = AB$. Write the (i, j) element of A as a^i_j, etc. Then the definition of matrix multiplication becomes

$$c^i_j = a^i_k b^k_j \ .$$

Here i is an index between 1 and m, j an index between 1 and p, and k is a dummy index which is summed from 1 to n.

As another example, if $y = f(u_1, \ldots, u_n)$ and $u_j = g_j(x_1, \ldots, x_m)$, the Chain Rule can be written

$$\frac{\partial y}{\partial x_k} = \frac{\partial y}{\partial u_j} \frac{\partial u_j}{\partial x_k}$$

looking more like the one-variable Chain Rule and suggesting that the ∂u_j can just be "canceled." This illustrates how the summation convention can sometimes mislead a user into trying unjustified algebraic manipulations.

III.11.18.4. Summation notation can be sometimes confusing and occasionally ambiguous. It is easy to forget that a summation is notationally omitted, or to be unsure when a summation is implied and when it isn't. It is telling that [BG80], which uses the convention, devotes a full page to warnings about its pitfalls, how certain letters must be designated as dummy variables for the convention, and how to indicate when the convention is not being used.

III.11.18.5. The summation convention is akin to the notational conventions used by logicians to write formulas and statements without using parentheses. In either case, experts undoubtedly develop a feel and facility for the notation and find it convenient and natural. But I have not, and will not use either convention in my book; I much prefer to stick to standard mathematical notation, which gives clear and unambiguous, if somewhat more complicated, formulas.

III.11.19. Exercises

III.11.19.1. Show that the tensor product of the 0 vector space (module) with any other vector space (module) is the zero vector space (module).

III.11.19.2. Let $\mathbb{F}$ be a field, regarded as a vector space over itself, and V a vector space over $\mathbb{F}$. Prove directly from the definition of tensor product (i.e. without using III.11.14.13.) that $\mathbb{F} \otimes V$ is naturally isomorphic to V. Generalize to modules over a ring.

III.11.19.3. Use III.11.19.2. and III.11.14.28. to show that if W is an n-dimensional vector space over $\mathbb{F}$, and V is any vector space over $\mathbb{F}$, then $V \otimes W \cong V \oplus \cdots \oplus V$ (n summands). Use this to give an alternate proof of III.11.14.14.. [It can be shown without AC that every finite-dimensional vector space has a basis.]

III.11.19.4. Let $\{V_j, \phi_{jk}\}$ be a directed system of vector spaces over $\mathbb{F}$, and let W be a vector space over $\mathbb{F}$. Then $\{V_j \otimes W, \phi_{jk} \odot 1\}$ is a directed system of vector spaces (where 1 is the identity map on W). Show that there is a natural isomorphism

$$\varinjlim(V_j \otimes W, \phi_{jk} \odot 1) \cong \left[\varinjlim(V_j, \phi_{jk})\right] \otimes W \ .$$

Generalize to modules. (This result as written uses the AC to insure that the direct limits exist; but the statement can be modified to say that *if* the direct limit of the V_j exists, then the direct limit of the $V_j \otimes W$ also exists and the isomorphism holds. Thus the application in III.11.19.5.(c) does not require the AC.)

III.11.19.5. Give an alternate proof of III.11.14.29. not requiring the AC, i.e. that every vector space is a flat module (III.11.15.20.), as follows.

(a) Use III.11.19.3. to show that a finite-dimensional vector space is flat. (More generally, any free module over a ring is flat.)

(b) Use III.11.19.4. to show that a direct limit of flat vector spaces (or flat modules over a ring) is flat.

(c) Conclude that every vector space is flat. [Any vector space is the direct limit of its finite-dimensional subspaces, directed by inclusion.]

It cannot be proved without AC that every vector space is a free module (i.e. has a basis), or even a projective module.

III.11.19.6. [Extension of Scalars] Let S be a (unital) ring, and R a (unital) subring of S. Let M be a left R-module. Regard S as a right R-module (). Then ${}_SM = S \otimes_R M$ is a left S-module, called the *extension of M to S*. Similarly, if N is a right R-module, regard S as a left R-module; then $N_S = N \otimes_R S$ is a right S-module. As a special case, let $\mathbb{K}$ be a field, and $\mathbb{F}$ a subfield of $\mathbb{K}$. Let V be a vector space over $\mathbb{F}$. Regard $\mathbb{K}$ as a vector space over $\mathbb{F}$. Then ${}_{\mathbb{K}}V = V_{\mathbb{K}} = \mathbb{K} \otimes_{\mathbb{F}} V$ is a vector space over $\mathbb{K}$.

(a) Show that $(\mathbb{R}^n)_{\mathbb{C}}$ is naturally isomorphic to $\mathbb{C}^n$.

(b) More generally, if $\mathbb{F}$ is a subfield of $\mathbb{K}$ and V is a finite-dimensional vector space over $\mathbb{F}$ with $dim_{\mathbb{F}}(V) = n$, then $V_{\mathbb{K}}$ is a finite-dimensional vector space over $\mathbb{K}$ with $dim_{\mathbb{K}}(V_{\mathbb{K}}) = n$.

III.11.19.7. Let R be a ring.

(a) Define a new ring $R^\dagger$ as follows. Let $R^\dagger = R \oplus \mathbb{Z}$ as an additive group. Define a multiplication on $R^\dagger$ by

$$(x, m)(y, n) = (xy + nx + my, mn)$$

(to motivate this, think of (x, m) as $x + m1$). Show that $R^\dagger$ is a unital ring, and that R is isomorphic to the subring $\{(x, 0) : x \in \mathbb{R}\}$.

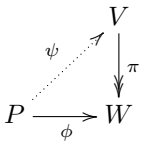
(b) If R is unital, show that $R^\dagger \cong R \oplus \mathbb{Z}$ as rings. (**Caution:** the decomposition is not the one in the definition of $R^\dagger$!)

(c) Show that if M is a (left or right) R-module, the R-action can be extended to a unital $R^\dagger$-action in a unique way making M an $R^\dagger$-module.

III.11.19.8. Let R be a unital commutative ring, and form the polynomial ring $R[X]$ in one variable. If M is a (left) R-module and T is a fixed element of $\mathcal{L}(M)$, define an action of $R[X]$ on M by letting the constants act via the given R-module action, and $X \cdot v = T(v)$ for $v \in M$. Show that this extends uniquely to an action making M a (left) $R[X]$-module.

This construction is especially interesting if R is a field $\mathbb{F}$, so M is a vector space over $\mathbb{F}$. This shows that module theory comes into ordinary linear algebra in a natural way in the analysis of linear operators. In this case, $\mathbb{F}[X]$ is a Principal Ideal Domain, and there is a nice theory of finitely generated modules over PIDs which can be used to obtain properties of operators. In particular, if $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$, this leads to canonical forms for linear operators on finite-dimensional vector spaces.

III.11.19.9. **Projective Modules.** Let R be a (unital) ring. A left R-module P is *projective* if, whenever V, W are left R-modules, π is an R-linear map from V onto W, and $\phi \in \mathcal{L}(P, W)$, there is a $\psi \in \mathcal{L}(P, V)$ with $\phi = \pi \circ \psi$, i.e. ϕ *lifts* to V (P is a projective object in the category of left R-modules). Pictorially:



(a) Show that a free module is projective. [Use the module version of III.11.3.20..]

(b) Show that a direct summand in a projective module is projective. In particular, a direct summand in a free module is projective.

(c) If P is a projective module, V any module, and π a surjective R-linear map from V to P, then there is a (necessarily injective) $\iota \in \mathcal{L}(P, V)$ with $\pi \circ \iota$ the identity map on P. [Let $W = P$ and ϕ the identity map.]

(d) Show the converse of (b): every projective module is a direct summand of a free module. [Every module is a quotient of a free module.] Thus a module is projective if and only if it is a direct summand of a free module. A module is finitely generated and projective if and only if it is a direct summand in a free module with a finite basis.

(e) Show that a left R-module P is projective if and only if every short exact sequence

$$0 \longrightarrow Y \longrightarrow X \longrightarrow P \longrightarrow 0$$

of left R-modules splits.

III.11.19.10. Let R be a unital ring. If F is a free left R-module with a finite basis $\{x_1, \ldots, x_n\}$, so that $F \cong R^n$, we would like to say that F has *dimension* n.

(a) If R is commutative, show that dimension is well defined for free R-modules with finite bases. [Suppose $R^n \cong R^m$ with $n \neq m$. Then there are matrices A and B over R with A $m \times n$ and B $n \times m$ with $AB = I_m$ and $BA = I_n$. Fill out A and B to square matrices by adding rows or columns of 0s, and apply III.11.10.53. to obtain a contradiction.]

(b) Show that dimension is not necessarily well defined if R is noncommutative. [Try the case $R = \mathcal{L}(V)$ for V an infinite-dimensional vector space.]

III.11.19.11. Let V be a finite-dimensional vector space over $\mathbb{F}$, and $k \in \mathbb{N}$. Show that the isomorphism $\Phi : \mathcal{T}^k(V^*) \to (\mathcal{T}^k(V))^*$ (III.11.14.23.) drops to an isomorphism from $\bigwedge^k(V^*)$ to $(\bigwedge^k(V))^*$. [Note that these spaces have the same dimension, so it suffices to show the map is well defined and injective. Use $\mathcal{A}^k(V^*)$.]

III.11.19.12. Let $\mathbb{F}$ be a field. Regarding $\mathbb{F}$ as a one-dimensional vector space over itself, show that $\mathcal{T}(\mathbb{F}) \cong \mathbb{F}[X]$, the polynomial ring over $\mathbb{F}$, and that $\bigwedge(\mathbb{F}) \cong \mathbb{F}[X]/(X^2)$.

III.11.19.13. Let V be a vector space over $\mathbb{F}$.

(a) Show that $\curlywedge$ is associative and that $(\bigwedge(V), \curlywedge)$ is a graded algebra isomorphic to $(\bigwedge(V), \wedge)$.

(b) More generally, let ψ be a function from $\mathbb{N} \cup \{0\}$ to $\mathbb{F}^\times = \mathbb{F} \setminus \{0\}$ with $\psi(0) = \psi(1) = 1$, and for each k, ℓ define $\wedge_\psi : \bigwedge^k(V) \times \bigwedge^\ell(V) \to \bigwedge^{k+\ell}(V)$ by

$$(v_1 \wedge \cdots \wedge v_k) \wedge_\psi (w_1 \wedge \cdots \wedge w_\ell) = \frac{\psi(k+\ell)}{\psi(k)\psi(\ell)} v_1 \wedge \cdots \wedge v_k \wedge w_1 \wedge \cdots \wedge w_\ell \ .$$

Show that $\wedge_\psi$ is well defined and associative, and that $(\bigwedge(V), \wedge_\phi)$ is a graded algebra isomorphic to $(\bigwedge(V), \wedge)$. Note that $\curlywedge = \wedge_\psi$, where $\psi(k) = k!$. If $\mathbb{F} = \mathbb{R}$, the function $\psi(k) = \sqrt{k!}$ is especially interesting, cf. IX.4.1.10..

(c) ([AMR88], attributed to P. Chernoff and J. Robbin) Let $c : (\mathbb{N} \cup \{0\}) \times (\mathbb{N} \cup \{0\}) \to \mathbb{F}^\times$ be a function with $c(0, k) = c(k, 0) = 1$ for all k such that the operation $\wedge_c : \bigwedge^k(V) \times \bigwedge^\ell(V) \to \bigwedge^{k+\ell}(V)$ defined by

$$(v_1 \wedge \cdots \wedge v_k) \wedge_c (w_1 \wedge \cdots \wedge w_\ell) = c(k, \ell) v_1 \wedge \cdots \wedge v_k \wedge w_1 \wedge \cdots \wedge w_\ell$$

makes $(\bigwedge(V), \wedge_c)$ into an associative graded algebra. Show that there is a function $\psi : \mathbb{N} \cup \{0\} \to \mathbb{F}^\times$ with $\psi(0) = \psi(1) = 1$ such that $c(k, \ell) = \frac{\psi(k+\ell)}{\psi(k)\psi(\ell)}$ for all k and ℓ (and in particular $c(k, \ell) = c(\ell, k)$ for all k and ℓ), so that $(\bigwedge(V), \wedge_c) \cong (\bigwedge(V), \wedge)$.

III.11.19.14. Prove the following theorem by induction on n:

THEOREM. Let R be a unital commutative ring and $n \in \mathbb{N}$. Let $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ be the standard basis of R^n. Then there is a unique alternating n-linear function ϕ from $R^n \times \cdots \times R^n$ (n factors) to R with $\phi(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n) = 1$.

III.12. The Banach-Tarski Paradox

"These are old fond paradoxes to make fools laugh i' the alehouse."

Shakespeare[10]

The Banach-Tarski Paradox is one of the best-known but least well-understood results in mathematics. There are several popular versions. The basic version is:

THEOREM. [BANACH-TARSKI PARADOX, BASIC VERSION] A closed ball of radius r in $\mathbb{R}^3$ can be partitioned into a finite number of disjoint subsets which can be moved and reassembled to form two disjoint closed balls of radius r.

In fact, it can be done with five pieces, one of which is a single point. The word "cut" is often used instead of "partition," but this can be misleading; see III.12.7.3..

A version which has caught the public fancy and which is stated in popular books is:

THEOREM. [BANACH-TARSKI PARADOX, POPULAR VERSION] A closed ball the size of a pea can be partitioned into a finite number of disjoint subsets which can be moved and reassembled to form a closed ball the size of the sun.

However, contrary to what is written in some books, this cannot be done with five pieces: in fact, a very large number of pieces is necessary (cf. III.12.8.2.). This result is a special case of a more general result which is even more remarkable (III.12.8.1.).

It is hard to categorize what branch of mathematics this "paradox" belongs to (it is of course not really a paradox since it is logically sound, but it seems to fly in the face of intuition). It is not really analysis, or geometry or topology; it really has to do with the *algebraic* structure of the group of rigid motions of $\mathbb{R}^3$ (specifically, that this group contains a free group on two generators).

While versions of the Paradox are connected with both Lebesgue measure and with group amenability, as well as other advanced topics, the basic Banach-Tarski Paradox is really just undergraduate mathematics, but assembled in a brilliant way.

The Banach-Tarski Paradox depends on the Axiom of Choice. Its conclusion is sufficiently unintuitive that it is often considered to be evidence that the AC must be false. However, if the AC is assumed to be false (so this paradox disappears), other conclusions can be drawn which are arguably even more unintuitive, e.g. (), (). Thus it seems questionable whether this Paradox should really be considered convincing refutation of the AC; it is actually more evidence of the AC's infinite subtlety.

The Banach-Tarski Paradox was found by S. BANACH and A. TARSKI in 1924 [**?**]. It is based in an essential way on an earlier result of HAUSDORFF (III.12.6.1.), so much so that the Paradox should properly be called the "Hausdorff-Banach-Tarski Paradox." The excellent book [Wag93], on which our exposition is largely modeled, contains far more about the topic than will be covered here.

III.12.1. Rigid Motions

III.12.1.1. DEFINITION. Fix $n \in \mathbb{N}$. A *rigid motion* of $\mathbb{R}^n$ is an orientation-preserving affine isometry of $\mathbb{R}^n$.

The rigid motions of $\mathbb{R}^n$ form a group under composition.

[10] *Othello*, Act II, Scene I.

III.12.1.2. There are two basic kinds of rigid motions:

Translations: functions of the form $T_{\boldsymbol{v}}$, where $T_{\boldsymbol{v}}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{v}$.

Special Orthogonal Transformations: linear transformations given by multiplication by an orthogonal matrix () of determinant 1. These are precisely the rigid motions fixing the origin $\mathbf{0}$.

The set of translations form a subgroup of the group of rigid motions isomorphic to the additive group $\mathbb{R}^n$, and the special orthogonal transformations a subgroup $SO(n)$.

III.12.1.3. There are affine isometries of $\mathbb{R}^n$ such as reflections across a hyperplane which do not preserve orientation and which we will not consider to be rigid motions. Thus, for example, an orthogonal transformation with determinant -1 is not a rigid motion under our definition.

III.12.1.4. If F is a rigid motion of $\mathbb{R}^n$, then $U = T_{\boldsymbol{v}} \circ F$ is a special orthogonal transformation, where $\boldsymbol{v} = -F(\mathbf{0})$; thus $F = T_{-\boldsymbol{v}} \circ U$. So every rigid motion can be written canonically as a special orthogonal transformation followed by a translation. In particular, the two subgroups together generate the group of rigid motions.

III.12.1.5. A special orthogonal transformation of $\mathbb{R}^2$ is just a rotation around the origin by some fixed angle, i.e. $SO(2)$ is an abelian group isomorphic to the group $\mathbb{R}/\mathbb{Z}$, or to the multiplicative group $\mathbb{T}$ of complex numbers of absolute value 1.

If U is a special orthogonal transformation of $\mathbb{R}^3$, then U must have 1 as an eigenvalue (this is true more generally in $\mathbb{R}^n$ for any odd n). Thus U leaves fixed some unit vector $\boldsymbol{u}$, and gives a special orthogonal transformation of the plane through the origin orthogonal to $\boldsymbol{u}$. It follows that U is a rotation by a fixed angle around the axis consisting of the line through $\mathbf{0}$ and $\boldsymbol{u}$. (The situation is not nearly so simple in $\mathbb{R}^n$ for $n > 3$.)

III.12.1.6. Here are some other facts about the group of rigid motions of $\mathbb{R}^n$. We will not need to use these facts to derive the Banach-Tarski Paradox.

(i) The group of rigid motions has a natural topology, the topology of uniform convergence on compact subsets (u.c. convergence), making it into a topological group. It is homeomorphic to $\mathbb{R}^n \times SO(n)$ as a topological space (but not isomorphic as a topological group!), and thus is locally compact and locally Euclidean (i.e. a Lie group ()). Since $\mathbb{R}^n$ and $SO(n)$ are path-connected, so is the group of rigid motions. The subgroup $\mathbb{R}^n$ of translations is normal, and the group of rigid motions is the semidirect product $\mathbb{R}^n \rtimes SO(n)$ (cf. XXIII.2.2.6.(iii)).

(ii) Every isometry of $\mathbb{R}^n$ is affine (XX.2.2.1.). The group of isometries of $\mathbb{R}^n$ is a topological group under the topology of u.c. convergence, and the group of rigid motions is precisely the connected component of the identity in this group, an open subgroup of index 2. The group of isometries is the semidirect product $\mathbb{R}^n \rtimes O(n)$.

III.12.2. Rigid Equivalence

III.12.2.1. Definition. Fix $n \in \mathbb{N}$, and let A and B be subsets of $\mathbb{R}^n$. Then A and B are *rigidly equivalent* in $\mathbb{R}^n$, written $A \approx B$, if there is a rigid motion F of $\mathbb{R}^n$ with $F(A) = B$.

III.12.2.2. Since the rigid motions of $\mathbb{R}^n$ are bijections and form a group under composition, rigid equivalence is indeed an equivalence relation on subsets of $\mathbb{R}^n$. If $A \approx B$ and one of the sets is bounded, so is the other (in fact, they have the same diameter).

Here are some fundamental examples which motivate the constructions in the paradoxes:

III.12.2.3. Examples. (i) A set can be rigidly equivalent to a proper subset of itself, even in $\mathbb{R}^1$: $\mathbb{N}$ is rigidly equivalent to $\mathbb{N} \setminus \{1\}$ under the translation T_1. This is just a geometric version of the characteristic property of an infinite (Dedekind-infinite) set that it is equipotent with a proper subset of itself.

(ii) A bounded subset of $\mathbb{R}^1$ cannot be rigidly equivalent to a proper subset of itself (Exercise III.12.9.1.). But it is possible in $\mathbb{R}^2$ by wrapping the example of (i) around the unit circle. Let

$$A = \{(\cos n, \sin n) : n \in \mathbb{N} \cup \{0\}\} .$$

Then A is a dense subset of the unit circle. Let U be rotation by angle 1 radian; then $U(A) = A \setminus \{(1, 0)\}$. It is more efficient to regard A as $\{e^{in} : n \in \mathbb{N} \cup \{0\}\}$ and taking U as multiplication by e^i. (This example works, of course, because π is irrational.)

(iii) Let C be a countable subset of the unit circle in $\mathbb{R}^2$. Choose an angle θ such that $e^{in\theta}z \notin C$ for every $z \in C$ and $n \in \mathbb{N}$ (there is such a θ, in fact many, since there are only countably many θ disqualified by the requirement). Set

$$A = \{e^{in\theta}z : n \in \mathbb{N} \cup \{0\}, z \in C\} .$$

Then A is a countable subset of the unit circle containing C. If U is rotation by θ (i.e. multiplication by $e^{i\theta}$), then $U(A) = A \setminus C$.

III.12.3. Equivalence by Dissection

III.12.4. Equidecomposability

III.12.4.1. Rigid equivalence is an equivalence relation which is, well, too rigid to be directly useful for the Banach-Tarski Paradox. For example, since a rigid motion is a homeomorphism, a closed set cannot be rigidly equivalent to a nonclosed one. But there is a more flexible equivalence relation which is just right:

III.12.4.2. Definition. Fix $n \in \mathbb{N}$. Two subsets A and B of $\mathbb{R}^n$ are *equidecomposable*, written $A \sim B$, if there is an $m \in \mathbb{N}$ and partitions $\{A_1, \ldots, A_m\}$ of A and $\{B_1, \ldots, B_m\}$ of B, each partition consisting of pairwise disjoint subsets, such that $A_j \approx B_j$ for $1 \leq j \leq m$. (The rigid motions implementing the rigid equivalences may vary for different j.)

III.12.4.3. Proposition. Fix $n \in \mathbb{N}$. Then equidecomposability is an equivalence relation on subsets of $\mathbb{R}^n$.

Proof: Equidecomposability is obviously reflexive and symmetric. To prove transitivity, suppose $A, B, C \subseteq \mathbb{R}^n$ with $A \sim B$ and $B \sim C$. Let $\{A_1, \ldots, A_m\}$ and $\{B_1, \ldots, B_m\}$ be partitions of A and B respectively, and F_j a rigid motion for each j, $1 \leq j \leq m$, with $F_j(A_j) = B_j$; and let $\{B'_1, \ldots, B'_p\}$ and $\{C_1, \ldots, C_p\}$ be

partitions of B and C respectively and G_k $(1 \le k \le p)$ rigid motions with $G_k(B'_k) = C_k$. For each j and k set

$$A_{jk} = F_j^{-1}(B_j \cap B'_k)$$
$$C_{jk} = G_k(B_j \cap B'_k) \ .$$

Then $\{A_{jk} : 1 \le j \le m, 1 \le k \le p\}$ and $\{C_{jk} : 1 \le j \le m, 1 \le k \le p\}$ are partitions of A and C, respectively, into a finite number of pairwise disjoint subsets, and $G_k \circ F_j$ is a rigid motion of $\mathbb{R}^n$ with $(G_k \circ F_j)(A_{jk}) = C_{jk}$ for each (j, k). Thus $A \sim C$.

Note that if m and p pieces are needed for the two equidecompositions, the proof shows that the composed equidecomposition can be done with mp pieces. This is not necessarily the minimum number of pieces needed.

III.12.4.4. Equidecomposablility, unlike rigid equivalence, is preserved under (finite) disjoint unions: if $\{A_1, \dots, A_m\}$ and $\{B_1, \dots, B_m\}$ are finite collections of subsets of $\mathbb{R}^n$ each of which consists of pairwise disjoint sets, and $A_j \sim B_j$ for each j, then

$$\bigcup_{j=1}^{m} A_j \sim \bigcup_{j=1}^{m} B_j \ .$$

The following crucial examples show how much more flexible equidecomposability is than rigid equivalence:

III.12.4.5. Examples. (i) If $A \approx B$, then $A \sim B$ (using $m = 1$).
(ii) Let $S^1 = \mathbb{T}$ be the unit circle in $\mathbb{R}^2$, and let C be a countable subset of S^1. Let $A = A_1$ be the countable subset defined in III.12.2.3. for which $A \approx A \setminus C$. Set $A_2 = S^1 \setminus A_1$. Then $\{A_1, A_2\}$ is a partition of S^1 into disjoint subsets, and $A_2 \approx A_2$ via the identity map, so

$$S^1 = A_1 \cup A_2 \sim (A_1 \setminus C) \cup A_2 = S^1 \setminus C \ .$$

Thus a circle is equidecomposable with a circle with any finite or countable subset removed.
(iii) With a bit of care, we can do the same for the 2-sphere S^2, the unit sphere in $\mathbb{R}^3$; this example will be important for the Paradox. Let C be a countable subset of S^2. First choose a line L through the origin not passing through any point of C (there are uncountably many lines through $\mathbf{0}$, and only countably many of them pass through a point of C). Then choose an angle θ such that rotation around the axis L by $n\theta$ does not send any point of C to itself or another point of C, for any $n \in \mathbb{N}$ (this condition only disqualifies countably many θ once L is fixed, so there is even a dense set of θ not disqualified). Let U be the special orthogonal transformation of $\mathbb{R}^3$ given by rotation by θ around the axis L, and set

$$A = \bigcup_{n=0}^{\infty} U^n(C) \ .$$

Then $A_1 = A$ is a countable subset of S^2, and $U(A) = A \setminus C$. If $A_2 = S^2 \setminus A$, then

$$S^2 = A_1 \cup A_2 \sim (A_1 \setminus C) \cup A_2 = S^2 \setminus C$$

so S^2 is equidecomposable with S^2 with any countable subset removed.

One of the most remarkable properties of equidecomposability is that it satisfies the following analog of the Schröder-Bernstein Theorem of set theory ():

III.12.4.6. THEOREM. [SCHRÖDER-BERNSTEIN THEOREM FOR EQUIDECOMPOSABILITY] Let A and B be subsets of $\mathbb{R}^n$. If there are subsets $A' \subseteq A$ and $B' \subseteq B$ with $A \sim B'$ and $B \sim A'$, then $A \sim B$.

PROOF: The proof is nearly identical to BERNSTEIN'S proof of the Schröder-Bernstein Theorem (). Suppose $\{A_1, \ldots, A_m\}$ and $\{B'_1, \ldots, B'_m\}$ are partitions of A and B', and F_j a rigid motion with $F_j(A_j) = B'_j$ for $1 \leq j \leq m$, and that $\{A'_1, \ldots, A'_p\}$ and $\{B_1, \ldots, B_p\}$ are partitions of A' and B, and G_k is a rigid motion with $G_k(B_k) = A'_k$ for $1 \leq k \leq p$. Set $X_1 = A \setminus A'$ and $X_{1,j,0} = A_j \cap X_1$ for $1 \leq j \leq m$. Similarly, set $Y_1 = B \setminus B'$ and $Y_{1,0,k} = B_k \cap Y_1$. For $1 \leq j \leq m$ set $Y_{2,j,0} = F_j(X_{1,j,0})$ and $Y_2 = \cup_{j=1}^m Y_{2,j,0}$, and for $1 \leq k \leq p$ set $X_{2,0,k} = G_k(Y_{1,0,k})$ and $X_2 = \cup_{k=1}^p X_{2,0,k}$. For $1 \leq j \leq m$ set $X_{2,j,k} = X_{2,0,k} \cap A_j$ and $X_{2,j,0} = \cup_{k=1}^p X_{2,j,k}$, and similarly set $Y_{2,j,k} = Y_{2,j,0} \cap B_k$ for $1 \leq k \leq p$ and $Y_{2,0,k} = \cup_{j=1}^m Y_{2,j,k}$. Repeat the construction inductively to jet $X_{i,j,k}$, $X_{i,j,0} = \cup_{k=1}^p X_{i,j,k}$, $X_{i,0,k} = \cup_{j=1}^m X_{i,j,k}$,

$$X_i = \bigcup_{j=1}^{p} X_{i,j,0} = \bigcup_{k=1}^{p} X_{i,0,k} = \bigcup_{j=1}^{m} \bigcup_{k=1}^{p} X_{i,j,k}$$

and $Y_{i,j,k}$, $Y_{i,j,0} = \cup_{k=1}^p Y_{i,j,k}$, $Y_{i,0,k} = \cup_{j=1}^m Y_{i,j,k}$,

$$Y_i = \bigcup_{j=1}^{p} Y_{i,j,0} = \bigcup_{k=1}^{p} Y_{i,0,k} = \bigcup_{j=1}^{m} \bigcup_{k=1}^{p} Y_{i,j,k}$$

such that $Y_{i+1,j,0} = F_j(X_{i,j,0})$ for each j and $X_{i+1,0,k} = G_k(Y_{i,0,k})$ for each K.

Set $X_\infty = A \setminus \cup_{i=1}^\infty X_i$ and $Y_\infty = B \setminus \cup_{i=1}^\infty Y_i$, and set $X_{\infty,j} = X_\infty \cap A_j$ and $Y_{\infty,j} = F_j(X_{\infty,j}$. Since $Y_{\infty,j} \subseteq B'_j$, the $Y_{\infty,j}$ are pairwise disjoint. We claim $\cup_{j=1}^m Y_{\infty,j} = Y_\infty$. First suppose $y \in Y_{\infty,j}$, i.e. $y = F_j(x)$ for some $x \in X_{\infty,j}$. Since $y \in B'_j$, $y \in B'$ so $y \notin Y_1 = B \setminus B'$. If $y \in Y_{i+1}$ for some i, then $y = F_j(x)$ for some $z \in X_{i,j,0}$, then $z \neq x$ but $y = F_j(x) = F_j(z)$, a contradiction since F_j is injective. Thus $y \in Y_\infty$. Conversely, if $y \in Y_\infty$, then $y \notin Y_1$, so $y \in B'$, hence $y \in B'_j$ for some j, i.e. $y = F_j(x)$ for some $x \in A_j$. If $x \in X_i$ for some i, then $x \in X_{i,j,0}$, so $y \in Y_{i+1,j,0} \subseteq Y_{i+1}$, contradicting that $y \in Y_\infty$. Thus $\{Y_{\infty,1}, \ldots, Y_{\infty,m}\}$ give a partition of Y_∞.

Now define, for $1 \leq j \leq m$, $1 \leq k \leq n$,

$$C_{1,j} = \bigcup_{i \text{ odd}} X_{i,j,0} \cup X_{\infty,j} \subseteq A$$

$$C_{2,k} = \bigcup_{i \text{ even}} X_{i,0,k} \subseteq A$$

$$D_{1,j} = \bigcup_{i \text{ even}} Y_{i,j,0} \cup Y_{\infty,j} \subseteq B$$

$$D_{2,k} = \bigcup_{i \text{ odd}} Y_{i,0,k} \subseteq B$$

Then $\{C_{1,1},\dots,C_{1,m},C_{2,1},\dots,C_{2,p}\}$ is a partition of A, $\{D_{1,1},\dots,D_{1,m},D_{2,1},\dots,D_{2,p}\}$ is a partition of B, $F_j(C_{1,j}) = D_{1,j}$ for each j, and $G_k^{-1}(C_{2,k}) = D_{2,k}$ for each k. Therefore $A \sim B$. 👍

Note that if m pieces are required for the equidecomposition of A and B', and p pieces required for the equidecomposition of A' and B, then the proof shows that the equidecomposition of A and B can be done with $m+p$ pieces (this is not necessarily the minimum number required).

III.12.5. Free Groups and Paradoxical Decompositions

For the basics of group theory, see ().

III.12.5.1. If a and b are two symbols, we can form "words" by concatenating a's and b's, e.g. aaa, ba, $aababbba$, etc. We can "multiply" words by concatenation:

$$(aababa)(abbbabba) = aababaabbbabba$$

etc. This is a binary operation on words, which is easily seen to be associative. If we allow the "empty word" (usually denoted e), it is an identity for this operation. This set of words with concatenation is called the *free semigroup* on the generators $\{a,b\}$.

III.12.5.2. We want to make a group in the same way, but there are some complications. First, we must have inverses a^{-1} and b^{-1} for a and b, with $a^{-1}a = aa^{-1} = b^{-1}b = bb^{-1} = e$. Thus we will take words in a, b, a^{-1}, b^{-1}. But distinct words can represent the same group element: whenever the consecutive symbols $a^{-1}a$, aa^{-1}, $b^{-1}b$, or bb^{-1} appear in a word, the word can be *contracted* by eliminating the pair. This can be done repeatedly if necessary, eventually reaching a *reduced word* in which no such consecutive pair appears. Since contraction of a word can in general be done in many ways, and the concatenation of two reduced words is not necessarily reduced, the following facts are not entirely obvious, and take a little work to prove; see e.g. ().

III.12.5.3. Proposition. (i) Every word x reduces to a *unique* reduced word $r(x)$.

(ii) If x and y are words, then $r(xy) = r(r(x)r(y))$.

(iii) If x, y, z are words, then $r(((r(x)r(y))r(z)) = r(r(x)(r(y)r(z)))$.

III.12.5.4. Thus we have a well-defined associative binary operation on reduced words by $x \cdot y = r(xy)$, making the set of reduced words, including the empty word e, into a semigroup with identity e. (We will write the operation without the dot.)

In fact, it is a group: to make the inverse of a (reduced) word, take the inverse of each letter and write them in the reverse order, e.g.

$$(ab^{-1}aba^{-1}b^{-1})^{-1} = bab^{-1}a^{-1}ba^{-1}$$

since it is easily seen that

$$ab^{-1}aba^{-1}b^{-1}bab^{-1}a^{-1}ba^{-1}$$

reduces to the empty word, starting in the middle.

III.12.5.5. DEFINITION. The group of all reduced words in a, b, a^{-1}, b^{-1} is called the *free group on generators* $\{a, b\}$, denoted $\mathbb{F}_2$.

III.12.5.6. A similar definition can be made of a free group on any set of generators. The group clearly depends up to isomorphism only on the number of generators, not on the letters used; denote the free group on n generators by $\mathbb{F}_n$. (It is not obvious, but true, that free groups on different numbers of generators are not isomorphic.) We can even define a free group $\mathbb{F}_1$ on one generator, and $\mathbb{F}_1$ is naturally isomorphic to the additive group $\mathbb{Z}$.

III.12.5.7. Free groups (on more than one generator) are highly nonabelian. $\mathbb{F}_n$ has a universal property: if $\mathcal{G}$ is any group and $x_1, \ldots, x_n \in \mathcal{G}$, there is a unique group-homomorphism $\phi : \mathbb{F}_n \to \mathcal{G}$ sending the k'th generator to x_k for $1 \leq k \leq n$.

Paradoxical Subdivisions of Groups

III.12.5.8. If $\mathcal{G}$ is a group and $\mathcal{A} \subseteq \mathcal{G}$, we define the *left translation* of $\mathcal{A}$ by $g \in \mathcal{G}$ to be $g\mathcal{A} = \{gx : x \in \mathcal{A}\}$. Right translations can be defined similarly, but in this section we will work only with left translations.

III.12.5.9. The fundamental "paradox" about group translations is that a translation of certain subsets of a free group (and many similar groups) become much "larger." Specifically, the group can be decomposed into a finite number of pieces which can be divided into two sets, such that translates of the pieces in each set cover the whole group. This somewhat unintuitive phenomenon, which cannot happen in "nice" groups, is the basis for the set paradoxes.

III.12.5.10. DEFINITION. Let $\mathcal{G}$ be a group. An (m, n)-*paradoxical decomposition* of $\mathcal{G}$ is a collection $\{\mathcal{A}_1, \ldots, \mathcal{A}_m; \mathcal{B}_1, \ldots, \mathcal{B}_n\}$ of pairwise disjoint subsets of $\mathcal{G}$ and elements $g_1, \ldots, g_m, h_1, \ldots, h_n$ of elements of $\mathcal{G}$ such that

$$\bigcup_{j=1}^{m} \mathcal{A}_j \cup \bigcup_{k=1}^{n} \mathcal{B}_k = \mathcal{G}$$

$$\{g_1\mathcal{A}_1, \ldots, g_m\mathcal{A}_m\} \text{ and } \{h_1\mathcal{B}_1, \ldots, h_n\mathcal{B}_n\} \text{ are pairwise disjoint}$$

$$\bigcup_{j=1}^{m} g_j\mathcal{A}_j = \bigcup_{k=1}^{n} h_k\mathcal{B}_k = \mathcal{G} \ .$$

The group $\mathcal{G}$ is *paradoxical* if it has an (m, n)-paradoxical decomposition for some (m, n).

III.12.5.11. The group $\mathbb{F}_2$ has a $(2, 2)$-paradoxical decomposition. A simple decomposition which almost works is to take $\mathcal{A}_1$ and $\mathcal{A}_2$ to be the sets of reduced words beginning with a and a^{-1} respectively, and $\mathcal{B}_1$ and $\mathcal{B}_2$ the sets of reduced words beginning with b and b^{-1}. We then have that $a\mathcal{A}_2$ is precisely the set of reduced words which do *not* begin with a (including the empty word e), so that $e\mathcal{A}_1 = \mathcal{A}_1$ and $a\mathcal{A}_2$ are disjoint and $e\mathcal{A}_1 \cup a\mathcal{A}_2 = \mathbb{F}_2$. Similarly, $e\mathcal{B}_1 = \mathcal{B}_1$ and $b\mathcal{B}_2$ are disjoint and $e\mathcal{B}_1 \cup b\mathcal{B}_2 = \mathbb{F}_2$. But this decomposition does not quite work: $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{B}_1, \mathcal{B}_2\}$ is a pairwise disjoint collection of subsets of $\mathbb{F}_2$, but their union is $\mathbb{F}_2 \setminus \{e\}$, not all of $\mathbb{F}_2$.

The decomposition can be fixed by putting e in $\mathcal{A}_1$ and moving all words of the form a^{-n} (i.e. words made of just a string of a^{-1}'s) from $\mathcal{A}_2$ to $\mathcal{A}_1$. The conditions are now all met.

III.12.5.12. More generally, any group containing a subgroup isomorphic to $\mathbb{F}_2$ has a (2,2)-paradoxical decomposition (decompose right cosets of the free subgroup). Conversely, it is known that any group with a (2,2)-paradoxical decomposition must contain a copy of $\mathbb{F}_2$ ().

III.12.5.13. There are groups which are paradoxical but do not contain a free subgroup. It is known that a group is paradoxical if and only if it is not amenable (). We will not give a definition of an amenable group here; see (). Amenable groups are ones which are "nice" in a great many equivalent ways. Many well-behaved groups are amenable:

Any finite group is amenable.

Any abelian group is amenable.

Any extension or direct limit of amenable groups is amenable. In particular, every solvable group is amenable.

There are amenable groups which are not in the smallest class of groups containing the finite and abelian groups and closed under extensions and direct limits (the *elementary amenable groups*).

Group Actions and Paradoxical Sets

We can extend the notion of paradoxical decompositions to group actions.

III.12.5.14. DEFINITION. Let X be a set, and $\mathcal{G}$ be a group acting on X (as a group of bijections of X). A subset E of X is *paradoxical* with respect to the action (*$\mathcal{G}$-paradoxical*) if there is a family $\{A_1, \ldots, A_m, B_1, \ldots, B_n\}$ of pairwise disjoint subsets of E and group elements $g_1, \ldots, g_m, h_1, \ldots, h_n$ such that

$$\bigcup_{j=1}^{m} g_j A_j = \bigcup_{k=1}^{n} h_k B_k = E \ .$$

III.12.5.15. Such a collection can be modified to another such collection for which

$$\bigcup_{j=1}^{m} A_j \cup \bigcup_{k=1}^{n} B_k = E$$

and such that $\{g_1 A_1, \ldots, g_m A_m\}$ and $\{h_1 B_1, \ldots, h_n B_n\}$ are each pairwise disjoint (Exercise III.12.9.4.). If these conditions are met, the A_j, B_k, g_j, h_k are a *$\mathcal{G}$-paradoxical decomposition* of E.

III.12.5.16. A group $\mathcal{G}$ is paradoxical if and only if it is $\mathcal{G}$-paradoxical for the action of $\mathcal{G}$ on itself by left translation. Some other examples:

(i) The Banach-Tarski Paradox (basic version) exactly says that a closed ball centered at the origin in $\mathbb{R}^n$ is $SO(n)$-paradoxical.

(ii) Let κ be a cardinal (). Then κ satisfies $\kappa = 2\kappa$ if and only if a set X of cardinality κ is $\mathcal{G}$-paradoxical with $m = n = 1$, where $\mathcal{G}$ is the group of permutations of X (bijections from X to X); cf. Exercise III.12.9.5.. In particular, a countably infinite set is $\mathcal{G}$-paradoxical.

III.12.6. The Hausdorff Paradox

The fundamental "paradox" found by HAUSDORFF in 1914 [Hau14] is:

III.12.6.1. THEOREM. [HAUSDORFF PARADOX] Let S^2 be the unit sphere in $\mathbb{R}^3$. Then there is a countable subset C of S^2 such that $S^2 \setminus C$ is equidecomposable with two disjoint copies of $S^2 \setminus C$. The equidecomposition can be done with four pieces.

The fundamental underlying result used to prove this theorem is:

III.12.6.2. THEOREM. $SO(3)$ contains a subgroup isomorphic to the free group $\mathbb{F}_2$ on two generators.

III.12.6.3. HAUSDORFF'S proof of this theorem gave an explicit pair of 3×3 special orthogonal matrices which generate a subgroup of $SO(3)$ isomorphic to $\mathbb{F}_2$ (actually he gave two matrices which generate a copy of the free product $\mathbb{Z}_2 * \mathbb{Z}_3$, which contains a copy of $\mathbb{F}_2$). His argument used that the cosine of the angle between the axes of rotation is chosen transcendental. We give a simpler argument due to S. ŚWIERCZKOWSKI [Ś58] (cf. [Wag93]).

PROOF: Let U and V be rotations around the z- and x-axes respectively with angle $\cos^{-1} \frac{1}{3}$, i.e.

$$U = \begin{bmatrix} \frac{1}{3} & -\frac{2\sqrt{2}}{3} & 0 \\ \frac{2\sqrt{2}}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & -\frac{2\sqrt{2}}{3} \\ 0 & \frac{2\sqrt{2}}{3} & \frac{1}{3} \end{bmatrix}.$$

We will show that the subgroup $\mathcal{G}$ of $SO(3)$ generated by U and V is isomorphic to $\mathbb{F}_2$. There is a homomorphism $\phi : \mathbb{F}_2 \to \mathcal{G}$ with $\phi(a) = U$ and $\phi(b) = V$; ϕ sends a word in a, b, a^{-1}, b^{-1} to the corresponding word in U, V, U^{-1}, V^{-1}, so we need to show that ϕ is injective, i.e. no nontrivial reduced word in U, V, U^{-1}, V^{-1} is equal to the identity I in $SO(3)$.

The proof is straightforward and elementary, but there are a number of cases, which we leave to the interested reader. We only outline the argument. Some details can be found in [Wag93, 2.1].

1. If W is a reduced word ending in V or V^{-1} with $W = I$, then $U^{-1}WU$ (reduced) is a nontrivial word which also equals I. Thus we only need to show that no reduced word ending in U or U^{-1} is equal to I.

2. Let $\mathbf{i} = (1, 0, 0)$ be the first standard basis vector. We claim that if W is a word ending in U or U^{-1}, then $W\mathbf{i}$ is of the form $3^{-k}(x, y\sqrt{2}, z)$ for integers x, y, z, with $k \in \mathbb{N}$ the length of W. By calculation, show that if $W\mathbf{i}$ is of this form, so are $UW\mathbf{i}$, $U^{-1}W\mathbf{i}$, $VW\mathbf{i}$, and $V^{-1}W\mathbf{i}$; thus the result follows by induction on the length of W.

3. Show that the y in the formula for $W\mathbf{i}$ is not divisible by 3, by the same method using twelve calculations (which can be combined into similar groups). Separate cases must be done depending on the first letter of W. One needs W to be reduced for this step. This is the hardest step (but not too hard).

4. It follows that if W is any reduced word ending in U or U^{-1}, then the second coordinate of $W\mathbf{i}$ is nonzero, and thus $W\mathbf{i} \neq \mathbf{i}$, $W \neq I$, completing the proof. 👍

III.12.6.4. In fact more is true: almost *every* pair of special orthogonal 3×3 matrices generates a copy of $\mathbb{F}_2$, in the sense that the pairs which do not, i.e. for which there is a nontrivial relation in the subgroup they generate, form a meager subset (XIII.12.1.1.) of $SO(3) \times SO(3)$. See [Wag93] for a proof. The group $SO(3)$ even contains a free group on $2^{\aleph_0}$ generators.

III.12.6.5. We now turn to the proof of Theorem III.12.6.1.. This is the place where the Axiom of Choice is used. The proof is very similar to the construction of VITALI'S nonmeasurable set (XVII.4.10.5.), although no measure theory will be needed for this construction.

Let $\mathcal{G}$ be a subgroup of $SO(3)$ isomorphic to $\mathbb{F}_2$. First note that any element of $SO(3)$ sends S^2 onto S^2, and any element other than the identity has exactly two fixed points (the points of intersection of the axis of rotation with S^2). Let C be the set of all points in S^2 which are fixed points of elements of $\mathcal{G}$. Then C is countable since $\mathcal{G}$ is countable. Set $D = S^2 \setminus C$ for ease of notation. Then $\mathcal{G}$ acts on D by $g \cdot x = U_g(x)$, and no element of $\mathcal{G}$ other than the identity has a fixed point (i.e. the action is *free*). Thus D partitions into the orbits of the action, i.e. for $x \in D$, $[x] = \{g \cdot x : g \in \mathcal{G}\}$. Alternatively, we may define an equivalence relation $\sim$ on D by $x \sim y$ if $y = g \cdot x$ for some $g \in \mathcal{G}$; the orbits are the equivalence classes.

Using the AC, let Q be a set containing exactly one element from each equivalence class. Then the sets $g \cdot Q = \{g \cdot x : x \in Q\}$ for $g \in \mathcal{G}$ are pairwise disjoint and $D = \cup_{g \in \mathcal{G}} (g \cdot Q)$.

If a and b are the generators of $\mathcal{G}$, let $\{\mathcal{G}_1, \ldots, \mathcal{G}_4\}$ be the paradoxical decomposition of $\mathcal{G}$ defined in (). For $1 \leq j \leq 4$ set

$$D_j = \bigcup_{g \in \mathcal{G}_j} g \cdot Q \ .$$

Then the D_j are pairwise disjoint, $D_1 \cup D_2 \cup D_3 \cup D_4 = D$, $(a \cdot D_1) \cup D_2 = D$, and $(b \cdot D_3) \cup D_4 = D$. Let T be translation by two units in one coordinate direction, $E_1 = U_a(D_1)$, $E_2 = D_2$, $E_3 = (T \circ U_b)(D_3)$, and $E_4 = T(D_4)$. Then $\{E_1, E_2, E_3, E_4\}$ are pairwise disjoint, $E_1 \cup E_2 = D$, $E_3 \cup E_4 = T(D)$, and $E_j \approx D_j$ $(1 \leq j \leq 4)$. Thus

$$D = D_1 \cup D_2 \cup D_3 \cup D_4 \sim E_1 \cup E_2 \cup E_3 \cup E_4 = D \cup T(D) \ .$$

This completes the proof of Theorem III.12.6.1.. 

III.12.6.6. None of the four sets $D_1 - D_4$ (or $E_1 - E_4$) can be $\mathscr{H}^2$-measurable, where $\mathscr{H}^2$ is two-dimensional Hausdorff measure on $\mathbb{R}^3$ (), and in particular none can be a Borel set. For D is a Borel set, hence $\mathscr{H}^2$-measurable, and if, say, D_1 is $\mathscr{H}^2$-measurable, then so is E_1 and therefore $D_2 = D \setminus E_1$ is too, and

$$\mathscr{H}^2(D_1 \cup D_2) = \mathscr{H}^2(D_1) + \mathscr{H}^2(D_2) = \mathscr{H}^2(E_1) + \mathscr{H}^2(D_2) = \mathscr{H}^2(D) = 4\pi$$

so $D_3 \cup D_4 = D \setminus (D_1 \cup D_2)$ is an $\mathscr{H}^2$-null set, and therefore D_3, D_4, E_3, E_4 are all $\mathscr{H}^2$-null sets, contradicting that $\mathscr{H}^2(E_3 \cup E_4) = 4\pi$. The arguments for the other sets are identical. ($\mathscr{H}^2$ can be replaced by "surface area measure" on S^2 (), which is just the restriction of $\mathscr{H}^2$ to S^2.) Thus the sets must be very bizarre and cannot be explicitly described.

III.12.6.7. HAUSDORFF did not take the final step III.12.4.5.(iii) which makes the paradox more interesting (he also phrased the paradox somewhat differently). If S_1^2 and S_2^2 are disjoint 2-spheres of radius 1, and C_1 and C_2 the images of the countable set C in S_1^2 and S_2^2 respectively, by III.12.4.5.(iii) we have

$$S^2 \sim S^2 \setminus C \sim (S_1^2 \setminus C_1) \cup (S_2^2 \setminus C_2) \sim S_1^2 \cup S_2^2 \ .$$

This equidecomposition cannot, however, be done with four pieces.

III.12.6.8. The result can be extended to S^n for $n > 2$, because $SO(n)$ contains a copy of $SO(3)$ and thus a copy of $\mathbb{F}_2$. But the extension is trickier since the set of fixed points of a higher-dimensional special orthogonal transformation is more complicated in general than just two points, and even a special orthogonal transformation of infinite order can have periodic points. It can, however, be done quite nicely by an inductive argument.

III.12.6.9. REMARK. We do not quite need the full force of the AC for the Paradox: we only need it for sets (including index sets) of cardinality $\leq 2^{\aleph_0}$. Thus, for example, the Continuum Hypothesis suffices.

III.12.7. The Banach-Tarski Paradox, Basic Version

We now derive the basic Banach-Tarski Paradox, which we repeat here. We will not obtain the sharp version with five pieces, and in fact not worry about how many pieces we need.

III.12.7.1. THEOREM. [BANACH-TARSKI PARADOX, BASIC VERSION] A closed ball of radius r in $\mathbb{R}^3$ is equidecomposable with the disjoint union of two closed balls of radius r.

PROOF: We may take $r = 1$ without loss of generality. Let K be the closed unit ball in $\mathbb{R}^3$ with the origin removed. Let $\{D_1, \ldots, D_m\}$ be pairwise disjoint subsets of S^2 with $\cup_{j=1}^m D_j = S^2$, and $\{E_1, \ldots, E_m\}$ pairwise disjoint moved copies of $D_1, \ldots, D_m$ with $\cup_{j=1}^m E_j$ two disjoint copies of S^2, say $E_j = F_j(D_j)$ for a rigid motion F_j. For each j let K_j be the subset of K obtained by extending each point of D_j radially to (but not including) $\mathbf{0}$. Then $\{K_1, \ldots, K_m\}$ are pairwise disjoint, and $\cup_{j=1}^m K_j = K$. The sets look something like the yucca in Figure III.3.

It is clear that $\{F_1(K_1), \ldots, F_m(K_m)\}$ are pairwise disjoint and that $\cup_{j=1}^m F_j(K_j)$ is two disjoint copies of K, since the rigid motions are just rotations or a rotation followed by a single fixed translation, and do the right thing on the surface sphere.

We thus have K equidecomposable with a disjoint union $K' \cup K''$ of two copies of K. To finish the argument we need to show that $K \sim B$, where B is the entire closed unit ball. But this is easily done using III.12.2.3.(ii): Let C be a circle in, say, the xy-plane, passing through $\mathbf{0}$, of small enough radius that it is contained in B. We have $C \approx C \setminus \{\mathbf{0}\}$, so

$$B = C \cup (B \setminus C) \sim (C \setminus \{0\}) \cup (B \setminus C) = B \setminus \{\mathbf{0}\} = K$$

$$B \sim K \sim K' \cup K'' \sim B' \cup B'' \ .$$





Figure III.3: A paradoxical yucca.

III.12.7.2. Using III.12.6.8., the result can be extended by the same argument to $\mathbb{R}^n$, $n > 3$.

III.12.7.3. The sets in the decomposition cannot all be Lebesgue measurable, by an argument as in III.12.6.6., and in particular cannot all be Borel sets. Thus the statement in popular books that a ball can be "cut" into five pieces and reassembled into two balls must be taken with a grain of salt: the pieces (except the singleton) are dense "clouds" of points which are too complicated to even be explicitly described (they are even more complicated than the "yucca" when done with five pieces).

III.12.7.4. Thus a closed ball is equidecomposable with two disjoint balls of the same size. But why stop with two? The second ball can be further subdivided to make two new balls disjoint from the first, so a ball is equidecomposable with three pairwise disjoint balls of the same size. By inductively repeating the decomposition, a closed ball is equidecomposable with m pairwise disjoint closed balls of the same size, for any $m \in \mathbb{N}$. Note, though, that the minimum number of pieces needed grows as m increases; see III.12.8.2. and III.12.9.3..

III.12.8. The Banach-Tarski Paradox, General Version

The most general version of the Banach-Tarski Paradox we will state is:

III.12.8.1. THEOREM. [BANACH-TARSKI PARADOX, GENERAL VERSION] Let $n \in \mathbb{N}$, $n \geq 3$. If A and B are bounded subsets of $\mathbb{R}^n$ with nonempty interior, then A and B are equidecomposable.

The popular version of the Paradox given in the introduction of this section is an immediate corollary.

III.12.8.2. Note that A and B are extremely general: they do not even have to be Lebesgue measurable (each is just the union of an open ball with an arbitrary bounded set). The result gives no information about the number of pieces needed for given A and B except that it can be done with finitely many; a very large number may be needed (a careful examination of the proofs will give a generally very large upper bound for the minimum number necessary). A lower bound can be easily obtained. Assume, without loss of generality, that $\lambda^*(B) \geq \lambda^*(A)$, where λ^* is Lebesgue outer measure in $\mathbb{R}^n$ (XVII.4.2.1.). If the equidecomposition is done with m pieces $A_1, \ldots, A_m, B_1, \ldots, B_m$, with $A_j \approx B_j$ for each j, then each A_j satisfies $\lambda^*(A_j) \leq \lambda^*(A)$. Thus, by subadditivity and rigid motion invariance of Lebesgue outer measure,

$$\lambda^*(B) \leq \sum_{j=1}^{m} \lambda^*(B_j) = \sum_{j=1}^{m} \lambda^*(A_j) \leq m\lambda^*(A)$$

and so

$$m \geq \frac{\lambda^*(B)}{\lambda^*(A)}$$

(note that $0 < \lambda^*(A), \lambda^*(B) < +\infty$ since A and B are bounded and have nonempty interior). Thus $\frac{\lambda^*(B)}{\lambda^*(A)}$ is a lower bound for the number of pieces necessary (it is potentially even much smaller than the actual minimum needed).

In the case of the popular version of the Paradox, I estimate the ratio of the volume of the sun to the volume of a pea to be roughly 10^{30}, so at least 10^{30} pieces are necessary, and probably far more (the upper bound for the minimum from the proofs is on the order of $(10^{30})^{\log_2 5} \approx 10^{70}$, cf. III.12.9.3.).

The proof of Theorem III.12.8.1. requires two lemmas:

III.12.8.3. LEMMA. Let A and B be subsets of $\mathbb{R}^n$. If $A \sim B$, then for any subset A' of A there is a subset B' of B with $A' \sim B'$.

PROOF: Let $\{A_1, \ldots, A_m\}$ and $\{B_1, \ldots, B_m\}$ be partitions of A and B respectively, and F_j $(1 \leq j \leq m)$ rigid motions with $F_j(A_j) = B_j$. For each j set $A'_j = A_j \cap A'$ and $B'_j = F_j(A'_j)$, and set $B' = \cup_{j=1}^{m} B'_j$. Then $\{A'_1, \ldots, A'_m\}$ and $\{B'_1, \ldots, B'_m\}$ are partitions of A' and B' respectively into disjoint subsets, and $A'_j \approx B'_j$ for each j, so $A' \sim B'$. 👌

The next lemma is the essential step in the proof of III.12.8.1..

III.12.8.4. LEMMA. Let B be a closed ball in $\mathbb{R}^n$. If A is any bounded subset of $\mathbb{R}^n$, then A is equidecomposable with a subset of B.

PROOF: Since A is bounded, it can be covered by a finite number $\{B_1, \ldots, B_m\}$ of closed balls the same size as B. Let $\{D_1, \ldots, D_m\}$ be a collection of pairwise disjoint closed balls the same size as B. Then $B \sim D := \cup_{j=1}^m D_j$ by III.12.7.4., so by Lemma III.12.8.3. it suffices to show that A is equidecomposable with a subset of D.

For each j let F_j be a rigid motion (translation) with $F_j(B_j) = D_j$. Set $A_1 = A \cap B_1$, and inductively set

$$A_j = A \cap \left[B_j \setminus \bigcup_{k=1}^{j-1} B_k \right]$$

and $C_j = F_j(A_j) \subseteq D_j$. Let $C = \cup_{j=1}^m C_j$. Then $\{A_1, \ldots, A_m\}$ and $\{C_1, \ldots, C_m\}$ are partitions of A and C respectively, and $A_j \approx C_j$ for each j, so $A \sim C$.

III.12.8.5. To prove III.12.8.1., since A is bounded and B contains a closed ball D, by Lemma III.12.8.4. A is equidecomposable with a subset of B. Similarly, B is equidecomposable with a subset of A. Thus $A \sim B$ by the Schröder-Bernstein Theorem for equidecomposability III.12.4.6..

Note that we have only given a proof of III.12.7.4., and hence of III.12.8.4. and III.12.8.1., for $n = 3$.

III.12.9. Exercises

III.12.9.1. (a) Show that a rigid motion of $\mathbb{R}^1$ is a translation.

(b) If A is a bounded subset of $\mathbb{R}^1$ and $A \approx B$ for some $B \subseteq A$, then $B = A$. [Consider the action of the rigid motion on $\sup(A)$ and $\inf(A)$.]

(c) Does (b) remain true if $\approx$ is replaced by $\sim$?

III.12.9.2. [Wag93] If A and B are bounded subsets of $\mathbb{R}^2$ and $A \sim B$, show that $\lambda^*(A) = \lambda^*(B)$. Thus nothing like the Banach-Tarski Paradox is possible in $\mathbb{R}^2$. (This is "explained" by the fact that $SO(2)$ is abelian and thus not paradoxical.)

III.12.9.3. (a) By successively dividing each ball into two balls, show that a ball of radius r in $\mathbb{R}^3$ is equidecomposable with a disjoint union of 2^k balls of radius r using 5^k pieces. (Since splitting a ball in two can be done using five pieces and one of the pieces is a singleton, by a more careful analysis the bound should be reducible to a little over 4^k pieces.)

(b) A ball of radius r in $\mathbb{R}^3$ can be split into m balls of radius r using roughly $m^{\log_2 5} \approx m^{2.32}$ pieces. (This should be reducible to a little over m^2 pieces.)

III.12.9.4. Let $\mathcal{G}$ be a group acting on a set X, $E \subseteq X$, and $\{A_1, \ldots, A_m, B_1, \ldots, B_n\}$ and $g_1, \ldots, g_m, h_1, \ldots, h_n$ be as in Definition III.12.5.14..

(a) Show that $A_1, \ldots, A_m, B_1, \ldots, B_n$ can be modified so that $\{g_1 A_1, \ldots, g_m A_m\}$ and $\{h_1 B_1, \ldots, h_n B_n\}$ are each pairwise disjoint and the other conditions are still satisfied (fairly easy).

(b) Show that $A_1, \ldots, A_m, B_1, \ldots, B_n$ can be additionally modified so that

$$\bigcup_{j=1}^{m} A_j \cup \bigcup_{k=1}^{n} B_k = E$$

and the other conditions are still satisfied (not so easy; see [Wag93,]).

III.12.9.5. Using the Schröder-Bernstein Theorem, prove the statement in III.12.5.16.(ii).

III.12.9.6. Carry out the calculations to complete the proof of III.12.6.2..

Chapter IV

Sequences and Series

Infinite sequences and infinite series can be regarded as the beginning of analysis. The idea that a sequence of numbers can approach a limiting value, or that an infinite set of numbers can be "added", is the principal concept that underlies the whole development of the subject. The notion of limit of a sequence is a subtle one which takes some getting used to, but which must be mastered before calculus and analysis of any type can be understood.

Sequences and their limits appear in various contexts throughout mathematics. The most basic situation is sequences of real (or complex) numbers, which is where we begin. The subject then moves on to sequences of points in a metric space, and then to sequences in topological spaces; in fact, the entire subject of topology can be regarded as a way of giving a general setting for studying limits of sequences. Another (intertwined) direction of development is the theory of limits of sequences of functions. The study of infinite series of numbers or functions may be regarded as a special case of the theory of sequences, or really the same theory from a different point of view, although much of this theory predated the advent of a careful theory of sequences and developed in parallel.

In this chapter, we will develop the theory of sequences and their limits, first for numbers and then in more general settings, and the theory of infinite series of numbers. Infinite sequences and series of functions will be postponed until the calculus chapter.

IV.1. Sequences

Recall () that a *sequence* in a set X is a function from $\mathbb{N}$ to X. The usual notation uses a subscript for the argument of the function, which is called the *index* of the sequence. Thus the value of the function (sequence) $\boldsymbol{a}$ at the number n is normally written a_n instead of $a(n)$; a_n is called the *n'th term* of the sequence. Occasionally the index of a sequence takes values in a set of the form $\{n \in \mathbb{Z} : n \geq n_0\}$ instead of in $\mathbb{N}$.

The notation (x_n) is commonly used to denote a sequence; the notation $(x_n)_{n\in\mathbb{N}}$ is a more precise variation. Some authors use the notation $\langle x_n \rangle$, or something similar. Note the difference between (x_n) and x_n: (x_n) refers to the entire sequence, while x_n denotes the n'th term of the sequence for a specific single n. We sometimes write a formula for the n'th term, surrounded by parentheses, to define a sequence; thus, for example, $\left(\frac{1}{n}\right)$ or, more precisely, $\left(\frac{1}{n}\right)_{n\in\mathbb{N}}$, denotes the sequence (x_n) with $x_n = \frac{1}{n}$ for each n.

IV.1.1. Sequences of Numbers

In this section and the subsequent subsections, we will concentrate on sequences of numbers, real or (occasionally) complex. The structure of the real numbers makes possible some definitions and operations which do not make sense for sequences in general sets.

Terms and Range of Sequences

IV.1.1.1. If (x_n) is a sequence of numbers (or a sequence in a general set X), a careful distinction must be made between the terms of the sequence and the set $\{x_n : n \in \mathbb{N}\}$. This set is the range of the function from $\mathbb{N}$ to $\mathbb{R}$ (or X). For example, consider the sequence $((-1)^n) = (-1, 1, -1, 1, \dots)$. There are infinitely many terms in this sequence, as there are in any sequence, but the range of the sequence is the two-point set $\{1, -1\}$. The terms x_2 and x_4 are different terms of the sequence, even though these terms have the same numerical value. The most extreme example is a *constant sequence*, where all the terms of the sequence have the same numerical value; they are nonetheless distinct terms.

This distinction causes complications when we try to depict sequences on the number line. Fortunately, sequences which arise naturally in applications tend to have terms with distinct values, i.e. as a function the sequence is injective, and for such sequences the distinction largely disappears, but not entirely: for example, if the order of the terms is rearranged a different sequence is obtained, even though it has the same range.

Bounded and Unbounded Sequences

IV.1.1.2. Definition. A sequence (x_n) of real numbers is *bounded above* if there is an $M \in \mathbb{R}$ such that $x_n \leq M$ for all n. The sequence (x_n) is *bounded below* if there is an $M \in \mathbb{R}$ such that $M \leq x_n$ for all n. The sequence (x_n) is *bounded* if it is both bounded above and bounded below, i.e. if there is an $M \in \mathbb{R}$ such that $|x_n| \leq M$ for all n. A sequence which is not bounded is *unbounded*.

In other words, the sequence (x_n) is bounded [bounded above, bounded below] if and only if the range $\{x_n : n \in \mathbb{N}\}$ is bounded [bounded above, bounded below] as a subset of $\mathbb{R}$.

Algebraic Operations on Sequences

IV.1.2. Subsequences and Tails

IV.1.3. Limits of Sequences

IV.1.3.1. Now we turn to the fundamental concept of analysis, the notion of what it means for a sequence to *converge* to a limit. The idea is that the sequence (x_n) approaches the limit (limiting value) a if eventually, i.e. from some point on, the terms of the sequence are approximately equal to a to within any given tolerance. Pictorially, for sequences of numbers, on the number line the number x_n should become almost indistinguishable from the limiting value a for all sufficiently large n:

The terms of the sequence, however, do not ever have to be *exactly* equal to the limit, and "usually" are not. Indeed, the notion of limit is most interesting for sequences which are not eventually constant.

It is important to note that "most" sequences of numbers do not approach a limiting value, i.e. do not converge. In fact, it is quite special for a sequence to have a limit.

IV.1.3.2. A way of making this notion precise is to say that the sequence (x_n) approaches the limit a if, no matter how tiny an interval I around a we specify, all the terms of the sequence eventually lie within the interval I. The smaller the interval I, the longer we might have to wait before the terms of the sequence settle down into I, but eventually they must for any I:

We are thus led to the following formal definition.

IV.1.3.3. Definition. [Limit of a Sequence, Version 1] Let (x_n) be a sequence in $\mathbb{R}$, and $a \in \mathbb{R}$. The sequence (x_n) *converges* to a, or a is the *limit* of the sequence (x_n), if for every interval I in $\mathbb{R}$ containing a in its interior, there is an $N \in \mathbb{N}$ such that $x_n \in I$ for all $n \geq N$.

IV.1.3.4. It turns out to usually be more useful in practice to rephrase the definition slightly. It is only necessary to consider open intervals which are *centered* at a, since any interval containing a in its interior contains a smaller open interval centered at a. It is customary in analysis to use the Greek letter epsilon (ϵ), or the Greek letter delta (δ), to denote the length of the interval to the left or right of a; thus any (bounded) interval centered at a is of the form

$$(a - \epsilon, a + \epsilon)$$

for some positive real number ϵ, and conversely any $\epsilon > 0$ defines such an interval:

Thus we can rephrase the definition in an equivalent way:

IV.1.3.5. Definition. [Limit of a Sequence, Version 2] Let (x_n) be a sequence in $\mathbb{R}$, and $a \in \mathbb{R}$. The sequence (x_n) *converges* to a, or a is the *limit* of the sequence (x_n), if for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$a - \epsilon < x_n < a + \epsilon$$

for all $n \geq N$.

There is one last slight notational variation. Note that $|x_n - a| < \epsilon$ if and only if $a - \epsilon < x_n < a + \epsilon$. Thus we can restate the definition in its usual form:

IV.1.3.6. Definition. [Limit of a Sequence, Final Version] Let (x_n) be a sequence in $\mathbb{R}$, and $a \in \mathbb{R}$. The sequence (x_n) *converges* to a, or a is the *limit* of the sequence (x_n), if for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$|x_n - a| < \epsilon$$

for all $n \geq N$.

The sequence (x_n) *converges* if there is a real number a such that a is the limit of the sequence (x_n). If there is no such a, the sequence (x_n) *diverges*.

IV.1.3.7. It is unimportant in this definition whether or not we use strict inequality in the $|x_n - a| < \epsilon$ and/or the $n \geq N$, or whether we require that $N \in \mathbb{N}$; see Exercise IV.1.12.1..

IV.1.3.8. NOTATION: There are two standard notations for the limit of a sequence. If (x_n) converges to a, we write

$$x_n \to a$$

or

$$\lim_{n\to\infty} x_n = a \ .$$

The last expression is read "the limit of x_n as n goes to infinity is a."

Note that we write $\lim_{n\to\infty} x_n = a$, not $\lim_{n\to\infty} x_n \to a$.

IV.1.3.9. The definition of limit is often described as a game or bet. If I claim that $x_n \to a$ and you challenge me, we play a game. You give me a positive number ϵ, and I must produce an N that works in the definition for the ϵ you give me. If I can, I win; if I can't, you win. The smaller the ϵ you give me, the harder it will be for me to find an N that works, and the larger it will have to be in general. But I don't have to find the N until you give me your ϵ; my N can (and in general will) depend on what your ϵ is.

A proof that $x_n \to a$ really consists of a demonstration that I can always win the game; I must convince you that no matter what ϵ you give me, I can always find an N that works, so that you realize it would be pointless for you to challenge me with any specific ϵ. On the other hand, a disproof that $x_n \to a$ normally consists of demonstrating a specific positive ϵ which is small enough that there is no N that works for it.

Note that for a given ϵ, there is more than one possible N that works: the N can always be increased with no penalty. If there is any N that works, there is always a smallest N, but it may be difficult to find the smallest N in specific situations, and it is usually not necessary; a larger N which is easier to find works just as well. (There are instances, e.g. in some numerical work, where it is important to find the smallest, or at least very nearly the smallest, N that works in a given setting, but such situations are rather unusual in analysis.)

In the same way, if a certain N works for one ϵ for a particular limit, the same N works for any larger ϵ, so there is no penalty in first making the ϵ smaller before finding an N that works. It may be hard to believe that it can sometimes be useful to do this, but it is; see, for example, ().

IV.1.3.10. EXAMPLES. (i) Let $x_n = \frac{1}{n}$ for all n. This is one of the most important limits, and will appear constantly. It seems intuitively clear that $x_n \to 0$. To prove this, we must show that for any $\epsilon > 0$, there is an N such that

$$|x_n - 0| = \left|\frac{1}{n} - 0\right| = \left|\frac{1}{n}\right| = \frac{1}{n} < \epsilon$$

for all $n \geq N$. So suppose $\epsilon > 0$ is given. In order to have $\frac{1}{n} < \epsilon$, it is sufficient (and also necessary) that $n > \frac{1}{\epsilon}$. Thus we can choose N to be any natural number greater than $\frac{1}{\epsilon}$. There is such an N by the Archimedean property of $\mathbb{R}$ (). Then if $n \geq N$, we have

$$|x_n - 0| = \frac{1}{n} \leq \frac{1}{N} < \epsilon$$

so this N works for the given ϵ.

Note that if ϵ is very small, then $\frac{1}{\epsilon}$ is very large, so N will have to be large; the smaller the ϵ, the larger the N will have to be. For example, if $\epsilon = 10^{-9}$, we will have to take N = 1,000,000,001 (or any larger integer).

Note that none of the terms of this sequence are exactly equal to the limit 0.

(ii) Similarly, suppose $x_n = 1 + \frac{1}{n}$. We guess that $x_n \to 1$. To prove this, we must show that, if $\epsilon > 0$, there is an N such that

$$|x_n - 1| = \left| \left(1 + \frac{1}{n}\right) - 1 \right| = \left| \frac{1}{n} \right| < \epsilon$$

for all $n \geq N$. Thus, by an identical argument to the one in (i), we can choose N to be any integer greater than $\frac{1}{\epsilon}$.

(iii) Here is an example which is so simple that it is hard. Suppose (x_n) is a constant sequence with $x_n = a$ for all n. We certainly expect that $x_n \to a$. If $\epsilon > 0$, we have

$$|x_n - a| = |a - a| = 0 < \epsilon$$

for *any* n. Thus we can take $N = 1$ no matter what ϵ is, so $x_n \to a$. This is practically the only example where the N does not depend on ϵ.

(iv) Suppose $x_n = (-1)^n$ for all n. Since this sequence alternates ± 1, we might expect that the sequence has two limits, 1 and -1. However, this is incorrect; in fact, this sequence diverges, i.e. has no limit, since to have a limit a *all* the terms of the sequence from some point on must be close to a. (We will later () define cluster points for sequences; this sequence has cluster points 1 and -1.)

For example, let us show that (x_n) does not converge to 1. We have $|x_n - 1| = 0$ if n is even, but $|x_n - 1| = 2$ for n odd. If we are given an $\epsilon > 0$, we may or may not be able to find an N that works. If the person challenging us is stupid enough to give us $\epsilon = 3$ (or any $\epsilon > 2$), we can find an N that works: in fact, $N = 1$ works for such an ϵ. But if our challenger is smart enough to give us an ϵ which is ≤ 2, we cannot find an N that works, since no matter what N we take there will be an odd n which is greater than N, and for this n we have

$$|x_n - 1| = 2 \geq \epsilon \ .$$

Thus we have shown that (x_n) does not converge to 1. An almost identical argument shows that (x_n) does not converge to -1.

However, this does not yet show that the sequence (x_n) diverges, i.e. that $\lim_{n\to\infty} x_n$ does not exist. To do this, we must show that for any $a \in \mathbb{R}$, the sequence does not converge to a. Fix an $a \in \mathbb{R}$. It suffices to exhibit an $\epsilon > 0$ for which there is no N that works for this ϵ and a. Take $\epsilon = 1$ (any smaller positive number also works). Suppose there is an N that works, i.e.

$$|x_n - a| < 1$$

for all $n \geq N$. If we take n to be an odd number greater than N, we have that

$$|x_n - a| = |-1 - a| = |1 + a| < 1$$

and similarly if $n \geq N$ is even, we have that

$$|x_n - a| = |1 - a| < 1 \ .$$

Thus we necessarily have the inequalities $|1 + a| < 1$ and $|1 - a| < 1$, so using the triangle inequality we have

$$2 = |(1 + a) + (1 - a)| \leq |1 + a| + |1 - a| < 1 + 1$$

which is a contradiction. Thus the sequence cannot converge to any $a \in \mathbb{R}$, and hence diverges.

IV.1.3.11. In many books and references, the phrase "closer and closer" is used to informally describe the limiting process, i.e. $x_n \to a$ if x_n gets "closer and closer" to a as $n \to \infty$. But this phrasing is too imprecise to be an adequate definition, and is even not entirely correct. For example, as SAUNDERS MACLANE was fond of pointing out, the sequence $\left(1 + \frac{1}{n}\right)$ gets closer and closer to 0 as $n \to \infty$. The missing idea is the notion of eventually becoming *arbitrarily* close. We might add that if, for example, $x_n = 1$ for all n, then $x_n \to 1$, but x_n does not get "closer and closer" to 1 as n increases since one cannot get any closer to 1 than 1 itself. And, anyway, if $x_n \to a$ the distance between x_n and a need not monotonically decrease as n increases, e.g. if $x_n = \frac{1}{n^2}$ for n odd and $x_n = \frac{1}{n}$ for n even (with $a = 0$), so x_n does not really get "closer and closer" to a. Thus it is best not to use this phrase at all even though it has some conceptual legitimacy.

IV.1.3.12. Indeed, the primary intellectual step forward (primarily due to CAUCHY, although reaching full fruition only with WEIERSTRASS) which made a satisfactory definition of limit possible was abandonment of the "dynamic" idea of *approach*, which previously dominated thought and discourse on the subject and was an apparently insurmountable obstacle to making a careful and rigorous definition, replacing it with a "static" one of simply being approximately equal within an arbitrarily specified tolerance. The dynamic point of view still has considerable intuitive value, however.

Uniqueness of Limits

We have used some suggestive language in Definition IV.1.3.6.: we have referred to "*the* limit of a (convergent) sequence." To justify this language, we must show that a sequence cannot have more than one limit:

IV.1.3.13. THEOREM. The limit of a convergent sequence is unique: if (x_n) is a sequence in $\mathbb{R}$, $a, b \in \mathbb{R}$, and $x_n \to a$ and $x_n \to b$, then $a = b$.

The idea of the proof is simple: if I and J are intervals around a and b respectively, and both $x_n \to a$ and $x_n \to b$ are true, then for all sufficiently large n we must have $x_n \in I$, and similarly for all sufficiently large n we have $x_n \in J$. But if $a \neq b$, then if I and J are short enough they are disjoint, so we cannot simultaneously have $x_n \in I$ and $x_n \in J$ for any n.

The details of the proof give a good example of how the definition of limit is used, and also of some of the common basic techniques used in analysis proofs. A reader unfamiliar with analysis proofs should be sure to examine the proof carefully to understand how it works logically and how it reflects the informal outline of the previous paragraph.

PROOF: We will prove the statement by contradiction. Suppose $a \neq b$, i.e. $|a - b| > 0$. Set $\epsilon = \frac{1}{3}|a - b|$; then $\epsilon > 0$. Since $x_n \to a$, there is an N_1 such that $|x_n - a| < \epsilon$ for all $n \geq N_1$. Similarly, there is an N_2 such that $|x_n - b| < \epsilon$ for all $n \geq N_2$. Set $n = \max(N_1, N_2)$. Then $|x_n - a| < \epsilon$ and $|x_n - b| < \epsilon$ for this n, so using the triangle inequality () we have

$$|a - b| = |(a - x_n) + (x_n - b)| \leq |a - x_n| + |x_n - b| < \epsilon + \epsilon = 2\epsilon = \frac{2}{3}|a - b|$$

and dividing both ends by $|a - b|$, we obtain $1 < \frac{2}{3}$, which is a contradiction.

See IV.1.12.2. for an alternate version of the proof by direct argument instead of contradiction.

IV.1.3.14. The "trick" of adding and subtracting the same thing within an absolute value and using the triangle inequality is one of the most important techniques in analysis, and will be seen in a great many arguments. (It is said that the difference between a "trick" and a "technique" is that a trick is a procedure which is used once, or a few times, and a technique a procedure which is used multiple times. This is definitely a technique under this distinction.)

IV.1.3.15. There is a subtle point connected with the proof of IV.1.3.13., which can be confusing to beginners. If I am trying to prove that $x_n \to a$, I am not allowed to choose the ϵ; I must show that there is an N that works for any ϵ someone else may decide to give me. But if I am *assuming* that $x_n \to a$ and want to use this to prove something else, as in the proof of IV.1.3.13., I am free to choose any ϵ I please, as long as it is positive, and I am assured (by my assumption) that there is an N that works for this ϵ.

Boundedness and Convergence

A convergent sequence is bounded:

IV.1.3.16. PROPOSITION. Let (x_n) be a convergent sequence in $\mathbb{R}$. Then (x_n) is bounded: there is an $M \in \mathbb{R}$ such that $|x_n| \leq M$ for all n.

PROOF: Suppose $x_n \to a$. Let N be a natural number that works in the definition of limit for $\epsilon = 1$, i.e. $|x_n - a| < 1$ for all $n \geq N$. Then $a - 1 < x_n < a + 1$ for all $n \geq N$, so $|x_n| \leq |a| + 1$ for all $n \geq N$. Set

$$M = \max(|x_1|, |x_2|, \ldots, |x_{N-1}|, |a| + 1)$$

(this is a finite set of numbers, so the maximum exists). Then $|x_n| \leq M$ for all n.

IV.1.3.17. The converse is false: a bounded sequence does not converge in general. A simple example is $((-1)^n)$. But a bounded sequence always has a convergent subsequence ().

A companion result, showing that the limit of a convergent sequence lies within any closed interval containing the sequence, is very important:

IV.1.3.18. THEOREM. Let (x_n) be a convergent sequence in $\mathbb{R}$, with $x_n \to c$. Then

(i) If $a \in \mathbb{R}$, $a < c$, then there is an $N \in \mathbb{N}$ such that $a < x_n$ for all $n \geq N$.

(ii) If $b \in \mathbb{R}$, $c < b$, then there is an $N \in \mathbb{N}$ such that $x_n < b$ for all $n \geq N$.

(iii) If $a \in \mathbb{R}$, $a \leq x_n$ for all n or more generally if there is an N such that $a \leq x_n$ for all $n \geq N$, then $a \leq c$.

(iv) If $b \in \mathbb{R}$, $x_n \leq b$ for all n or more generally if there is an N such that $x_n \leq b$ for all $n \geq N$, then $c \leq b$.

(v) If $a, b \in \mathbb{R}$, $a \leq x_n \leq b$ for all n or more generally if there is an N such that $a \leq x_n \leq b$ for all $n \geq N$, then $a \leq c \leq b$.

PROOF: (i): Set $\epsilon = c - a > 0$. Then there is an N such that

$$a = c - \epsilon < x_n$$

for all $n \geq N$.

(ii): Nearly identical to (i), setting $\epsilon = b - c$.

(iii) and (iv) are the contrapositives of (ii) and (i) respectively, and (v) is a combination of (iii) and (iv). 👍

IV.1.3.19. **Caution:** The forms of the inequalities in this result are very important. In (i) and (ii) strict inequality in the hypothesis is essential. In (iii)-(v), if we have strict inequality in the hypothesis, the result applies, but we do not necessarily get strict inequality in the conclusion. For example, if $x_n = \frac{1}{n}$, then $0 < x_n$ for all n, but the limit is not positive. ✋

Here is a simple corollary of (i) and (ii) which is useful:

IV.1.3.20. COROLLARY. Let (x_n) be a sequence of real numbers, and $c \in \mathbb{R}$. If $x_n \to c$ and $c \neq 0$, then there is an $N \in \mathbb{N}$ such that $x_n \neq 0$ for all $n \geq N$.

PROOF: Apply IV.1.3.18.(i) with $a = 0$ if $c > 0$, and IV.1.3.18.(ii) with $b = 0$ if $c < 0$. 👍

This corollary also has a simple direct proof which works in $\mathbb{C}$ as well as in $\mathbb{R}$: Set $\epsilon = \frac{|c|}{2}$ and choose N as in the definition of limit for this ϵ.

Limits and Subsequences

Algebraic Operations and Limits

The algebraic operations (addition, multiplication, division) behave nicely with respect to limits.

IV.1.3.21. THEOREM. Let (x_n) and (y_n) be sequences in $\mathbb{R}$. Suppose $x_n \to a$ and $y_n \to b$. Then

(i) $x_n + y_n \to a + b$.

(ii) $x_n y_n \to ab$.

(iii) If $y_n \neq 0$ for all n and $b \neq 0$, then $\frac{x_n}{y_n} \to \frac{a}{b}$.

IV.1.3.22. The theorem is often informally stated as "the limit of a sum [product, quotient] is the sum [product, quotient] of the limits," and written

$$\lim_{n\to\infty}(x_n+y_n)=\left(\lim_{n\to\infty}x_n\right)+\left(\lim_{n\to\infty}y_n\right)$$

$$\lim_{n\to\infty}(x_ny_n)=\left(\lim_{n\to\infty}x_n\right)\left(\lim_{n\to\infty}y_n\right)$$

$$\lim_{n\to\infty}\left(\frac{x_n}{y_n}\right)=\frac{\lim\limits_{n\to\infty}x_n}{\lim\limits_{n\to\infty}y_n}$$

(provided $\lim\limits_{n\to\infty}y_n\neq 0$).

However, stating the result this way can be misleading if one is not careful. The results only hold under the assumption that the sequences (x_n) and (y_n) converge, i.e. the limits on the right side exist; existence of the limits on the left in these expressions does not mean the limits on the right exist and that the limits on the left can be calculated in the indicated way. So care must be exercised not to try to apply the converse of the statements of IV.1.3.21..

The following special case of IV.1.3.21.(ii) (with (y_n) a constant sequence) is used very frequently:

IV.1.3.23. Corollary. Let (x_n) be a sequence of real numbers, and $a\in\mathbb{R}$. If $x_n\to a$, then for any $c\in\mathbb{R}$ the sequence (cx_n) converges to ca. In other words, if (x_n) converges, then (cx_n) converges and

$$\lim_{n\to\infty}(cx_n)=c\cdot\left(\lim_{n\to\infty}x_n\right)\ .$$

IV.1.3.24. We first give the proof of IV.1.3.21.(i). We will give the reasoning steps in great detail since the technique is a very common one which can be confusing to beginners; note the comments in IV.1.3.15..

Let $\epsilon>0$ be given. We must find an N (i.e. show that one exists) such that

$$|(x_n+y_n)-(a+b)|<\epsilon$$

for all $n\geq N$. We assume that $x_n\to a$, so for any $\epsilon_1>0$ (not necessarily the given ϵ) there is an N_1 such that $|x_n-a|<\epsilon_1$ for all $n\geq N_1$. Similarly, since $y_n\to b$, for any $\epsilon_2>0$ (not necessarily the same as either ϵ or ϵ_1) there is an N_2 such that $|y_n-b|<\epsilon_2$ for all $n\geq N_2$. Thus, if we fix an ϵ_1 and an ϵ_2 and take the corresponding N_1 and N_2, and set $N=\max(N_1,N_2)$, we have

$$|(x_n+y_n)-(a+b)|=|(x_n-a)+(y_n-b)|\leq|x_n-a|+|y_n-b|<\epsilon_1+\epsilon_2$$

for all $n\geq N$. We are given ϵ, but we are free to choose ϵ_1 and ϵ_2 however we please, and there are automatically corresponding N_1 and N_2, hence an N that works for the limit of the sum for $\epsilon_1+\epsilon_2$. All we need to do is choose ϵ_1 and ϵ_2 so that $\epsilon_1+\epsilon_2\leq\epsilon$, and the process gives us an N that works for the limit of the sum for the given ϵ. The simplest thing is just to choose

$$\epsilon_1=\epsilon_2=\frac{\epsilon}{2}\ .$$

All of this could be regarded as preliminary "scratch work" giving us the idea of the proof. The way we would write the proof formally is:

PROOF: Let $\epsilon > 0$. Since $x_n \to a$, there is an N_1 such that $|x_n - a| < \frac{\epsilon}{2}$ for all $n \geq N_1$. Since $y_n \to b$, there is an N_2 such that $|y_n - b| < \frac{\epsilon}{2}$ for all $n \geq N_2$. Set $N = \max(N_1, N_2)$. Then, for all $n \geq N$,

$$|(x_n + y_n) - (a + b)| = |(x_n - a) + (y_n - b)| \leq |x_n - a| + |y_n - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and so, since $\epsilon > 0$ is arbitrary, $x_n + y_n \to a + b$. 👍

IV.1.3.25. Now we turn to the proof of IV.1.3.21.(ii). The argument is somewhat similar, but more complicated. We go through the reasoning more quickly.

As before, for any ϵ_1 and ϵ_2 we have an N such that $|x_n - a| < \epsilon_1$ and $|y_n - b| < \epsilon_2$ for all $n \geq N$. Then, by the adding and subtracting technique, we have

$$|x_n y_n - ab| = |(x_n y_n - a y_n) + (a y_n - ab)| \leq |x_n y_n - a y_n| + |a y_n - ab| = |y_n||x_n - a| + |a||y_n - b| < |y_n|\epsilon_1 + |a|\epsilon_2$$

for all $n \geq N$. Thus, if we are given an $\epsilon > 0$, we want to choose ϵ_1 and ϵ_2 so that

$$|y_n|\epsilon_1 + |a|\epsilon_2 \leq \epsilon$$

for all n. We might try

$$\epsilon_1 = \frac{\epsilon}{2|y_n|} \quad \text{and} \quad \epsilon_2 = \frac{\epsilon}{2|a|}$$

but there is a serious problem with the first of these since ϵ_1 cannot depend on n. (There is a slight problem with the other one too since we could have $a = 0$.)

Fortunately, there is an easy way past the difficulty. Since (y_n) converges, it is bounded (IV.1.3.16.); hence there is an M such that $|y_n| \leq M$ for all n. We might as well choose M so that $|a| \leq M$ to ease notation; and we can assume $M > 0$ since there is no penalty in increasing M (and the case where we can take $M = 0$ is pretty trivial anyway). Then we can take

$$\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2M}$$

for this M. Thus the formal proof is:

PROOF: Let $\epsilon > 0$. Since (y_n) converges, there is an M such that $|y_n| \leq M$ for all n by IV.1.3.16.; we may also assume $M > 0$ and $|a| \leq M$ by increasing M if necessary. Since $x_n \to a$, there is an N_1 such that $|x_n - a| < \frac{\epsilon}{2M}$ for all $n \geq N_1$, and since $y_n \to b$ there is an N_2 such that $|y_n - b| < \frac{\epsilon}{2M}$ for all $n \geq N_2$. Set $N = \max(N_1, N_2)$. If $n \geq N$, then

$$|x_n y_n - ab| = |(x_n y_n - a y_n) + (a y_n - ab)| \leq |x_n y_n - a y_n| + |a y_n - ab| = |y_n||x_n - a| + |a||y_n - b|$$

$$\leq M|x_n - a| + M|y_n - b| < M\frac{\epsilon}{2M} + M\frac{\epsilon}{2M} = \epsilon$$

and so, since $\epsilon > 0$ is arbitrary, $x_n y_n \to ab$. 👍

IV.1.3.26. Finally, we prove (iii). It will actually suffice to prove the following special case:

LEMMA. Let (y_n) be a convergent sequence of real numbers, $y_n \to b$, with all $y_n \neq 0$ and $b \neq 0$. Then $\frac{1}{y_n} \to \frac{1}{b}$.

Then (iii) follows by combining the lemma with (ii). (Note also that the hypothesis in (iii) that $y_n \neq 0$ for all n is almost automatic: by IV.1.3.20., $y_n \neq 0$ for all sufficiently large n if $b \neq 0$.)

To prove the lemma, note that

$$\frac{1}{y_n} - \frac{1}{b} = \frac{b - y_n}{by_n}$$

for all n. Thus we have

$$\left|\frac{1}{y_n} - \frac{1}{b}\right| = \frac{1}{|by_n|}|y_n - b|$$

for all n. If we can show that the sequence $\left(\frac{1}{|by_n|}\right)$ is bounded, we can easily proceed. Note that since $b \neq 0$, there is an N_1 such that $|y_n - b| < \frac{|b|}{2}$ for all $n \geq N_1$, and thus $|y_n| > \frac{|b|}{2}$ for $n \geq N_1$. Set

$$M = \max(\frac{1}{|by_1|}, \ldots, \frac{1}{|by_{N_1-1}|}, \frac{2}{|b|^2})$$

and note that $\frac{1}{|by_n|} \leq M$ for all n (and that $M > 0$).

Now suppose $\epsilon > 0$ is given. There is an N such that

$$|y_n - b| < \frac{\epsilon}{M}$$

for all $n \geq N$. Then, for $n \geq N$,

$$\left|\frac{1}{y_n} - \frac{1}{b}\right| = \frac{1}{|by_n|}|y_n - b| \leq M\frac{\epsilon}{M} = \epsilon\ .$$

IV.1.3.27. Note that IV.1.3.21.(iii) does not apply if $b = 0$, even if $y_n \neq 0$ for all n. It turns out that if $b = 0$ and $a \neq 0$, then $\lim_{n\to\infty}\left(\frac{x_n}{y_n}\right)$ cannot exist (since if it did, and equaled c, we would have

$$a = \lim_{n\to\infty}\left(y_n\frac{x_n}{y_n}\right) = \left(\lim_{n\to\infty} y_n\right)\left(\lim_{n\to\infty}\left(\frac{x_n}{y_n}\right)\right) = bc$$

by (ii), which is impossible). But if $a = b = 0$, the limit of the quotient sequence can exist, but cannot be computed simply in general. This is a type of *indeterminate form*. One of the important parts of calculus and analysis is the resolution of indeterminate forms of various types; see (), (), ...

Limits in $\mathbb{C}$

Limits in $\mathbb{C}$ behave essentially identically to those in $\mathbb{R}$, although the geometric interpretation is a little different (it is essentially like the interpretation in $\mathbb{R}^n$ described in ()). The definition of convergence in $\mathbb{C}$ is identical to the final version (IV.1.3.6.) with $\mathbb{R}$ replaced by $\mathbb{C}$ throughout and no other changes:

IV.1.3.28. Definition. [Limit of a Complex Sequence] Let (z_n) be a sequence in $\mathbb{C}$, and $c \in \mathbb{C}$. The sequence (z_n) *converges* to c, or c is the *limit* of the sequence (z_n), if for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$|z_n - c| < \epsilon$$

for all $n \geq N$.

The sequence (z_n) *converges* if there is a complex number c such that c is the limit of the sequence (z_n). If there is no such c, the sequence (z_n) *diverges*.

IV.1.3.29. All the results about real sequences except those depending on the ordering of $\mathbb{R}$ (e.g. IV.1.3.18. and the results on monotone sequences) hold for sequences in $\mathbb{C}$, with identical proofs. As an alternate approach, we may reduce the theory of sequences in $\mathbb{C}$ to the theory for $\mathbb{R}$ by considering real and imaginary parts.

IV.1.3.30. Proposition. Let (z_n) be a sequence in $\mathbb{C}$. For each n, write $z_n = x_n + iy_n$ with $x_n, y_n \in \mathbb{R}$. Then (x_n) and (y_n) are real sequences. If $c = a + ib \in \mathbb{C}$ with $a, b \in \mathbb{R}$, then $z_n \to c$ if and only if $x_n \to a$ and $y_n \to b$.

Proof: Suppose $z_n \to c$. Let $\epsilon > 0$. There is an N such that $|z_n - c| < \epsilon$ for all $n \geq N$. Since the real and imaginary parts of a complex number are not larger than the absolute value,

$$|x_n - a| \leq |(x_n - a) + i(y_n - b)| = |z_n - c| < \epsilon$$

for all $n \geq N$, so $x_n \to a$. Similarly, $y_n \to b$.

Conversely, suppose $x_n \to a$ and $y_n \to b$. Let $\epsilon > 0$. Then there is an N_1 such that

$$|x_n - a| < \frac{\epsilon}{2}$$

for all $n \geq N_1$, and there is an N_2 such that

$$|y_n - b| < \frac{\epsilon}{2}$$

for all $n \geq N_2$. Set $N = \max(N_1, N_2)$. Then, if $n \geq N$,

$$|z_n - c| = |(x_n - a) + i(y_n - b)| \leq |x_n - a| + |y_n - b| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and thus $z_n \to c$. 👍

IV.1.4. Infinite Limits

IV.1.5. Monotone Sequences

IV.1.6. Subsequences and Cluster Points of Sequences

"Most" sequences, even most bounded sequences, do not converge.

IV.1.7. The Bolzano-Weierstrass Theorem

The Bolzano-Weierstrass Theorem is one of the key facts of real analysis. It is closely related to the Heine-Borel Theorem and compactness of closed bounded subsets of Euclidean space. There are various proofs; we give one which requires no Choice. See VI.1.3.6. for another proof (which also requires no choice).

IV.1.7.1. THEOREM. [BOLZANO-WEIERSTRASS THEOREM] Every bounded sequence of real numbers has a convergent subsequence.

IV.1.7.2. Since any bounded monotone sequence converges (), the Bolzano-Weierstrass Theorem is an immediate consequence of the next result.

IV.1.7.3. THEOREM. Every sequence of real numbers has a monotone subsequence.

IV.1.7.4. DEFINITION. Let (x_n) be a sequence of real numbers. A term x_k is *peaking* for the sequence if $x_n \leq x_k$ for all $n > k$.

With this definition, we prove IV.1.7.3..

PROOF: Let (x_n) be a sequence of real numbers. We construct a monotone subsequence as follows.

Case 1: Suppose the sequence has infinitely many peaking terms. Then the peaking terms form a subsequence which is monotone decreasing (not necessarily strictly decreasing!)

Case 2: Suppose the sequence has only finitely many peaking terms. If x_k is the last peaking term, set $n_1 = k + 1$ (set $n_1 = 1$ if there are no peaking terms). Suppose $n_1, \ldots, n_m$ have been defined. Let n_{m+1} be the smallest $n > n_m$ for which $x_n > x_{n_m}$ (there will be such an n since x_{n_m} is not peaking). In this way a subsequence is defined recursively, and this subsequence is strictly increasing. 👍

IV.1.8. Limit Superior and Limit Inferior

In this subsection, we define two generalizations of limit for sequences of real numbers, limit superior and limit inferior. Every real sequence has a limit superior and a limit inferior, even sequences which do not converge; in fact, these notions are most interesting for nonconvergent sequences (they both just reduce to the usual limit for convergent sequences).

These notions make essential use of the ordering on $\mathbb{R}$, and thus do not generalize to $\mathbb{C}$, $\mathbb{R}^n$, or general metric spaces.

Limit Superior and Limit Inferior for Bounded Sequences

We first consider only bounded sequences for simplicity. We give the formal definitions, then explain them.

IV.1.8.1. DEFINITION. Let (x_n) be a bounded sequence in $\mathbb{R}$. Define

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty}\left[\sup_{k\geq n} x_k\right] = \inf_n\left[\sup_{k\geq n} x_k\right]$$

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty}\left[\inf_{k\geq n} x_k\right] = \sup_n\left[\inf_{k\geq n} x_k\right]$$

called the *limit superior* and *limit inferior* of (x_n) respectively.

IV.1.8.2. To explain these definitions, first note that the set $\{x_k : k \geq n\}$ is a bounded nonempty set of numbers for any n, hence has a supremum s_n and infimum t_n. As n increases, s_n decreases and t_n increases since the set $\{x_k : k \geq n\}$ decreases. Thus (s_n) is a bounded decreasing (nonincreasing) sequence, hence converges to its infimum; similarly, (t_n) increases and converges to its supremum. These limits are $\limsup_{n\to\infty} x_n$ and $\liminf_{n\to\infty} x_n$ respectively.

The $\limsup_{n\to\infty} x_n$ and $\liminf_{n\to\infty} x_n$ are often written $\overline{\lim}_{n\to\infty} x_n$ and $\underline{\lim}_{n\to\infty} x_n$ respectively. In all notations the "$n \to \infty$" is frequently omitted whenever there is no possibility of confusion.

IV.1.8.3. EXAMPLES. (i) Let $x_n = (-1)^n$. For each n we have $\sup_{k\geq n} x_k = 1$ and $\inf_{k\geq n} x_k = -1$, so $\limsup_{n\to\infty} x_n = 1$ and $\liminf_{n\to\infty} x_n = -1$.

(ii) Let $x_n = (-1)^n\left(1 + \frac{1}{n}\right)$. Then for n odd we have $\sup_{k\geq n} x_k = 1 + \frac{1}{n+1}$ and $\inf_{k\geq n} x_k = -1 - \frac{1}{n}$, and for n even we have $\sup_{k\geq n} x_k = 1 + \frac{1}{n}$ and $\inf_{k\geq n} x_k = -1 - \frac{1}{n+1}$, so $\limsup_{n\to\infty} x_n = 1$ and $\liminf_{n\to\infty} x_n = -1$.

(iii) Let $x_n = \sin n$. Then $-1 < x_n < 1$ for all n. There is a sequence (n_j) in $\mathbb{N}$ with $\sin n_j \to 1$ (), so for any n we have $\sup_{k\geq n} x_k = 1$ and $\limsup_{n\to\infty} x_n = 1$. Similarly, there is a sequence (m_j) in $\mathbb{N}$ with $\sin m_j \to -1$, so for any n we have $\inf_{k\geq n} x_k = -1$ and $\liminf_{n\to\infty} x_n = -1$.

IV.1.8.4. We obviously have $\liminf_{n\to\infty} x_n \leq \limsup_{n\to\infty} x_n$ since $t_n \leq s_n$ for each n. We will show (IV.1.8.9.) that $\limsup_{n\to\infty} x_n$ and $\liminf_{n\to\infty} x_n$ are just the largest and smallest cluster points () of the sequence (x_n), which is usually the best way to think of them. To begin with, we make the following observation:

IV.1.8.5. PROPOSITION. Let (x_n) be a bounded sequence in $\mathbb{R}$, and let $a \in \mathbb{R}$.

(i) If for every $\epsilon > 0$ there are only finitely many x_n satisfying $x_n > a + \epsilon$, then $\limsup_{n\to\infty} x_n \leq a$.

(ii) If for every $\epsilon > 0$ there are infinitely many x_n satisfying $x_n > a - \epsilon$, then $\limsup_{n\to\infty} x_n \geq a$.

(iii) If for every $\epsilon > 0$ there are only finitely many x_n satisfying $x_n < a - \epsilon$, then $\liminf_{n\to\infty} x_n \geq a$.

(iv) If for every $\epsilon > 0$ there are infinitely many x_n satisfying $x_n < a + \epsilon$, then $\liminf_{n\to\infty} x_n \leq a$.

PROOF: (i): Let $\epsilon > 0$, and fix N so that $x_k \leq a+\epsilon$ for all $k \geq N$. Then, for all $n \geq N$, $s_n = \sup_{k\geq n} x_k \leq a+\epsilon$, so $\limsup_{n\to\infty} x_n = \lim_{n\to\infty} s_n \leq a + \epsilon$. Since this is true for every $\epsilon > 0$, $\limsup_{n\to\infty} x_n \leq a$.

(ii) For any $\epsilon > 0$, we have $s_n > a - \epsilon$ for every n, so each s_n satisfies $s_n \geq a$ and thus $\inf_n s_n \geq a$.
(iii) and (iv) are essentially identical with inequalities reversed. 👍

A simple rephrasing is often useful:

IV.1.8.6. COROLLARY. Let (x_n) be a bounded sequence. Then
(i) $\limsup_{n\to\infty} x_n = \inf\{b \in \mathbb{R} : x_n < b \text{ for all sufficiently large } n\}$.
(ii) $\liminf_{n\to\infty} x_n = \sup\{a \in \mathbb{R} : x_n > a \text{ for all sufficiently large } n\}$.

PROOF: For (i), let $B = \inf\{b \in \mathbb{R} : x_n < b \text{ for all sufficiently large } n\}$ and $L = \limsup_{n\to\infty} x_n$. Then we have $L \leq B$ by IV.1.8.5.(i) and $L \geq B$ by IV.1.8.5.(ii). The proof of (ii) is analogous. 👍

The next result gives a useful way of thinking of lim sup and lim inf.

IV.1.8.7. THEOREM. Let (x_n) be a bounded sequence.
(i) Let $L = \limsup_{n\to\infty} x_n$. Then L satisfies

(a) For any $\epsilon > 0$ there are infinitely many n such that $x_n > L - \epsilon$

(b) For any $\epsilon > 0$ there are are only finitely many n such that $x_n > L + \epsilon$

and L is the unique real number with these two properties.
(ii) Let $l = \liminf_{n\to\infty} x_n$. Then l satisfies

(a) For any $\epsilon > 0$ there are infinitely many n such that $x_n < l + \epsilon$

(b) For any $\epsilon > 0$ there are are only finitely many n such that $x_n < l - \epsilon$

and l is the unique real number with these two properties.

PROOF: (i): By IV.1.8.5. (i)–(ii), any number satisfying (ia)–(ib) must be equal to L. On the other hand, if $\epsilon > 0$, then $s_n = \sup_{k\geq n} x_k$ satisfies $s_n \geq L$ for all n, so for each n there must be a $k \geq n$ with $x_k > L - \epsilon$, i.e. $x_k > L - \epsilon$ for infinitely many k. But if $x_k > L + \epsilon$ for infinitely many k, then $s_n > L + \epsilon$ for all n, and thus $L = \inf_n s_n \geq L + \epsilon$, a contradiction.
(ii): Almost identical with inequalities reversed. 👍

IV.1.8.8. COROLLARY. Let (x_n) be a bounded sequence in $\mathbb{R}$. Then $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n$ if and only if $\lim_{n\to\infty} x_n$ exists, in which case

$$\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n = \lim_{n\to\infty} x_n \ .$$

PROOF: If $\lim_{n\to\infty} x_n$ exists and equals L, then L satisfies all four conditions, so

$$L = \liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n \ .$$

Conversely, if $\liminf_{n\to\infty} x_n = \limsup_{n\to\infty} x_n = L$, then L satisfies (ib) and (iib), hence $\lim_{n\to\infty} x_n$ exists and equals L. 👍

Finally, we obtain the characterization of lim sup and lim inf in terms of cluster points:

IV.1.8.9. THEOREM. Let (x_n) be a bounded sequence. Then $\limsup_{n\to\infty} x_n$ is the largest cluster point of (x_n) and $\liminf_{n\to\infty} x_n$ is the smallest cluster point of (x_n).

PROOF: Let $L = \limsup_{n\to\infty} x_n$. If c is a cluster point of (x_n), then for any $\epsilon > 0$ there are infinitely many x_n satisfying $x_n > c - \epsilon$, so $c \le L$ by IV.1.8.5.(ii). On the other hand, for any $\epsilon > 0$ there are infinitely many x_n satisfying $x_n > L - \epsilon$ by IV.1.8.7.(ia), and all but finitely many of these also satisfy $x_n < L + \epsilon$ by IV.1.8.7.(ib), so L is a cluster point of (x_n). The proof for lim inf is analogous. 👍

Tail Dependence

If (x_n) is a bounded sequence, then $\limsup_{n\to\infty} x_n$ and $\liminf_{n\to\infty} x_n$, like $\lim_{n\to\infty} x_n$ if it exists, depend only on the tails of the sequence. We state two precise versions of this, whose simple proofs are left to the reader (),().

IV.1.8.10. PROPOSITION. Let (x_n) be a bounded sequence, and (y_n) any tail sequence of (x_n). Then $\limsup_{n\to\infty} x_n = \limsup_{n\to\infty} y_n$ and $\liminf_{n\to\infty} x_n = \liminf_{n\to\infty} y_n$.

IV.1.8.11. PROPOSITION. Let (x_n) and (y_n) be bounded sequences, with $x_n = y_n$ for all sufficiently large n. Then $\limsup_{n\to\infty} x_n = \limsup_{n\to\infty} y_n$ and $\liminf_{n\to\infty} x_n = \liminf_{n\to\infty} y_n$.

We can improve the last result:

IV.1.8.12. PROPOSITION. Let (x_n) and (y_n) be bounded sequences, with $\lim_{n\to\infty} (x_n - y_n) = 0$. Then $\limsup_{n\to\infty} x_n = \limsup_{n\to\infty} y_n$ and $\liminf_{n\to\infty} x_n = \liminf_{n\to\infty} y_n$.

A direct proof can be given, but the result is an immediate corollary of IV.1.8.17..

Algebraic and Order Properties

We begin with two simple observations, whose proofs are left to the reader. The first generalizes IV.1.8.13.:

IV.1.8.13. PROPOSITION. Let (x_n) and (y_n) be bounded sequences, with $x_n \leq y_n$ for all sufficiently large n. Then $\limsup_{n\to\infty} x_n \leq \limsup_{n\to\infty} y_n$ and $\liminf_{n\to\infty} x_n \leq \liminf_{n\to\infty} y_n$.

The next observation gives a symmetry between lim sup and lim inf:

IV.1.8.14. PROPOSITION. Let (x_n) be a bounded sequence. Then $\limsup_{n\to\infty}(-x_n) = -\liminf_{n\to\infty} x_n$ and $\liminf_{n\to\infty}(-x_n) = -\limsup_{n\to\infty} x_n$.

IV.1.8.15. The limit of a sum of two convergent sequences is the sum of the limits (). However, the analogous statement does not hold for lim sup and lim inf. For example, let $x_n = (-1)^n$ and $y_n = (-1)^{n+1}$ for all n. Then $\limsup_{n\to\infty} x_n = \limsup_{n\to\infty} y_n = 1$ and $\liminf_{n\to\infty} x_n = \liminf_{n\to\infty} y_n = -1$. But $x_n + y_n = 0$ for all n, so

$$\limsup_{n\to\infty}(x_n + y_n) = 0 \neq 2 = \limsup_{n\to\infty} x_n + \limsup_{n\to\infty} y_n$$

$$\liminf_{n\to\infty}(x_n + y_n) = 0 \neq -2 = \liminf_{n\to\infty} x_n + \liminf_{n\to\infty} y_n \ .$$

The reason we should not expect equality is that the n's for which the x_n and y_n are close to the lim sup may well be different. However, we do have inequalities:

IV.1.8.16. THEOREM. Let (x_n) and (y_n) be bounded sequences. Then

$$\liminf_{n\to\infty} x_n + \liminf_{n\to\infty} y_n \leq \liminf_{n\to\infty}(x_n+y_n) \leq \limsup_{n\to\infty} x_n + \liminf_{n\to\infty} y_n \leq \limsup_{n\to\infty}(x_n+y_n) \leq \limsup_{n\to\infty} x_n + \limsup_{n\to\infty} y_n \ .$$

Of course, the x_n and y_n may be interchanged in the middle, i.e. the middle expression can be replaced by $\liminf_{n\to\infty} x_n + \limsup_{n\to\infty} y_n$.

PROOF: For the first inequality, let $l = \liminf_{n\to\infty} x_n$ and $l' = \liminf_{n\to\infty} y_n$. Let $\epsilon > 0$. Then there is an N such that $x_n > l - \frac{\epsilon}{2}$ and $y_n > l' - \frac{\epsilon}{2}$ for all $n \geq N$. Then $x_n + y_n > l + l' - \epsilon$ for all $n \geq N$, so $\liminf_{n\to\infty}(x_n + y_n) \geq l + l'$ by IV.1.8.5.(iii).
For the second inequality, let $L = \limsup_{n\to\infty} x_n$ and $l' = \liminf_{n\to\infty} y_n$. Let $\epsilon > 0$. Then there is an N such that $x_n < L + \frac{\epsilon}{2}$ for all $n \geq N$. For infinitely many such n, $y_n < l' + \frac{\epsilon}{2}$. Thus for infinitely many n, $x_n + y_n < L + l' + \epsilon$, and thus $\liminf_{n\to\infty}(x_n + y_n) \leq L + l'$ by IV.1.8.5.(iv).
The proofs of the other two inequalities are analogous. 👍

We do get equalities if one of the sequences converges:

IV.1.8.17. Corollary. Let (x_n) be a bounded sequence and (y_n) a convergent sequence. Then

$$\limsup_{n\to\infty}(x_n + y_n) = \limsup_{n\to\infty} x_n + \lim_{n\to\infty} y_n$$

$$\liminf_{n\to\infty}(x_n + y_n) = \liminf_{n\to\infty} x_n + \lim_{n\to\infty} y_n \ .$$

Proof: For the first equality, we have

$$\limsup_{n\to\infty} x_n + \lim_{n\to\infty} y_n = \limsup_{n\to\infty} x_n + \liminf_{n\to\infty} y_n \le \limsup_{n\to\infty}(x_n+y_n) \le \limsup_{n\to\infty} x_n + \limsup_{n\to\infty} y_n = \limsup_{n\to\infty} x_n + \lim_{n\to\infty} y_n \ .$$

The second is analogous, using the remark at the end of Theorem IV.1.8.16.. 👍

Unbounded Sequences

IV.1.8.18. The same definition of lim sup and lim inf makes sense, properly interpreted, for unbounded sequences of real numbers and even extended real numbers; in this case the lim sup and/or the lim inf may take the values $\pm\infty$, as well as real-number values. It is simplest to just note the following criteria for when they take infinite values:

IV.1.8.19. Proposition. Let (x_n) be a sequence of real numbers. Then
(i) $\limsup_{n\to\infty} x_n = +\infty$ if and only (x_n) is not bounded above.
(ii) $\liminf_{n\to\infty} x_n = -\infty$ if and only (x_n) is not bounded below.

Proof: (i): $\limsup_{n\to\infty} x_n < +\infty$ if and only if $s_n = \sup_{k\ge n} x_k$ is finite for some n, i.e. a tail of (x_n) is bounded above. But this implies that the entire sequence is bounded above. Conversely, if $x_n \le M$ for all n, then $s_n \le M$ for all n and $\limsup_{n\to\infty} x_n \le M$.
The proof of (ii) is analogous. 👍

IV.1.8.20. Proposition. Let (x_n) be a sequence of real numbers. Then
(i) $\limsup_{n\to\infty} x_n = -\infty$ if and only if $\lim_{n\to\infty} x_n = -\infty$ (in the sense of ()).
(ii) $\liminf_{n\to\infty} x_n = +\infty$ if and only if $\lim_{n\to\infty} x_n = +\infty$.

The simple proof is left to the reader.

IV.1.8.21. All the results about lim sup and lim inf for bounded sequences, suitably interpreted, hold also for unbounded sequences, although some of the proofs need to be modified. Details are left to the reader.

IV.1.9. Cauchy Sequences

Recall the definition of a Cauchy sequence in III.5.2. and the discussion there of why they are interesting and important.

IV.1.9.1. Definition. Let (x_n) be a sequence in $\mathbb{R}$. Then (x_n) is a *Cauchy sequence* if for every $\epsilon > 0$ there is an N such that $|x_n - x_m| < \epsilon$ for all $n, m > N$.

IV.1.9.2. Note that there is no mention of a limit in the definition: it is only regarding the terms of the sequence (informally, they become arbitrarily close together far out in the sequence). But we have:

IV.1.9.3. Proposition. Any convergent sequence in $\mathbb{R}$ is a Cauchy sequence.

Proof: Let (x_n) be a convergent sequence, with $x_n \to a \in \mathbb{R}$. Let $\epsilon > 0$. Then there is an N such that $|x_n - a| < \frac{\epsilon}{2}$ for all $n > N$. Then, if $n, m > N$,

$$|x_n - x_m| = |(x_n - a) + (a - x_m)| \leq |x_n - a| + |a - x_m| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

so (x_n) is a Cauchy sequence. 👍

IV.1.9.4. We thus get a necessary condition for convergence of a sequence in $\mathbb{R}$: it must be a Cauchy sequence. This condition is often called the *Cauchy criterion*.

IV.1.9.5. Cauchy sequences in general metric spaces do not necessarily converge: for example, a Cauchy sequence in $\mathbb{Q}$ need not converge (in $\mathbb{Q}$), cf. IV.1.12.7.; this is the substance of the construction in III.5.2.. However, Cauchy sequences in $\mathbb{R}$ do always converge (IV.1.9.8.), so the Cauchy criterion is both necessary and sufficient for convergence of sequences in $\mathbb{R}$.

IV.1.9.6. Lemma. Every Cauchy sequence in $\mathbb{R}$ is bounded.

Proof: Let (x_n) be a Cauchy sequence. Fix N such that $|x_n - x_m| < 1$ for all $n, m > N$. Let $M = \max\{|x_1|, \ldots, |x_N|, |x_{N+1}| + 1\}$. Then $|x_n| \leq M$ for all n. 👍

IV.1.9.7. Lemma. Let (x_n) be a Cauchy sequence in $\mathbb{R}$. If (x_n) has a convergent subsequence, then the sequence converges (to the same limit, of course).

Proof: Let (x_n) be a Cauchy sequence, and suppose there is a subsequence (x_{n_k}) with $x_{n_k} \to a$. Let $\epsilon > 0$. Fix N such that $|x_n - x_m| < \frac{\epsilon}{2}$ for $n, m > N$, and fix K such that $|x_{n_k} - a| < \frac{\epsilon}{2}$ for $k > K$. Choose $k > K$ such that $n_k > N$. Then, for any $m > N$,

$$|x_m - a| = |(x_m - x_{n_k}) + (x_{n_k} - a)| \leq |x_m - x_{n_k}| + |x_{n_k} - a| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

so $x_m \to a$. 👍

The next theorem is one of the principal facts about the real numbers:

IV.1.9.8. THEOREM. Every Cauchy sequence in $\mathbb{R}$ converges. Thus a sequence in $\mathbb{R}$ converges in $\mathbb{R}$ if and only if it is a Cauchy sequence. (We say $\mathbb{R}$ is *complete*.)

PROOF: This is an immediate consequence of the previous two lemmas and the Bolzano-Weierstrass Theorem: if (x_n) is a Cauchy sequence, it is bounded, hence has a convergent subsequence, hence converges. 👍

IV.1.10. Sequences in $\mathbb{R}^n$

IV.1.11. Sequences in Metric Spaces

IV.1.12. Exercises

IV.1.12.1. Show that in the definition of limit (IV.1.3.6., **??**) it is unimportant whether strict inequality is used in two places: if (x_n) is a sequence in $\mathbb{R}$, then the following conditions are equivalent to the condition in IV.1.3.6..

(i) For every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$|x_n - a| \leq \epsilon$$

for all $n \geq N$. [For one direction, decrease ϵ before finding N.]

(ii) For every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$|x_n - a| < \epsilon$$

for all $n > N$. [For one direction, replace N by $N + 1$.]

(iii) For every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$|x_n - a| \leq \epsilon$$

for all $n > N$.

(iv) Show that the condition $N \in \mathbb{N}$ can be replaced by $N \in \mathbb{R}$ without changing the definition of limit, i.e. that the following condition is equivalent to the condition in IV.1.3.6.:

For every $\epsilon > 0$, there is an $N \in \mathbb{R}$ such that

$$|x_n - a| < \epsilon$$

for all $n \in \mathbb{N}$, $n \geq N$. [For one direction, use the Archimedean property of $\mathbb{R}$ ().]

Do the same for the conditions in (i)–(iii). Show analogous results for limits of sequences in metric spaces.

IV.1.12.2. Give an alternate version of the proof of IV.1.3.13. by direct argument, as follows. Suppose $x_n \to a$ and $x_n \to b$. Let $\epsilon > 0$.

(a) Find N_1 and N_2 so that $|a_n - a| < \frac{\epsilon}{2}$ for all $n \geq N_1$ and $|x_n - b| < \frac{\epsilon}{2}$ for all $n \geq N_2$.

(b) If $n > \max(N_1, N_2)$, use the triangle inequality to conclude that

$$|a - b| \leq |a - a_n| + |a_n - b| < \epsilon \ .$$

(c) Since this is true for all $\epsilon > 0$, conclude that $|a - b| = 0$.

IV.1.12.3. Let (x_n) and (y_n) be sequences converging to the same limit a. Alternate or shuffle the x_n and y_n into a single sequence (z_n) by taking $z_n = x_m$ if $n = 2m - 1$ is odd, and $z_n = y_m$ if $n = 2m$ is even. Give two proofs that $z_n \to a$, one directly from the definition of limit and one from ().

IV.1.12.4. Let (x_n) be a sequence. If (y_n) is a rearrangement of (x_n), i.e. $y_n = x_{\phi(n)}$ for a bijection (permutation) $\phi : \mathbb{N} \to \mathbb{N}$, show that (y_n) converges if and only if (x_n) converges, and $\lim\limits_{n\to\infty} y_n = \lim\limits_{n\to\infty} x_n$.

IV.1.12.5. Prove the following theorem ([**?**]; cf. []):

THEOREM.

(i) Let (α_n) be a sequence of positive numbers satisfying $\alpha_{n+m} \leq \alpha_n \alpha_m$ for all n and m. Set $\beta_n = \alpha_n^{1/n}$. Then $\lim\limits_{n\to\infty} \beta_n$ exists (as a real number) and

$$\lim_{n\to\infty} \beta_n = \inf_n \beta_n \ .$$

(ii) Let (a_n) be a sequence of real numbers satisfying $a_{n+m} \leq a_n + a_m$ for all n and m. Set $b_n = \frac{a_n}{n}$. Then $\lim\limits_{n\to\infty} b_n$ exists as a real number or $-\infty$, and

$$\lim_{n\to\infty} b_n = \inf_n b_n \ .$$

(iii) Let (a_n) be a sequence of real numbers satisfying $a_{n+m} \geq a_n + a_m$ for all n and m. Set $b_n = \frac{a_n}{n}$. Then $\lim\limits_{n\to\infty} b_n$ exists as a real number or $+\infty$, and

$$\lim_{n\to\infty} b_n = \sup_n b_n \ .$$

(a) Show that (i), (ii), and (iii) are equivalent. So it suffices to prove (iii).

(b) In (iii), replacing a_n by $a_n - na_1$ for all n, it suffices to assume $a_1 = 0$.

(c) For any $k, m \in \mathbb{N}$, $a_{km} \geq ka_m$, and hence $b_{km} \geq b_m$. In particular, $a_n \geq na_1 = 0$ for all n, and hence $b_n \geq 0$ for all n.

(d) Fix $m \in \mathbb{N}$. For $k \in \mathbb{N}$ and $1 \leq r \leq m$,

$$a_{km+r} \geq a_{km} + a_r \geq a_{km}$$

and thus

$$b_{km+r} = \frac{a_{km+r}}{km+r} \geq \frac{a_{km}}{(k+1)m} = \frac{k}{k+1} \cdot b_{km} \geq \frac{k}{k+1} \cdot b_m \ .$$

(e) Conclude that

$$\liminf_{n\to\infty} b_n \geq \sup_n b_n \ .$$

IV.1.12.6. Let (a_n) and (b_n) be sequences of real numbers, with (b_n) a strictly increasing sequence of nonzero numbers with $\lim_{n\to\infty} b_n = +\infty$. Suppose

$$\lim_{n\to\infty} \frac{a_{n+1} - a_n}{b_{n+1} - b_n}$$

exists (in either the ordinary or extended sense) and equals L. Show that

$$\lim_{n\to\infty} \frac{a_n}{b_n}$$

also exists (in the same sense) and equals L. [Show that for every $\epsilon > 0$ there is an m such that

$$(L - \epsilon)(b_n - b_m) < a_n - a_m < (L + \epsilon)(b_n - b_m)$$

for all $n > m$ by adding a telescoping sum.]

This result can be regarded as a discrete version of l'Hôpital's Rule Version 2d′ (V.9.2.12.).

IV.1.12.7. Define a sequence (x_n) in $\mathbb{Q}$ recursively by setting $x_1 = 1$ and

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right) .$$

(a) Show that (x_n) is a Cauchy sequence in $\mathbb{Q}$.

(b) Suppose $x_n \to x \in \mathbb{Q}$. Show that $x^2 = 2$. Conclude that (x_n) does not converge in $\mathbb{Q}$.

The sequence (x_n) is the sequence of approximations to $\sqrt{2}$ using Newton's Method () with $f(x) = x^2 - 2$ and $x_1 = 1$.

IV.1.12.8. [CJ99] For each n set $x_n = \sqrt[n]{n} = n^{1/n}$. Show that $\lim_{n\to\infty} x_n = 1$, as follows. [See V.9.3.3.(iii) for an alternate proof.]

(a) If $h > 0$ and $n \in \mathbb{N}$, $n \geq 2$, give two proofs that $(1 + h)^n > 1 + nh$, one using the Binomial Theorem (V.12.5.3.) and a direct one by induction on n. [See V.5.3.1. for a more general result.]

(b) Set $y_n = \sqrt{x_n}$. Then $y_n^n = \sqrt{n}$. Set $y_n = 1 + h_n$ with $h_n > 0$ for $n \geq 2$.

(c) Using (a), obtain for $n \geq 2$

$$\sqrt{n} = y_n^n = (1 + h_n)^n > 1 + nh_n$$

$$0 < h_n < \frac{\sqrt{n} - 1}{n} = \frac{1}{\sqrt{n}} - \frac{1}{n} .$$

(d) Conclude that

$$\lim_{n\to\infty} h_n = 0, \quad \lim_{n\to\infty} y_n = 1 = \lim_{n\to\infty} x_n .$$

IV.2. Infinite Series

> "No part of Mathematics suffers more from the triviality of its initial presentation than the great subject of series."
>
> *Alfred North Whitehead*[1]

Infinite series (first of numbers, then of functions) form one of the most important and useful topics in analysis. There are considerable subtleties in the theory, however, which are not immediately apparent.

An infinite series is an expression of the form

$$\sum_{k=1}^{\infty} a_k$$

(or slight variations of this general form, e.g. we may start the summation at 0 or another integer), where the a_k, called the *terms* of the series, are numbers or (later) functions.

The summation notation of infinite series suggests that the series represents a sum of numbers (or functions), hence has a numerical value. But an infinite series is, at least at first, just a formal expression and we do not necessarily associate to it a numerical value. We can associate a numerical "sum" to many series, but not all (the ones which have an associated sum are called *convergent*, and the ones which do not are called *divergent*). These "sums," however, do not always behave like finite sums. The following principle must be kept in mind at all times in this subject:

First Fundamental Principle of Infinite Series: An infinite series is NOT a finite sum and cannot be treated like one!

The following chart lists the main differences between infinite series and finite sums:

Finite Sum	Infinite Series
Always represents a number	May or may not converge
Purely algebraic expression	Requires limits and analysis to define sum
Associative and commutative laws hold	Laws of arithmetic do not hold in general

A lot of the theory of infinite series is devoted to establishing that under certain conditions, infinite series can be manipulated and evaluated roughly like finite sums. But operations on infinite series which would be routine or trivial for finite sums must always be carefully justified using the results of the theory. In *finding* results or formulas, it can be very useful to manipulate infinite series formally as though they were finite sums, but formulas obtained in this way cannot be considered valid unless justified by careful applications of theorems about infinite series; some such formulas turn out to be correct and others don't.

IV.2.1. Partial Sums and Convergence

IV.2.1.1. Suppose $\sum_{k=1}^{\infty} a_k$ is an infinite series. We can associate to this series two sequences. One obvious sequence is the sequence (a_k) of terms. The other, much more important, is the sequence (s_n) of partial sums:

[1] [?, p. 144].

IV.2.1.2. DEFINITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, and $n \in \mathbb{N}$. The *n'th partial sum* of the series is the finite sum

$$s_n = \sum_{k=1}^{n} a_k \ .$$

IV.2.1.3. Conversely, if (s_n) is any sequence of real numbers, there is a unique infinite series $\sum_{k=1}^{\infty} a_k$ such that (s_n) is the sequence of partial sums: set $a_1 = s_1$ and $a_k = s_k - s_{k-1}$ for $k > 1$. Thus in a sense the theory of infinite series is identical to the theory of sequences, although we usually take a somewhat different point of view. (This correspondence can be regarded as a discrete version of the Fundamental Theorem of Calculus ().)

Since the partial sums of an infinite series are finite sums, the usual rules of arithmetic and algebra can be used with them. We can also use facts and results about sequences, applied to the sequence of partial sums, to analyze series. In particular, we can define convergence for an infinite series:

IV.2.1.4. DEFINITION. The infinite series $\sum_{k=1}^{\infty} a_k$ *converges*, with *sum* $s \in \mathbb{R}$, if the sequence (s_n) of partial sums converges to s. If the sequence (s_n) does not converge, the series is said to *diverge*.

If the series $\sum_{k=1}^{\infty} a_k$ converges with sum s, we will often write

$$\sum_{k=1}^{\infty} a_k = s$$

and regard the infinite series as representing the number s. Of course, by the uniqueness of limits for convergent sequences (IV.1.3.13.), the sum of a convergent infinite series is uniquely determined. It is useful to think of the partial sums of a convergent series as approximations to the sum.

We reiterate the First Fundamental Principle: use of the word "sum" for a convergent infinite series does *not* mean that the "sum" is an actual sum in the algebraic sense.

IV.2.1.5. It is sometimes useful, if imprecise, to write an infinite series $\sum_{k=1}^{\infty} a_k$, especially a convergent series, as

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + \cdots \ .$$

We will mostly avoid this notation since it is imprecise and can be misleading, although with due care it can be helpful in working with the series. When writing an infinite series this way one must be especially careful not to treat it like a finite sum; even great mathematicians have occasionally been misled into errors by confusing such an expression with a finite sum and applying unjustifiable algebraic operations.

A sequence converges if and only if all "tail" sequences converge (). A similar statement holds for series:

IV.2.1.6. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. The following are equivalent:

(i) The series $\sum_{k=1}^{\infty} a_k$ converges.

(ii) For some $n \in \mathbb{N}$, the tail series $\sum_{k=n+1}^{\infty} a_k$ converges.

(iii) For every $n \in \mathbb{N}$, the tail series $\sum_{k=n+1}^{\infty} a_k$ converges.

PROOF: (i) $\Rightarrow$ (iii): Suppose the series converges, with sum s, and fix $n \in \mathbb{N}$. Let s_m be the m'th partial sum of the original series, and t_m the m'th partial sum of the series

$$\sum_{k=n+1}^{\infty} a_k \ .$$

The relationship is that $t_m = s_{n+m} - s_n$ for any m. We have that

$$\lim_{m\to\infty} t_m = \left[\lim_{m\to\infty} s_{n+m}\right] - s_n = s - s_n$$

so the series

$$\sum_{k=n+1}^{\infty} a_k$$

converges with sum $s - s_n$.
(iii) $\Rightarrow$ (ii) is trivial.
(ii) $\Rightarrow$ (i): If the tail series converges for some $n \in \mathbb{N}$, with sum t, we have

$$\lim_{m\to\infty} s_{n+m} = \left[\lim_{m\to\infty} t_m\right] + s_n = t + s_n$$

so the original series converges with sum $t + s_n$. 👍

IV.2.1.7. Note, however, that there is a difference between sequences and series: the limit of a tail of a sequence is the same as the limit of the whole sequence (i.e. the limit of a sequence depends only on the tail terms). But the sum of a tail of a convergent infinite series is not generally the same as the sum of the whole series, i.e. although the *convergence* of an infinite series depends only on tail behavior, the *sum* of the series depends on all the terms. ✋

A useful first observation is that for an infinite series to converge, it is necessary for the terms to approach zero (the key observation is that $s_{n+1} - s_n = a_{n+1}$ for any n):

IV.2.1.8. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. If the series converges, then

$$\lim_{k\to\infty} a_k = 0 \ .$$

PROOF: Let $\epsilon > 0$. Since $\lim_{n\to\infty} s_n$ exists and equals, say, s, there is an N such that for all $n \geq N$,

$$|s_n - s| < \frac{\epsilon}{2} \ .$$

If $n \geq N$, we also have $|s_{n+1} - s| < \frac{\epsilon}{2}$. So for all $n \geq N$ we have

$$|a_{n+1}| = |s_{n+1} - s_n| = |(s_{n+1} - s) + (s - s_n)| \leq |s_{n+1} - s| + |s - s_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and thus $\lim_{n\to\infty} a_n = 0$. 

IV.2.1.9. It is important to note that this result goes only one way: if the terms of an infinite series approach zero, the series may or may not converge. A simple example is $\sum_{k=1}^{\infty} \frac{1}{\sqrt{k}}$; we have $s_n \geq n \cdot \frac{1}{\sqrt{n}} = \sqrt{n}$ for each n, so the partial sums are unbounded and cannot converge. We will see many other examples later. 

The result IV.2.1.8. is commonly used in its contrapositive form:

IV.2.1.10. COROLLARY. [DIVERGENCE CRITERION] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. If the sequence (a_k) of terms does not converge to zero, then the infinite series diverges.

We can give an important refinement of IV.2.1.8., using the Cauchy Criterion (). The key observation here is that if $n < m$, then

$$s_m - s_n = \sum_{k=n+1}^{m} a_k \ .$$

IV.2.1.11. THEOREM. [CAUCHY CRITERION FOR SERIES] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. Then the series converges if and only if, for every $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that

$$\left| \sum_{k=n+1}^{m} a_k \right| < \epsilon$$

whenever $N \leq n < m$.

In other words, an infinite series converges if and only if the sum of arbitrarily many consecutive terms is small, provided only that the sum starts sufficiently far out in the sequence of terms. This result includes IV.2.1.8., which is the special case where $m = n + 1$.

As a variation, we have a version involving sums of "tails" of the series which refines part of IV.2.1.6.:

IV.2.1.12. Corollary. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. If the series converges, then for every $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that

$$\sum_{k=n+1}^{\infty} a_k$$

(which converges for all $n \in \mathbb{N}$ by IV.2.1.6.) satisfies

$$\left| \sum_{k=n+1}^{\infty} a_k \right| < \epsilon$$

whenever $n \geq N$.

Proof: Suppose the series converges. Let $\epsilon > 0$, and fix $N \in \mathbb{N}$ such that

$$\left| \sum_{k=n+1}^{m} a_k \right| < \frac{\epsilon}{2}$$

whenever $N \leq n < m$. We have, for $n \geq N$,

$$\left| \sum_{k=n+1}^{\infty} a_k \right| = \lim_{m \to \infty} \left| \sum_{k=n+1}^{m} a_k \right| \leq \frac{\epsilon}{2} < \epsilon \ .$$



This corollary can also be proved directly with a slight extension of the proof of IV.2.1.6..

Geometric Series

One of the most basic examples of an infinite series is the geometric series.

IV.2.1.13. Definition. Let $x \in \mathbb{R}$. The *geometric series* with ratio x is the infinite series

$$\sum_{k=1}^{\infty} x^k \ .$$

Eventually, we will want to regard x as a variable to obtain a function (power series), but for now x will be just regarded as a fixed number.

IV.2.1.14. PROPOSITION. The geometric series with ratio x converges if and only if $|x| < 1$. In this case, we have

$$\sum_{k=1}^{\infty} x^k = \frac{x}{1-x} \ .$$

PROOF: If $|x| \geq 1$, the series diverges by the Divergence Criterion since the terms do not go to zero. For any x, we have, for any $n \in \mathbb{N}$,

$$(1-x)(x + x^2 + \cdots + x^n) = x - x^{n+1}$$

by a simple calculation (or, more precisely, a simple proof by induction), so if s_n is the n'th partial sum of the geometric series with ratio x, we have

$$(1-x)s_n = x - x^{n+1}$$

and thus, for $x \neq 1$, we have

$$s_n = \frac{x}{1-x} - \frac{x^{n+1}}{1-x}$$

so, if $|x| < 1$, we have

$$\lim_{n\to\infty} s_n = \frac{x}{1-x} - \lim_{n\to\infty} \frac{x^{n+1}}{1-x} = \frac{x}{1-x}$$

and thus the series converges with sum $\frac{x}{1-x}$. 👍

IV.2.1.15. REMARKS. (i) This is one of the few infinite series where it is possible to find a simple formula for the exact sum. In many cases, it can be shown easily from one of the convergence tests that the series converges, but the exact sum is not easy (and often even impossible) to find.

(ii) The closely related series

$$\sum_{k=0}^{\infty} x^k$$

is often called the geometric series instead. This series just has one more term at the beginning, equal to 1; thus it also converges if and only if $|x| < 1$, and the sum is

$$1 + \frac{x}{1-x} = \frac{(1-x)+x}{1-x} = \frac{1}{1-x} \ .$$

Linear Combinations of Convergent Series

There is some nice algebraic behavior for convergent infinite series:

IV.2.1.16. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be convergent infinite series. Then the series

$$\sum_{k=1}^{\infty}(a_k + b_k)$$

converges, and

$$\sum_{k=1}^{\infty}(a_k + b_k) = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k \ .$$

PROOF: Let s_n and t_n be the n'th partial sum of the series $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ respectively. Then

$$\sum_{k=1}^{n}(a_k + b_k) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k = s_n + t_n$$

for any n by usual rules of algebra for finite sums. Then, using (), we obtain

$$\lim_{n\to\infty}\left[\sum_{k=1}^{n}(a_k + b_k)\right] = \lim_{n\to\infty}(s_n + t_n) = \lim_{n\to\infty} s_n + \lim_{n\to\infty} t_n = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k$$

and thus the series $\sum_{k=1}^{\infty}(a_k + b_k)$ converges and has the desired sum. 👍

IV.2.1.17. Note that this result also goes only one way: $\sum_{k=1}^{\infty}(a_k + b_k)$ can converge even if the two individual series diverge. A rather trivial example is to take $a_k = 1$ and $b_k = -1$ for all k. ✋

Similarly, we have a "distributive law" for infinite series:

IV.2.1.18. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series and $c \in \mathbb{R}$. If the series $\sum_{k=1}^{\infty} a_k$ converges, then the series

$$\sum_{k=1}^{\infty} ca_k$$

converges, and

$$\sum_{k=1}^{\infty} ca_k = c\left[\sum_{k=1}^{\infty} a_k\right] \ .$$

PROOF: If s_n is the n'th partial sum of $\sum_{k=1}^{\infty} a_k$, then for each n we have

$$\sum_{k=1}^{n} ca_k = c\left[\sum_{k=1}^{n} a_k\right] = cs_n$$

by the distributive law for finite sums. Thus, by (), we have

$$\lim_{n\to\infty}\left[\sum_{k=1}^{n} ca_k\right] = \lim_{n\to\infty} cs_n = c\left[\lim_{n\to\infty} s_n\right] = c\left[\sum_{k=1}^{\infty} a_k\right]$$

and the result follows. 👍

IV.2.1.19. Note that there is no corresponding result for term-by-term products; there is no simple relationship between the partial sums of $\sum_{k=1}^{\infty} a_k b_k$ and those of $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$. In fact, if $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ converge, the series $\sum_{k=1}^{\infty} a_k b_k$ may or may not converge (). (However, see Exercise (); see Exercise () for a somewhat related result.) ✋

Grouping and Contraction of Series

There is a limited sense in which the associative law applies in a convergent series.

IV.2.1.20. DEFINITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. A *grouping* or *contraction* of the series is an infinite series $\sum_{k=1}^{\infty} b_k$, where

$$b_k = \sum_{j=m_{k-1}+1}^{m_k} a_j$$

for some strictly increasing sequence $0 = m_0 < m_1 < m_2 < \cdots$ in $\mathbb{N} \cup \{0\}$.

Informally, the series $\sum_{k=1}^{\infty} b_k$ is obtained by inserting parentheses

$$(a_1 + \cdots + a_{m_1}) + (a_{m_1+1} + \cdots + a_{m_2}) + \cdots + (a_{m_k+1} + \cdots + a_{m_{k+1}}) + \cdots$$

Note that all the terms in the original series appear exactly once, and in the same order; they are simply grouped together.

IV.2.1.21. Proposition. Let $\sum_{k=1}^{\infty} a_k$ be a convergent infinite series. Then any grouping or contraction of the series also converges, and has the same sum.

Proof: Let $\sum_{k=1}^{\infty} b_k$ be a contraction of the series. Then every partial sum of the contraction is also a partial sum of the original series; hence the sequence of partial sums of $\sum_{k=1}^{\infty} b_k$ is a subsequence of the sequence of partial sums of $\sum_{k=1}^{\infty} a_k$, and thus converges to the same limit. 👍

IV.2.1.22. This result does not apply to divergent series. A famous example concerns the series $\sum_{k=1}^{\infty}(-1)^{k+1}$. We can contract this divergent series to ✋

$$(1-1)+(1-1)+(1-1)+\cdots$$

to obtain a convergent series with sum 0, or to

$$1+(-1+1)+(-1+1)+(-1+1)+\cdots$$

to obtain a convergent series with sum 1.

Before convergence of infinite series was well understood, several great mathematicians argued over what should be the correct sum of this series; some, including Leibniz and Euler, claimed that the true sum should be $\frac{1}{2}$, since among other things this is the average of the two "equally reasonable" groupings above (they had better reasons too; cf. ()). The Cesaro sum () of this series is $\frac{1}{2}$.

IV.2.1.23. Applying IV.2.1.21. repeatedly, we may insert parentheses at will in a convergent series (not changing the order of terms) without affecting convergence or changing the sum, so long as any two pairs of inserted parentheses are either nested or disjoint, i.e. only the patterns ([]), [()], ()[], or []() are allowed, not ([)] or [([]) (which do not make sense anyway).

If the order of the terms in a convergent series is changed, however, or if some (infinitely many) terms are deleted, the new series may not converge; see (), (), ...

IV.2.2. Nonnegative Series

IV.2.2.1. A *nonnegative series* is an infinite series in which all the terms are nonnegative. Nonnegative series behave somewhat more nicely than general infinite series, and criteria for convergence are considerably simpler to state. In this subsection and the next two, we examine convergence of nonnegative series and develop the standard tests for convergence.

IV.2.2.2. Of course, as for any infinite series, the terms of a convergent nonnegative infinite series must converge to zero. But a nonnegative series whose terms go to zero does not necessarily converge (IV.2.1.9.). In fact, we will develop a set of results that make precise the following informal principle:

Fundamental Principle of Nonnegative Series: A nonnegative series converges if and only if the terms of the series approach zero sufficiently rapidly.

IV.2.2.3. The important fact about the sequence of partial sums of a nonnegative series is that it is nondecreasing. Since a nondecreasing sequence converges if and only if it is bounded (), we obtain:

IV.2.2.4. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be a nonnegative infinite series. Then the series converges if and only if the sequence of partial sums is bounded.

IV.2.2.5. This result can fail for series which are not nonnegative. For example, in the geometric series

$$\sum_{k=1}^{\infty} (-1)^k$$

the partial sum s_n is -1 for n odd and 0 for n even, hence the sequence of partial sums is bounded. But the series does not converge by the Divergence Criterion.

Indeed, some authors distinguish between infinite series which are "divergent" (or "definitely divergent") in the sense that the limit of the partial sums is $+\infty$ or $-\infty$, and series which are "indefinitely divergent," or "oscillatory," in the sense that the sequence of partial sums has no limit in the extended real line.

The Comparison Test

The first test we develop for convergence of nonnegative series, which makes the Fundamental Principle more precise, is the Comparison Test. This test will, however, only have teeth once we develop a reasonable supply of "test series" to compare new series to. (So far the only examples we have are geometric series.)

IV.2.2.6. THEOREM. [COMPARISON TEST] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be nonnegative infinite series. Suppose $a_k \leq b_k$ for all k. Then

(i) If $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\sum_{k=1}^{\infty} a_k$ diverges, then $\sum_{k=1}^{\infty} b_k$ diverges.

PROOF: Statement (ii) is just the contrapositive of statement (i), so we need only prove (i). For $n \in \mathbb{N}$, let s_n and t_n be the n'th partial sum of $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ respectively. Then $s_n \leq t_n$ for all n. Since the sequence (t_n) converges, it is bounded (); any bound for (t_n) is also a bound for (s_n), hence (s_n) is also bounded. Thus $\sum_{k=1}^{\infty} a_k$ converges by IV.2.2.4.. 👍

We can give a slight extension in which finitely many terms can be ignored in each sequence:

IV.2.2.7. COROLLARY. [EXTENDED COMPARISON TEST] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be nonnegative infinite series. Suppose there are $p, q \in \mathbb{N}$ such that $a_{p+k} \leq b_{q+k}$ for all k (this will be satisfied, for example, if there is an $N \in \mathbb{N}$ for which $a_k \leq b_k$ for all $k > N$, by taking $p = q = N$). Then

(i) If $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\sum_{k=1}^{\infty} a_k$ diverges, then $\sum_{k=1}^{\infty} b_k$ diverges.

PROOF: Apply the Comparison Test to the tail series $\sum_{k=p+1}^{\infty} a_k$ and $\sum_{k=q+1}^{\infty} b_k$, and apply IV.2.1.6.. 👍

IV.2.2.8. The substance of these results is that a nonnegative series whose terms go to zero "more rapidly" than those of a convergent nonnegative series also converges, and a nonnegative series whose terms go to zero "more slowly" than those of a divergent nonnegative series also diverges.

The Harmonic Series

IV.2.2.9. Perhaps the most important single infinite series is the *harmonic series*

$$\sum_{k=1}^{\infty} \frac{1}{k} .$$

IV.2.2.10. It is not obvious whether the harmonic series converges. Its terms do go to zero, so the Divergence Criterion does not rule out convergence but does not guarantee it either. If we do some numerical experiments calculating some partial sums we obtain

$$s_{10} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{10} = 2.9289\cdots$$

$$s_{50} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{50} = 4.499\cdots$$
$$s_{100} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{100} = 5.187\cdots$$

and the partial sums might appear to be approaching a limit. If we go on and compute

$$s_{1000} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{1000} = 7.485\cdots$$
$$s_{10000} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{10000} = 9.787\cdots$$

this conclusion seems less obvious. Are the partial sums bounded?

IV.2.2.11. THEOREM. The harmonic series diverges, i.e. the partial sums are unbounded.

PROOF: We will give a quick proof later using the Integral Test, but here we give an alternate elementary proof based on the Comparison Test. Define a new series $\sum_{k=1}^{\infty} a_k$ by setting $a_1 = 1$ and $a_k = 2^{-r}$ for $2^{r-1} + 1 \le k \le 2^r$, $r \in \mathbb{N}$. Note that $a_k \le \frac{1}{k}$ for all k. If s_n is the n'th partial sum of the series $\sum_{k=1}^{\infty} a_k$, then for $m \in \mathbb{N}$ and $n = 2^m$ we have

$$s_n = s_{2^m} = \sum_{k=1}^{2^m} a_k = 1 + \sum_{r=1}^{m} \sum_{k=2^{r-1}+1}^{2^r} a_k = 1 + \sum_{r=1}^{m} [2^{r-1} \cdot 2^{-r}] = 1 + \sum_{r=1}^{m} \frac{1}{2} = 1 + \frac{m}{2} \ .$$

These are unbounded as $m \to \infty$, so $\sum_{k=1}^{\infty} a_k$ diverges, and hence the harmonic series also diverges by the Comparison Test. 👍

IV.2.2.12. The argument of this proof was discovered by the French mathematician NICOLE ORESME in the fourteenth century, and is one of the most significant pieces of mathematics surviving from medieval Europe. (ORESME also developed the theory of exponents, and was the first to draw graphs of functions.) Despite this argument, convergence of infinite series and its importance were not widely understood by mathematicians until the nineteenth century.

IV.2.2.13. The argument indicates that the partial sums of the harmonic series grow only logarithmically in n, and thus go to $+\infty$ very slowly. This conclusion will follow more carefully from an application of the Integral Test. In fact, it takes 15,092,688,622,113,788,323,693,563,264,538,101,449,859,497 $\approx 1.5 \times 10^{43}$ terms to make the partial sum exceed 100, for example [**?**]. For more information about the partial sums of the harmonic series, see the website http://mathworld.wolfram.com/HarmonicNumber.html.

The divergence of the harmonic series leads to some apparent paradoxes. For example, a stable stack of identical books, or bricks, etc., can be made (in principle) in which the top book overhangs the bottom one by an arbitrary amount: overhang the $(k-1)$'th book from the k'th (counting down from the top) by $\frac{1}{k}$ times the width of the book. For other "paradoxes" involving the harmonic series, see ().

The p-Series

The harmonic series is a member of an important class of infinite series to which a similar argument applies, the p-series:

IV.2.2.14. DEFINITION. Let $p \in \mathbb{R}$. The *p-series* with exponent p is the infinite series

$$\sum_{k=1}^{\infty} k^{-p} = \sum_{k=1}^{\infty} \frac{1}{k^p} \ .$$

The p-series with exponent p is a nonnegative series. The only interesting case is when $p > 0$, since for $p \leq 0$ the terms of the series do not approach zero, so the series automatically diverges.

IV.2.2.15. The harmonic series is the p-series with exponent 1. We know this series diverges (IV.2.2.11.), and by comparison we have that a p-series for $p < 1$ also diverges. We now show that a p-series with $p > 1$ converges. We will give another proof of these facts using the Integral Test (IV.2.3.5.).

IV.2.2.16. THEOREM. The p-series with exponent p converges if and only if $p > 1$.

PROOF: We have already done the cases $p \leq 1$, so assume $p > 1$. Define a new series $\sum_{k=1}^{\infty} a_k$ by setting $a_1 = 1$ and $a_k = 2^{-pr} = [2^{-r}]^p$ for $2^r \leq k \leq 2^{r+1} - 1$, $r \in \mathbb{N}$. Note that $a_k \geq \frac{1}{k^p}$ for all k. If s_n is the n'th partial sum of the series $\sum_{k=1}^{\infty} a_k$, then for $m \in \mathbb{N}$ and $n = 2^{m+1} - 1$ we have

$$s_n = s_{2^{m+1}-1} = \sum_{k=1}^{2^{m+1}-1} a_k = 1 + \sum_{r=1}^{m} \sum_{k=2^r}^{2^{r+1}-1} a_k = 1 + \sum_{r=1}^{m} [2^r \cdot 2^{-pr}] = 1 + \sum_{r=1}^{m} [2^{1-p}]^r \ .$$

These are partial sums of a geometric series with ratio $2^{1-p} < 1$, so these s_n converge. So the sequence of partial sums of this series is increasing and has a convergent subsequence and thus converges, and hence the p-series converges by the Comparison Test. 👍

The Limit Comparison Test

There is a variation of the Comparison Test which is often simpler to use:

IV.2.2.17. THEOREM. [LIMIT COMPARISON TEST] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be infinite series with positive terms. Suppose

$$L = \lim_{k \to \infty} \frac{a_k}{b_k}$$

exists as a (nonnegative) extended real number. Then:

(i) If $0 \leq L < +\infty$ and $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $0 < L \leq +\infty$ and $\sum_{k=1}^{\infty} b_k$ diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

PROOF: For (i), if $L < c < +\infty$, then we have $a_k < cb_k$ for all sufficiently large k. We have that $\sum_{k=1}^{\infty} cb_k$ converges, so $\sum_{k=1}^{\infty} a_k$ converges by the Extended Comparison Test. The proof of (ii) is almost identical: if $0 < d < L$, then $\sum_{k=1}^{\infty} db_k$ diverges, so $\sum_{k=1}^{\infty} a_k$ diverges. 👍

A refined version using lim sup and lim inf holds with the same proof:

IV.2.2.18. THEOREM. [LIMIT COMPARISON TEST, GENERAL VERSION] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be infinite series with positive terms.

(i) If $\limsup_{k\to\infty} \frac{a_k}{b_k} < +\infty$ and $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\liminf_{k\to\infty} \frac{a_k}{b_k} > 0$ and $\sum_{k=1}^{\infty} b_k$ diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

The next class of examples gives one of the most important applications of the Limit Comparison Test:

IV.2.2.19. EXAMPLE. Let $f(k)$ and $g(k)$ be functions which are finite sums of terms of the form $c_r k^r$ for $r \in \mathbb{R}$. (f and g need not be polynomials, i.e. the powers need not be nonnegative integers.) Suppose $g(k) \neq 0$ for all k. Let $c_p k^p$ be the term in f with largest power, and $d_q k^q$ the term in g with largest power (c_p and d_q are nonzero). Then:

(i) If $q - p > 1$, the series $\sum_{k=1}^{\infty} \frac{f(k)}{g(k)}$ converges.

(ii) If $q - p \leq 1$, the series $\sum_{k=1}^{\infty} \frac{f(k)}{g(k)}$ diverges.

For we can do a limit comparison test with the p-series $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k^{q-p}}$ since

$$\lim_{k\to\infty} \frac{k^{q-p} f(k)}{g(k)} = \frac{c_p}{d_q}$$

which is a nonzero finite number. (If $\frac{c_p}{d_q}$ is negative, first multiply the series by -1. The terms of the series will then eventually be strictly positive.)

The Border Between Convergence and Divergence

IV.2.2.20. The set of nonnegative sequences (a_k) converging to zero splits into two disjoint sets, those for which $\displaystyle\sum_{k=1}^{\infty} a_k$ converges and those for which it diverges. Any sequence dominated by a sequence in the first group is in the first group, and any sequence which dominates one in the second is in the second. Is there some identifiable border between the sets? Specifically, we can ask:

(i) Is there a sequence (a_k) such that $a_k \to 0$ most slowly among terms of convergent series, i.e. a nonnegative series $\displaystyle\sum_{k=1}^{\infty} b_k$ converges if and only if $\frac{b_k}{a_k}$ is bounded?

(ii) Is there a sequence (a_k) such that $a_k \to 0$ most rapidly among terms of divergent series, i.e. a nonnegative series $\displaystyle\sum_{k=1}^{\infty} b_k$ diverges if and only if $\frac{a_k}{b_k}$ is bounded?

(We could restrict to nonincreasing sequences to sharpen the questions and eliminate some obvious pathologies; the terms of a nonnegative series can be reordered without affecting convergence – see IV.2.8.4..)

Both questions have negative answers: for any convergent nonnegative series there is another convergent nonnegative series whose terms approach zero more slowly (Exercise IV.2.16.2.), and for every divergent nonnegative series there is a divergent nonnegative series whose terms go to zero more rapidly (Exercise IV.2.16.3.). The border cannot even be countably described: for any countable collection of convergent nonnegative series there is a convergent nonnegative series whose terms go to zero more slowly than all of them, and for any countable collection of divergent nonnegative series there is a divergent nonnegative series whose terms go to zero more rapidly than all of them (Exercise IV.2.16.4.). Thus, although a sequence like $\left(\frac{1}{k}\right)$ can be regarded as "close to the border between series convergence and divergence" (i.e. "just barely divergent"), there is no precise countable description of an exact border between convergence and divergence, and thus no foolproof countably describable test for convergence of nonnegative series.

Nonetheless, there are many useful tests which together work for most series arising in practice, as described in the next two sections.

IV.2.3. Tests for Convergence of Nonnegative Series

Besides the Comparison Tests of the previous section, which primarily illustrate the Fundamental Principle of Nonnegative Series, there are several standard tests for convergence of such series, along with a large number of specialized tests. In this subsection, we describe the three standard tests, the Integral Test, the Root Test, and the Ratio Test, and give a sampling of some of the more important specialized tests.

The Integral Test

Infinite series bear a close resemblance to improper integrals; in fact, infinite series can be regarded as a discrete analog of improper integrals. At least in the case of nonnegative series, this close relationship can be exploited to give a test for convergence of certain nonnegative series using improper integrals.

IV.2.3.1. Suppose f is a function on the interval $[1, \infty)$ which is (Riemann-)integrable on each finite subinterval (e.g. continuous or at least piecewise-continuous, or monotone). Recall () that the improper integral of f from 1 to ∞ is defined to be

$$\int_1^\infty f(t)\,dt = \lim_{b\to\infty}\left[\int_1^b f(t)\,dt\right]$$

if the limit exists (the improper integral is then said to *converge*). The finite integrals may be regarded as "partial integrals" and some sort of continuous analog of the partial sums of an infinite series.

If f is nonnegative on $[1, \infty)$, the "partial integral" from 1 to b is nondecreasing as b increases, so the improper integral converges if and only if the partial integrals are bounded; and

$$\int_1^\infty f(t)\,dt = \sup_{b>1}\left[\int_1^b f(t)\,dt\right]$$

if the supremum is finite. If f is nonincreasing, this will be the case if and only if $f(t) \to 0$ "sufficiently rapidly" as $t \to \infty$, in complete analogy with the Fundamental Principle of Nonnegative Series (IV.2.2.2.).

IV.2.3.2. If the function f is nonnegative and nonincreasing, we have that

$$f(k+1)\cdot 1 \le \int_k^{k+1} f(t)\,dt \le f(k)\cdot 1$$

for all k (see Figure ()). It follows that for any n, we have

$$\sum_{k=2}^{n+1} f(k) \le \int_1^{n+1} f(t)\,dt \le \sum_{k=1}^{n} f(k)\ .$$

Thus, if the nonnegative infinite series $\displaystyle\sum_{k=1}^{\infty} f(k)$ converges, we obtain that the supremum of the partial integrals from 1 to n is finite, and hence the improper integral $\displaystyle\int_1^\infty f(t)\,dt$ also converges. Conversely, we have

$$\sum_{k=1}^{n} f(k) \le f(1) + \int_1^n f(t)\,dt$$

for every n, so if the improper integral converges, the partial sums of the series $\displaystyle\sum_{k=1}^{\infty} f(k)$ are bounded and so the series converges. Thus we obtain:

IV.2.3.3. THEOREM. [INTEGRAL TEST] Let f be a function which is nonnegative and nonincreasing on the interval $[1, \infty)$. Then the infinite series $\sum_{k=1}^{\infty} f(k)$ converges if and only if the improper integral $\int_1^{\infty} f(t)\,dt$ converges, and we have

$$\int_1^{\infty} f(t)\,dt - f(1) \leq \sum_{k=1}^{\infty} f(k) \leq \int_1^{\infty} f(t)\,dt\ .$$

IV.2.3.4. Note that this test, when it applies, gives a complete necessary and sufficient condition for convergence of the series $\sum_{k=1}^{\infty} f(k)$. But although it gives some information about the sum, it does not directly allow a calculation of the exact sum. See ().

Of course, when using this test, as with any test, one must check that the hypotheses are satisfied.

Application: p-Series

We use the Integral Test to give an alternate proof of the convergence or divergence of the p-series (IV.2.2.14.); cf. IV.2.2.16..

IV.2.3.5. THEOREM. The p-series with exponent p converges if and only if $p > 1$.

PROOF: Suppose $p > 0$, $p \neq 1$. We apply the Integral Test to the function $f(t) = t^{-p}$, which is continuous, nonnegative, and nonincreasing on $[1, \infty)$. Then we have

$$\int_1^b t^{-p}\,dt = \left.\frac{t^{-p+1}}{-p+1}\right|_1^b = \frac{b^{1-p}}{1-p} - \frac{1}{1-p}\ .$$

If $p < 1$, $b^{1-p} \to +\infty$ as $b \to \infty$ and the improper integral diverges, so the series diverges. If $p > 1$, $b^{1-p} \to 0$ as $b \to \infty$ and the improper integral converges, so the series converges.
For completeness, we also do the case $p = 1$ using the integral test. We have

$$\int_1^b t^{-1}\,dt = \log t|_1^b = \log b - \log 1 = \log b$$

which goes to $+\infty$ as $b \to \infty$; hence the improper integral diverges and the series diverges too. 👍

The Ratio and Root Tests

The Ratio Test and the Root Test are two of the most widely-used tests for convergence of nonnegative series. They are rather crude, in the sense that the tests fail (give no conclusion) for many series; in fact, they are successful only for series which converge or diverge rather spectacularly. But when they do work they are often quite easy to apply, so are quite useful.

The Root Test and the Ratio Test are sometimes called *Cauchy's Test* and *d'Alembert's Test* respectively, after their originators (although it seems generous to attribute the Ratio Test to D'ALEMBERT; see [Gra00, p.116-117]).

IV.2.3.6. THEOREM. [BASIC ROOT TEST] Let $\sum_{k=1}^{\infty} a_k$ be a nonnegative infinite series. Then

(i) If there is a $u < 1$ such that $\sqrt[k]{a_k} = a_k^{1/k} \leq u$ for all sufficiently large k, then the series converges.

(ii) If $\sqrt[k]{a_k} = a_k^{1/k} \geq 1$ for infinitely many k, then the terms do not approach zero and the series diverges.

PROOF: We first show (ii). If $a_k^{1/k} \geq 1$ for infinitely many k, then $a_k \geq 1$ for infinitely many k, and thus we cannot have $\lim_{k\to\infty} a_k = 0$, and the series diverges by the Divergence Criterion.

For (i), if $a_k^{1/k} \leq u < 1$ for all $k \geq N$, then $a_k \leq u^k$ for all $k \geq N$. Comparing the series with the geometric series $\sum_{k=1}^{\infty} u^k$, which converges since $u < 1$, we conclude from the Extended Comparison Test that the series converges. 👍

The Root Test is usually phrased in terms of the lim sup, although the divergence part is not quite as strong (but often easier to check):

IV.2.3.7. COROLLARY. [ROOT TEST] Let $\sum_{k=1}^{\infty} a_k$ be a nonnegative infinite series, and set

$$r = \limsup_{k\to\infty} \sqrt[k]{a_k} = \limsup_{k\to\infty} a_k^{1/k} .$$

Then

(i) If $r < 1$, then the series converges.

(ii) If $r > 1$, then the terms do not approach zero and the series diverges.

(iii) If $r = 1$ but $\sqrt[k]{a_k} < 1$ for all sufficiently large k, then the series may or may not converge, and no conclusion may be drawn (the test fails).

In particular, if $\lim_{k\to\infty} a_k^{1/k}$ exists, the limit is the r of the theorem and the series converges or diverges according to whether $r < 1$ or $r > 1$. But the test applies even if this limit does not exist; recall that the lim sup always exists for any sequence.

PROOF: Part (i) is actually equivalent to IV.2.3.6.(i): suppose $r < 1$, and fix u with $r < u < 1$. Then there is an $N \in \mathbb{N}$ with $a_k^{1/k} < u$ for all $k \geq N$ (IV.1.8.6.), so the hypotheses of IV.2.3.6.(i) are satisfied. Conversely, if there is a u such that IV.2.3.6.(i) holds, then $r \leq u$ by IV.1.8.6..

For (ii), if $r > 1$, then $a_k^{1/k} > 1$ for infinitely many k, so the hypotheses of IV.2.3.6.(ii) are satisfied. (However, the two are not equivalent: the hypotheses of IV.2.3.6.(ii) do not imply those of IV.2.3.7.(ii).)

To show the test fails in case (iii), it suffices to consider p-series. For the p-series, we have

$$a_k^{1/k} = \frac{1}{[k^p]^{1/k}} = \frac{1}{[k^{1/k}]^p}$$

which has a limit of 1 for every p since $\lim_{k\to\infty} k^{1/k} = 1$ (). But the p-series can either converge or diverge depending on p. (Actually even the Basic Root Test fails to apply to p-series for $p > 0$.) 👍

IV.2.3.8. THEOREM. [BASIC RATIO TEST] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Then

(i) If there is a $u < 1$ such that $\frac{a_{k+1}}{a_k} \leq u$ for all sufficiently large k, then the series converges.

(ii) If $\frac{a_{k+1}}{a_k} \geq 1$ for all sufficiently large k, then the terms do not approach zero and the series diverges.

PROOF: The proof is quite similar to the proof of the Root Test.
(ii): Suppose there is an $N \in \mathbb{N}$ such that $\frac{a_{n+1}}{a_n} \geq 1$, $a_{n+1} \geq a_n$ for all $n \geq N$. Then the terms of the series are eventually nondecreasing and so, since they are positive, they cannot converge to zero. Thus the series diverges by the Divergence Criterion.
(i): Suppose there is an $N \in \mathbb{N}$ such that $\frac{a_{k+1}}{a_k} \leq u$ for all $k \geq N$. Then, for all $k \in \mathbb{N}$, we have

$$a_{N+k} \leq u a_{N+k-1} \leq u^2 a_{N+k-2} \leq \cdots \leq u^k a_N$$

and we may apply the Extended Comparison Test to the given series and the series $\sum_{k=1}^{\infty} a_N u^k$, which converges since $u < 1$. 👍

As with the Root Test, the Ratio Test is usually phrased in terms of lim sup (and lim inf):

IV.2.3.9. COROLLARY. [RATIO TEST] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Then

(i) If $\limsup_{k\to\infty} \frac{a_{k+1}}{a_k} < 1$, then the series converges.

(ii) If $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} > 1$, then the terms do not approach zero and the series diverges.

(iii) If $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} \leq 1 \leq \limsup_{k\to\infty} \frac{a_{k+1}}{a_k}$ but $\frac{a_{k+1}}{a_k} < 1$ for infinitely many k, then the series may or may not converge, and no conclusion may be drawn (the test fails).

A more special version often applies which is slightly easier to state:

IV.2.3.10. Corollary. [Ratio Test, Limit Version] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms, and suppose $r = \lim_{k\to\infty} \frac{a_{k+1}}{a_k}$ exists. Then

(i) If $r < 1$, then the series converges.

(ii) If $r > 1$, then the terms do not approach zero and the series diverges.

(iii) If $r = 1$ but $\frac{a_{k+1}}{a_k} < 1$ for infinitely many k, then the series may or may not converge, and no conclusion may be drawn (the test fails).

Proof: IV.2.3.10.(i)–(ii) are clearly special cases of IV.2.3.9.(i)–(ii). The proof of IV.2.3.9.(i)–(ii) is nearly identical to the proof of IV.2.3.7. from IV.2.3.6. and is left to the reader.

To show the test fails in case (iii), even in the case of IV.2.3.10., it suffices to consider p-series. For the p-series, we have

$$\frac{a_{k+1}}{a_k} = \frac{1/(k+1)^p}{1/k^p} = \left[\frac{k}{k+1}\right]^p$$

which has a limit of 1 for every p since $\lim_{k\to\infty} \frac{k}{k+1} = 1$. But the p-series can either converge or diverge depending on p.

There are even convergent positive series where $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} = 1 < \limsup_{k\to\infty} \frac{a_{k+1}}{a_k}$ (Exercise IV.2.16.5.) and divergent positive series where $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} < 1 = \limsup_{k\to\infty} \frac{a_{k+1}}{a_k}$ (Exercise IV.2.16.6.), and both convergent and divergent positive series where $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} < 1 < \limsup_{k\to\infty} \frac{a_{k+1}}{a_k}$ (Exercise IV.2.16.7.). 👍

To compare the strength of the Ratio Test and Root Test, we use the following fact:

IV.2.3.11. Proposition. Let (x_n) be a sequence of positive numbers. Then

$$\liminf_{n\to\infty} \frac{x_{n+1}}{x_n} \le \liminf_{n\to\infty} x_n^{1/n} \le \limsup_{n\to\infty} x_n^{1/n} \le \limsup_{n\to\infty} \frac{x_{n+1}}{x_n} \ .$$

Proof: We prove only the last inequality; the first is similar (Exercise ()). Let $r = \limsup_{n\to\infty} \frac{x_{n+1}}{x_n} \ge 0$. If $r = +\infty$, there is nothing to prove, so assume $r \in [0, \infty)$. Then, for any $s > r$, there is an $N \in \mathbb{N}$ such that $\frac{x_{n+1}}{x_n} < s$ for all $n \ge N$. For $m \in \mathbb{N}$, we have

$$x_{N+m} < s x_{N+m-1} < s^2 x_{N+m-2} < \cdots < s^m x_N$$

and so we obtain

$$x_n < s^n \frac{x_N}{s^N}$$

for all $n \geq N$ (note that $s > 0$ since $r \geq 0$). Taking n'th roots, we have

$$x_n^{1/n} < st^{1/n}$$

for all $n \geq N$, where $t = \frac{x_N}{s^N}$. So

$$\limsup_{n\to\infty} x_n^{1/n} \leq s \cdot \limsup_{n\to\infty} t^{1/n} = s$$

since $\lim_{n\to\infty} t^{1/n} = 1$ (). Since s is an arbitrary number $> r$, we have

$$\limsup_{n\to\infty} x_n^{1/n} \leq r \ .$$



IV.2.3.12. Any or all of the inequalities in IV.2.3.11. can be strict; see Exercise ().

IV.2.3.13. Corollary. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. If the Ratio Test gives a conclusion about the convergence or divergence of the series (i.e. the Ratio Test does not fail), then the Root Test also gives the same conclusion (in particular, it does not fail).

In other words, the Root Test is, at least in principle, a more broadly applicable test than the Ratio Test. It is in fact strictly more broadly applicable (Exercise ()). However, in practice the Ratio Test is frequently simpler to use than the Root Test, so there is utility in having both tests available.

IV.2.4. More Delicate Tests

There are several refinements of the Ratio Test which can be used for certain series for which the Ratio Test fails (see also IV.2.16.9. for a refinement of the Root Test). Most are special cases of the following blockbuster test:

IV.2.4.1. Theorem. [Kummer's Test] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms.

(i) If there are positive real numbers r and $p_1, p_2, \ldots$ such that

$$p_k \frac{a_k}{a_{k+1}} - p_{k+1} \geq r$$

for all sufficiently large k, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If there are positive real numbers $p_1, p_2, \dots$ such that

$$p_k \frac{a_k}{a_{k+1}} - p_{k+1} \le 0$$

for all sufficiently large k, and $\displaystyle\sum_{k=1}^{\infty} \frac{1}{p_k}$ diverges, then $\displaystyle\sum_{k=1}^{\infty} a_k$ diverges.

Note that the Basic Ratio Test is the special case where $p_k = 1$ for all k.

PROOF: For both (i) and (ii), by passing to a tail of the series we may assume the inequalities hold for all k.

(i): For each k, we have $a_{k+1} \le \frac{1}{r}[p_k a_k - p_{k+1} a_{k+1}]$, so for each $n > 1$,

$$s_n = \sum_{k=1}^{n} a_k = a_1 + \sum_{k=2}^{n} a_k \le a_1 + \frac{1}{r} \sum_{k=1}^{n-1} (p_k a_k - p_{k+1} a_{k+1})$$

$$= a_1 + \frac{1}{r} \left[\sum_{k=1}^{n-1} p_k a_k - \sum_{k=2}^{n} p_k a_k \right] = a_1 + \frac{p_1 a_1}{r} - \frac{p_n a_n}{r} \le a_1 + \frac{p_1 a_1}{r}$$

and so the partial sums are bounded and the series converges.

(ii): We have that

$$a_2 \ge \frac{p_1 a_1}{a_2},\ a_3 \ge \frac{p_2 a_2}{p_3} \ge \frac{p_1 a_1}{p_3},\ \dots\ ,\ a_k \ge \frac{p_1 a_1}{p_k},\ \dots$$

and so we can apply the Comparison Test to $\displaystyle\sum_{k=1}^{\infty} a_k$ and the divergent series

$$a_1 + \sum_{k=2}^{\infty} \frac{p_1 a_1}{p_k} \ .$$



IV.2.4.2. Kummer's Test is, in principle at least, a maximally applicable test: for any infinite series with strictly positive terms, there is a sequence (p_k) of positive numbers for which the test works to give either convergence or divergence (Exercise IV.2.16.11.). However, it can be extremely difficult to find such a sequence for a given series, so the general Kummer Test is of limited applicability in practice.

Special cases of Kummer's Test with specific sequences (p_n) are more useful. Here is one of the most widely-used ones:

IV.2.4.3. THEOREM. [RAABE'S TEST] Let $\displaystyle\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Then

(i) If $\displaystyle\liminf_{k\to\infty} k \left[\frac{a_k}{a_{k+1}} - 1 \right] > 1$, the series $\displaystyle\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $k\left[\dfrac{a_k}{a_{k+1}}-1\right]\le 1$ for all sufficiently large k, the series $\displaystyle\sum_{k=1}^{\infty} a_k$ diverges.

PROOF: Apply Kummer's Test with $p_k = k$ for all k and, for (i), r a number with $0 < r < \liminf\limits_{k\to\infty} k\left[\dfrac{a_k}{a_{k+1}}-1\right]-1$. 👍

A limit version is frequently applicable:

IV.2.4.4. COROLLARY. [RAABE'S TEST, LIMIT VERSION] Let $\displaystyle\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$L = \lim_{k\to\infty} k\left[\frac{a_k}{a_{k+1}}-1\right]$$

exists (as an extended real number). Then

(i) If $L > 1$, the series $\displaystyle\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $L < 1$, the series $\displaystyle\sum_{k=1}^{\infty} a_k$ diverges.

(iii) If $L = 1$, then the series $\displaystyle\sum_{k=1}^{\infty} a_k$ may or may not converge, and no conclusion may be drawn (the test fails).

Raabe's Test is often written in a slightly different but equivalent form (cf. Exercise IV.2.16.13.).

IV.2.4.5. COROLLARY. [RAABE'S TEST, VERSION 2] Let $\displaystyle\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$L = \lim_{k\to\infty} k\left[1-\frac{a_{k+1}}{a_k}\right]$$

exists (as an extended real number). Then

(i) If $L > 1$, the series $\displaystyle\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $L < 1$, the series $\displaystyle\sum_{k=1}^{\infty} a_k$ diverges.

(iii) If $L = 1$, then the series $\sum_{k=1}^{\infty} a_k$ may or may not converge, and no conclusion may be drawn (the test fails).

Yet another formulation of Raabe's Test is:

IV.2.4.6. THEOREM. [RAABE'S TEST, VERSION 3] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$\frac{a_{k+1}}{a_k} = 1 - \frac{\lambda}{k} + o\left(\frac{1}{k}\right)$$

for some fixed real number λ. Then

(i) If $\lambda > 1$, the series $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\lambda < 1$, the series $\sum_{k=1}^{\infty} a_k$ diverges.

(iii) If $\lambda = 1$, then the series $\sum_{k=1}^{\infty} a_k$ may or may not converge, and no conclusion may be drawn (the test fails).

PROOF: Applying IV.2.4.5., we have

$$k\left[1 - \frac{a_{k+1}}{a_k}\right] = \lambda + k \cdot o\left(\frac{1}{k}\right) = \lambda + o(1)$$

so we have $L = \lambda$. The converse implication is similar, and left as an exercise. 👍

IV.2.4.7. Note that Raabe's Test (even the limit version IV.2.4.4.) is more delicate than the Ratio Test: if $\limsup_{k\to\infty} \frac{a_{k+1}}{a_k} < 1$, then $L = +\infty$, and if $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} > 1$, then $L = -\infty$, so Raabe's Test applies (and gives the same conclusion, of course).

Raabe's test is delicate enough to handle p-series for $p \neq 1$: for the p-series we have

$$L = \lim_{k\to\infty} k\left[\frac{a_k}{a_{k+1}} - 1\right] = \lim_{k\to\infty} k\left[\frac{(k+1)^p}{k^p} - 1\right] = \lim_{k\to\infty} \frac{(k+1)^p - k^p}{k^{p-1}} = p$$

(See Exercise ().)

However, Raabe's Test does not work for the harmonic series (although the slight extension to Kummer's Test for $p_k = k$ does), nor does it work for the $\log^p$-series

$$\sum_{k=2}^{\infty} \frac{1}{k[\log k]^p}$$

which converges for $p > 1$ and diverges for $p \leq 1$ (). For these $\log^p$-series, a "second-order Raabe's Test" does apply:

IV.2.4.8. THEOREM. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$M = \lim_{k\to\infty} (\log k) \left(k \left[\frac{a_k}{a_{k+1}} - 1 \right] - 1 \right)$$

exists (as an extended real number). Then

(i) If $M > 1$, the series $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $M < 1$, the series $\sum_{k=1}^{\infty} a_k$ diverges.

(iii) If $M = 1$, then the series $\sum_{k=1}^{\infty} a_k$ may or may not converge, and no conclusion may be drawn (the test fails).

PROOF: Apply Kummer's Test with $p_k = k \log k$. (See Exercise ().) 👍

There is a variation of this second-order test similar to IV.2.4.6.. The proof is left as an exercise.

IV.2.4.9. THEOREM. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$\frac{a_{k+1}}{a_k} = 1 - \frac{1}{k} - \frac{\mu}{k \log k} + o\left(\frac{1}{k \log k} \right)$$

for some fixed real number μ. Then

(i) If $\mu > 1$, the series $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\mu < 1$, the series $\sum_{k=1}^{\infty} a_k$ diverges.

(iii) If $\mu = 1$, then the series $\sum_{k=1}^{\infty} a_k$ may or may not converge, and no conclusion may be drawn (the test fails).

IV.2.4.10. The second-order test will work for the $\log^p$-series with $p \neq 1$. It will not work for the $\log\log^p$-series

$$\sum_{k=3}^{\infty} \frac{1}{k \log k [\log\log k]^p}$$

(which again converges or diverges depending whether $p > 1$ or $p \leq 1$; cf. ()), but a "third-order Raabe Test" will (for $p \neq 1$), using

$$L = \lim_{k\to\infty} (\log\log k) \left[(\log k) \left(k \left[\frac{a_k}{a_{k+1}} - 1 \right] - 1 \right) - 1 \right] .$$

The process can be continued *ad infinitum* to higher orders. Details are left to the interested reader.

Gauss's Test

As part of his work on hypergeometric series (), Gauss developed an early test which is closely related to Raabe's test, but both the hypotheses and conclusion are slightly stronger. (Gauss actually only stated this test when the $O\left(\frac{1}{k^p}\right)$ term is a rational function in k.) Note that this test never fails (when the hypotheses are satisfied).

IV.2.4.11. Theorem. [Gauss's Test] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$\frac{a_{k+1}}{a_k} = 1 - \frac{\lambda}{k} + O\left(\frac{1}{k^p}\right)$$

for some fixed real number λ and some fixed $p > 1$. Then

(i) If $\lambda > 1$, the series $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If $\lambda \leq 1$, the series $\sum_{k=1}^{\infty} a_k$ diverges.

Proof: Since $O\left(\frac{1}{k^p}\right) = o\left(\frac{1}{k}\right)$, if $\lambda \neq 1$ the result follows from IV.2.4.6.. And if $\lambda = 1$, we have $O\left(\frac{1}{k^p}\right) = o\left(\frac{1}{k \log k}\right)$ so the result follows from IV.2.4.9. with $\mu = 0$. 👍

Cauchy's Condensation Test and Yermakov's Test

These two closely related tests reduce the question of convergence of a series with decreasing nonnegative terms to consideration of a "sparse" subseries. Cauchy's Condensation Test is a generalization of the argument used in IV.2.2.11..

IV.2.4.12. THEOREM. [CAUCHY'S CONDENSATION TEST] Let (c_k) be a strictly increasing sequence of positive integers such that there is a constant M with

$$c_{k+2} - c_{k+1} \leq M(c_{k+1} - c_k)$$

for all k. Then, if $\sum_{k=1}^{\infty} a_k$ is an infinite series with nonincreasing nonnegative terms, the two series

$$\sum_{k=1}^{\infty} a_k \quad \text{and} \quad \sum_{k=1}^{\infty} (c_{k+1} - c_k) a_{c_k}$$

either both converge or both diverge.

PROOF: If $c_1 > 1$, set $r = \sum_{k=1}^{c_1 - 1} a_k$; if $c_1 = 1$, set $r = 0$. Write s_n and t_n for the n'th partial sums of the two series respectively. Then, for $n \leq c_m$, we have

$$s_n \leq s_{c_m} \leq r + (a_{c_1} + a_{c_1+1} + \cdots + a_{c_2-1}) + (a_{c_2} + \cdots + a_{c_3-1}) + \cdots + (a_{c_m} + \cdots + a_{c_{m+1}-1})$$

$$\leq r + (c_2 - c_1)a_{c_1} + \cdots + (c_{m+1} - c_m)a_{c_m} = r + t_m$$

and so, if the second series converges, i.e. (t_m) is bounded, the sequence (s_n) is also bounded so the first series converges. On the other hand, if $n > c_m$, we have

$$s_n \geq s_{c_m} \geq (a_{c_1+1} + \cdots + a_{c_2}) + (a_{c_2+1} + \cdots + a_{c_3}) + \cdots + (a_{c_{m-1}+1} + \cdots + a_{c_m})$$

$$\geq (c_2 - c_1)a_{c_2} + (c_3 - c_2)a_{c_3} + \cdots + (c_m - c_{m-1})a_{c_m}$$

$$M s_n \geq (c_3 - c_2)a_{c_2} + \cdots + (c_{m+1} - c_m)a_{c_m} = t_m - a_{c_1}$$

and so if the first series converges, i.e. (s_n) is bounded, the sequence (t_m) is also bounded and the second series converges. 👍

The most commonly used case of this test is to take $c_k = 2^k$ for all k (then we can take $M = 2$). We obtain:

IV.2.4.13. Corollary. [Cauchy's Condensation Test, Common Version] If $\sum_{k=1}^{\infty} a_k$ is an infinite series with nonincreasing nonnegative terms, the two series

$$\sum_{k=1}^{\infty} a_k \quad \text{and} \quad \sum_{k=1}^{\infty} 2^k a_{2^k}$$

either both converge or both diverge.

This corollary suffices to show that the harmonic series converges, since the terms of the second series then become all 1, so the second series diverges by the Divergence Criterion. (This is essentially the argument of IV.2.2.11..) The corollary also suffices to show convergence and divergence of the series in ().

The test of Yermakov (or Ermakoff) is similar in nature, but is based on the Integral Test and gives a direct criterion for convergence or divergence without making a comparison with another series:

IV.2.4.14. Theorem. [Yermakov's Test] Let f be a nonincreasing strictly positive function on $[1, \infty)$, and ϕ a strictly increasing differentiable function on $[1, \infty)$ with $\phi(t) > t$ for all t. Then

(i) If $\limsup_{t\to\infty} \dfrac{\phi'(t)f(\phi(t))}{f(t)} < 1$, then $\sum_{k=1}^{\infty} f(k)$ converges.

(ii) If $\phi'(t)f(\phi(t)) \geq f(t)$ for all sufficiently large t (in particular, if $\liminf_{t\to\infty} \dfrac{\phi'(t)f(\phi(t))}{f(t)} > 1$), then $\sum_{k=1}^{\infty} f(k)$ diverges.

Proof: (i): It suffices to show that $\int_c^{\infty} f(t)\,dt$ converges for some c by the Integral Test. Under the hypothesis, there is an $r < 1$ and an a such that $\phi'(t)f(\phi(t)) < r \cdot f(t)$ for all $t \geq a$. Then, for any $b > a$ (we restrict to $b > \phi(a)$), we have

$$\int_a^b f(\phi(t))\phi'(t)\,dt < r \int_a^b f(t)\,dt$$

so, making the substitution $u = \phi(t)$ in the first integral, we obtain

$$\int_{\phi(a)}^{\phi(b)} f(u)\,du < r \int_a^b f(t)\,dt\ .$$

Writing the first integral as $\int_{\phi(a)}^{\phi(b)} f(t)\,dt$, we have

$$(1-r)\int_{\phi(a)}^{\phi(b)} f(t)\,dt < r\left[\int_a^b f(t)\,dt - \int_{\phi(a)}^{\phi(b)} f(t)\,dt\right]$$

$$= r\left[\int_a^{\phi(a)} f(t)\,dt + \int_{\phi(a)}^{b} f(t)\,dt - \int_{\phi(a)}^{b} f(t)\,dt - \int_b^{\phi(b)} f(t)\,dt\right] = r\left[\int_a^{\phi(a)} f(t)\,dt - \int_b^{\phi(b)} f(t)\,dt\right]$$

and, since $b < \phi(b)$, the last integral is positive, so we have

$$(1-r)\int_{\phi(a)}^{\phi(b)} f(t)\,dt < r\int_a^{\phi(a)} f(t)\,dt\ .$$

The right side is constant, so we have that the left side is bounded as $b \to +\infty$ (and hence $\phi(b) \to +\infty$). Thus $\displaystyle\int_{\phi(a)}^{\infty} f(t)\,dt$ converges.

(ii): Suppose $\phi'(t)f(\phi(t)) \geq f(t)$ for all $t \geq a$. Then we have, for any $b > a$,

$$\int_a^b f(\phi(t))\phi'(t)\,dt \geq \int_a^b f(t)\,dt\ .$$

Changing variables as before, we obtain

$$\int_{\phi(a)}^{\phi(b)} f(t)\,dt \geq \int_a^b f(t)\,dt$$

$$\int_{\phi(a)}^{b} f(t)\,dt + \int_b^{\phi(b)} f(t)\,dt \geq \int_a^{\phi(a)} f(t)\,dt + \int_{\phi(a)}^{b} f(t)\,dt$$

$$\int_b^{\phi(b)} f(t)\,dt \geq \int_a^{\phi(a)} f(t)\,dt\ .$$

The right side is a strictly positive constant, and if $\displaystyle\int_1^{\infty} f(t)\,dt$ converged, the left side would have to converge to 0 as $b \to \infty$. Thus $\displaystyle\int_1^{\infty} f(t)\,dt$ diverges, and the series also diverges by the Integral Test. 👍

This test is commonly written in the special case $\phi(t) = e^t$:

IV.2.4.15. Corollary. [Yermakov's Test, Common Version] Let f be a nonincreasing strictly positive function on $[1, \infty)$. Then

(i) If $\displaystyle\limsup_{t\to\infty} \frac{e^t f(e^t)}{f(t)} < 1$, then $\displaystyle\sum_{k=1}^{\infty} f(k)$ converges.

(ii) If $\displaystyle\liminf_{t\to\infty} \frac{e^t f(e^t)}{f(t)} > 1$, then $\displaystyle\sum_{k=1}^{\infty} f(k)$ diverges.

IV.2.4.16. This version is sufficient to handle all the p-series, $\log^p$-series, $\log\log^p$-series, etc., of () (Exercise ()). (Of course, these series can also be handled directly by the Integral Test, as well as by Cauchy's Condensation Test.)

IV.2.5. Absolute and Conditional Convergence

Convergence or divergence of an infinite series is a much more subtle matter if it has both positive and negative terms, since there can be complicated cancellation in the partial sums. In fact, there are convergent series whose terms go to zero arbitrarily slowly. But series which converge because their terms go to zero rapidly (*absolutely convergent series*) are better behaved and do not have the pathologies possible in the general case where convergence depends delicately on such cancellation.

IV.2.5.1. DEFINITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. The series $\sum_{k=1}^{\infty} a_k$ *converges absolutely* if $\sum_{k=1}^{\infty} |a_k|$ converges. The series $\sum_{k=1}^{\infty} a_k$ *converges conditionally* if $\sum_{k=1}^{\infty} a_k$ converges but $\sum_{k=1}^{\infty} |a_k|$ diverges.

Of course, convergence and absolute convergence are the same thing for nonnegative series (and, by IV.2.1.18., also for series with only negative terms). But there is a big difference for general series.

The first observation is that absolute convergence implies convergence. Buried in the proof is the Completeness Property of $\mathbb{R}$, in the form of the Cauchy Criterion; in fact this result is another equivalent characterization of completeness of $\mathbb{R}$ (cf. ()).

IV.2.5.2. PROPOSITION. Every absolutely convergent series converges.

PROOF: Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, and suppose $\sum_{k=1}^{\infty} |a_k|$ converges. Let $\epsilon > 0$, and fix N such that, whenever $N \leq n < m$, we have

$$\sum_{k=n+1}^{m} |a_k| < \epsilon$$

(IV.2.1.11.). Then, by the triangle inequality we have for $N \leq n < m$,

$$\left| \sum_{k=n+1}^{m} a_k \right| \leq \sum_{k=n+1}^{m} |a_k| < \epsilon$$

so $\sum_{k=1}^{\infty} a_k$ converges by the Cauchy Criterion (IV.2.1.11.). 👍

For an alternate proof, see IV.2.16.16..

IV.2.5.3. Thus there are only three possibilities for an infinite series $\sum_{k=1}^{\infty} a_k$:

(i) The series converges absolutely, i.e. $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} |a_k|$ both converge.

(ii) The series converges conditionally, i.e. $\sum_{k=1}^{\infty} a_k$ converges but $\sum_{k=1}^{\infty} |a_k|$ diverges.

(iii) The series diverges, i.e. both $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} |a_k|$ diverge.

Of course, conditional convergence is only a possibility when there are infinitely many positive terms and also infinitely many negative terms.

The Fundamental Principle of Nonnegative Series (IV.2.2.2.) asserts that a series converges absolutely if and only if its terms go to zero sufficiently rapidly.

We have not yet seen an example of a conditionally convergent series, but there are (rather degenerate) simple examples where the terms go to zero arbitrarily slowly:

IV.2.5.4. Example. Let (b_k) be a sequence of numbers converging to zero. For each k, set $a_{2k-1} = b_k$ and $a_{2k} = -b_k$. Then the partial sums s_n of the series $\sum_{k=1}^{\infty} a_k$ satisfy $s_n = 0$ for n even and $s_n = b_m$ if $n = 2m-1$ is odd. Thus $s_n \to 0$ and the series converges (with sum 0). The series $\sum_{k=1}^{\infty} a_k$ converges absolutely if and only if $\sum_{k=1}^{\infty} |b_k|$ converges; hence, for example, if $b_k = \frac{1}{k}$, the series $\sum_{k=1}^{\infty} a_k$ converges conditionally.

We will see more interesting examples of conditionally convergent series.

To further illustrate the nature of conditionally convergent series, we note:

IV.2.5.5. Proposition. Let $\sum_{k=1}^{\infty} a_k$ be a conditionally convergent series, and let $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ be the infinite series whose terms are the positive and negative terms of $\sum_{k=1}^{\infty} a_k$ respectively. Then $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ both diverge (i.e. the partial sums of both series are unbounded).

PROOF: The series $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ cannot both converge, since then we would have, for any n,

$$\sum_{k=1}^{n} |a_k| \le \sum_{k=1}^{\infty} b_k + \sum_{k=1}^{\infty} (-c_k)$$

and $\sum_{k=1}^{\infty} a_k$ would converge absolutely. If $\sum_{k=1}^{\infty} b_k$ diverges but $\sum_{k=1}^{\infty} c_k$ converges, there is an $m > 0$ such that all partial sums of $\sum_{k=1}^{\infty} c_k$ are $\ge -m$; for any M all sufficiently large partial sums of $\sum_{k=1}^{\infty} b_k$ are $\ge M$, so for all sufficiently large n we have

$$\sum_{k=1}^{n} a_k \ge M - m$$

and so $\sum_{k=1}^{\infty} a_k$ cannot converge. The argument in the other case is almost identical. 👍

What the proof actually shows is the following slightly more precise result:

IV.2.5.6. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with infinitely many terms of each sign, and let $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ be the infinite series whose terms are the positive and negative terms of $\sum_{k=1}^{\infty} a_k$ respectively. Then

(i) If $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ both converge, then $\sum_{k=1}^{\infty} a_k$ converges absolutely.

(ii) If one of $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ converges and the other diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

Tests for Absolute Convergence

Any of the tests for convergence of nonnegative series become tests for absolute convergence if absolute values are inserted. We explicitly state the Root and Ratio Tests, and the Comparison Test:

IV.2.5.7. THEOREM. [ROOT TEST] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, and set

$$r = \limsup_{k\to\infty} \sqrt[k]{|a_k|} = \limsup_{k\to\infty} |a_k|^{1/k} \ .$$

Then

(i) If $r < 1$, then the series converges absolutely.

(ii) If $r > 1$, then the terms do not approach zero and the series diverges.

(iii) If $r = 1$, then the series may converge absolutely, converge conditionally, or diverge, and no conclusion may be drawn (the test fails).

IV.2.5.8. THEOREM. [RATIO TEST] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with nonzero terms. Then

(i) If $\limsup_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| < 1$, then the series converges absolutely.

(ii) If $\liminf_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| > 1$, then the terms do not approach zero and the series diverges.

(iii) If $\liminf_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| \leq 1 \leq \limsup_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right|$, then the series may converge absolutely, converge conditionally, or diverge, and no conclusion may be drawn (the test fails).

The only part of these theorems which may not be obvious from earlier work is the assertion in (iii) of each that the series can converge conditionally. The series of IV.2.5.4. with $b_k = \frac{1}{k}$ is an example in both cases. (In fact, any conditionally convergent series must fall into case (iii) in each theorem.)

IV.2.5.9. THEOREM. [COMPARISON TEST] Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be infinite series. Suppose $|a_k| \leq |b_k|$ for all k. If $\sum_{k=1}^{\infty} b_k$ converges absolutely, then $\sum_{k=1}^{\infty} a_k$ converges absolutely.

IV.2.5.10. **Caution:** We do not get a comparison test for simple convergence of series, i.e. if $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ are infinite series, with $|a_k| \leq |b_k|$ for all k, and $\sum_{k=1}^{\infty} b_k$ converges, then we cannot conclude that $\sum_{k=1}^{\infty} a_k$ converges. There are many counterexamples from the Alternating Series Test (IV.2.5.13.), e.g. $a_k = \frac{1}{k}$ and $b_k = \frac{(-1)^k}{k}$ (IV.2.5.14.(i)).

Combining the Comparison Test and the Triangle Inequality with IV.2.1.16., we obtain:

IV.2.5.11. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be absolutely convergent infinite series. Then the series

$$\sum_{k=1}^{\infty}(a_k + b_k)$$

converges absolutely, and

$$\sum_{k=1}^{\infty}(a_k + b_k) = \sum_{k=1}^{\infty} a_k + \sum_{k=1}^{\infty} b_k \ .$$

It is more straightforward to see that if $\sum_{k=1}^{\infty} a_k$ is an absolutely convergent infinite series, and $c \in \mathbb{R}$, then $\sum_{k=1}^{\infty} ca_k$ is also absolutely convergent.

Alternating Series

There is a simple criterion for the convergence of infinite series in which the terms alternate sign. In such series the cancellation of positive and negative terms is regular enough to make the series converge under quite mild restrictions.

IV.2.5.12. DEFINITION. An *alternating series* is an infinite series in which successive terms are of opposite signs.

The next theorem is the main result of this subsection. It is a special case of Dirichlet's Test, but the proof is considerably simpler in this case, essentially reducing to the Nested Intervals Property of $\mathbb{R}$. This is also the most commonly used test for series with terms of varying sign.

IV.2.5.13. THEOREM. [ALTERNATING SERIES TEST] Let $\sum_{k=1}^{\infty} a_k$ be an alternating series. If

(i) The terms decrease in absolute value, i.e. $|a_{k+1}| \leq |a_k|$ for all k, and

(ii) $\lim_{k\to\infty} a_k = 0$,

then the series $\sum_{k=1}^{\infty} a_k$ converges. The sum lies between any two consecutive partial sums.

Note that this result says nothing about absolute or conditional convergence; either type can occur under the hypotheses.

PROOF: By discarding the first term of the series if necessary (cf. IV.2.1.6.), we may assume $a_1 > 0$, and hence $a_k > 0$ for k odd, $a_k < 0$ for k even. (This is not an essential restriction, but simplifies the notation of the proof.)

Let s_n be the n'th partial sum of the series. By (i), we have

$$s_2 \leq s_4 \leq s_6 \leq \cdots \leq s_5 \leq s_3 \leq s_1$$

and thus for any m we have $s_{2m} \leq s_k \leq s_{2m-1}$ for all $k \geq 2m-1$. (This can be proved carefully by induction on k.) Thus, if $k, j \geq 2m-1$, we have

$$|s_k - s_j| \leq s_{2m-1} - s_{2m} = |a_{2m}| \ .$$

For any $\epsilon > 0$, we have $|a_{2m}| < \epsilon$ for all sufficiently large m by (ii), so the sequence (s_n) is a Cauchy sequence and hence converges; the limit s satisfies $s_{2m} \leq s \leq s_{2m-1}$ for any m. 👍

IV.2.5.14. EXAMPLES. (i) The Alternating Series Test provides more interesting examples of conditionally convergent series than IV.2.5.4.; in fact, in the Alternating Series Test the terms can also go to zero arbitrarily slowly. For example, the alternating p-series

$$\sum_{k=1}^{\infty} \frac{(-1)^k}{k^p}$$

converges for any $p > 0$; the convergence is absolute for $p > 1$ and conditional for $0 < p \leq 1$.

(ii) In the Alternating Series Test, both hypotheses are necessary. Hypothesis (ii) is clearly necessary by the Divergence Criterion, and is easily remembered. It is more common to forget to check (i), but some version of (i) is necessary: for example, set $a_k = \frac{1}{k}$ for k odd and $a_k = -\frac{1}{k^2}$ for k even. Then the series $\sum_{k=1}^{\infty} a_k$ is alternating, and the terms approach zero; but the series of positive terms diverges and the series of negative terms converges, and hence $\sum_{k=1}^{\infty} a_k$ diverges by IV.2.5.6.. ✋

It does suffice, of course, that condition (i) (and even the fact that the series alternates) just hold from some point on in the series, by IV.2.1.6.; the statement that the sum of the series lies between successive partial sums then only holds beyond the point where the terms satisfy the hypotheses.

IV.2.6. Hypergeometric Series

Hypergeometric series are a generalization of geometric series which arise naturally in applications: many applied problems give rise to power series which are hypergeometric (). The theory of hypergeometric series was primarily developed by GAUSS.

IV.2.6.1. DEFINITION. A *hypergeometric series* is an infinite series $\sum_{k=1}^{\infty} a_k$ in which all a_k are nonzero, and for which there are polynomials p and q such that

$$\frac{a_{k+1}}{a_k} = \frac{p(k)}{q(k)}$$

for all k. The summation may also start at 0 or another integer.

IV.2.6.2. EXAMPLES. (i) A geometric series is a hypergeometric series where p and q are constant: if the ratio is x, take $p(k) = x$ (x is a constant!) and $q(k) = 1$.

(ii) For fixed x, the series

$$\sum_{k=0} \frac{x^k}{k!}$$

for e^x is hypergeometric: we have

$$\frac{a_{k+1}}{a_k} = \frac{k!x^{k+1}}{(k+1)!x^k} = \frac{x}{k+1}$$

so we may take $p(k) = x$ (constant) and $q(k) = k+1$. The power series for $\sin x$ and $\cos x$ (x fixed) are also hypergeometric.

(iii) The basic hypergeometric series considered by Gauss was

$$\sum_{k=0}^{\infty} \frac{\alpha(\alpha+1)\cdots(\alpha+k-1)\beta(\beta+1)\cdots(\beta+k-1)}{k!\gamma(\gamma+1)\cdots(\gamma+k-1)} x^k$$

where α, β, γ are real numbers which are not integers ≤ 0, x is a fixed real number, and the $k = 0$ term is taken to be 1. We then have

$$\frac{a_{k+1}}{a_k} = \frac{(\alpha+k)(\beta+k)x}{(1+k)(\gamma+k)}$$

so the series is hypergeometric. (The geometric series is the special case where $\alpha = 1$ and $\beta = \gamma$.)

IV.2.6.3. Since $p(k)$ and $q(k)$ have constant sign for sufficiently large k, it follows that the terms of a hypergeometric series either are eventually constant in sign or eventually alternate in sign.

IV.2.6.4. The Ratio Test is tailor-made for determining convergence of hypergeometric series, since

$$\lim_{k\to\infty} \frac{p(k)}{q(k)}$$

always exists (as an extended real number) and is easily computed. (In fact, the Ratio Test was originally specifically developed largely to analyze hypergeometric series.) Thus the series converges absolutely if this limit is less than one in absolute value, and diverges if greater than one. In particular, we have:

IV.2.6.5. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be a hypergeometric series with

$$\frac{a_{k+1}}{a_k} = \frac{p(k)}{q(k)}$$

for all k. Let c and d be the leading coefficients of p and q respectively. Then

(i) If the degree of q is greater than the degree of p, the series converges absolutely.

(ii) If the degree of p is greater than the degree of q, the series diverges.

(iii) If p and q have the same degree and $\left|\frac{c}{d}\right| < 1$, the series converges absolutely.

(iv) If p and q have the same degree and $\left|\frac{c}{d}\right| > 1$, the series diverges.

IV.2.6.6. Corollary. The basic hypergeometric series of IV.2.6.2.(iii) converges absolutely if $|x| < 1$ and diverges if $|x| > 1$.

The most interesting situation is when the Ratio Test fails, i.e. when p and q have the same degree and $\left|\frac{c}{d}\right| = 1$. (In the basic hypergeometric series, this is the case $|x| = 1$.) Gauss was led to develop his test (IV.2.4.11.), which always applies in this situation.

IV.2.6.7. Suppose p and q have the same degree m and $\left|\frac{c}{d}\right| = 1$. Then

$$\frac{|a_{k+1}|}{|a_k|} = \left|\frac{p(k)/c}{q(k)/d}\right| = \left|\frac{k^m + ak^{m-1} + r(k)}{k^m + bk^{m-1} + s(k)}\right|$$

where r and s are polynomials of degree $\leq m-2$. Since the quotient in the last term is positive for large k, the absolute values can be removed. We then have

$$1 - \frac{|a_{k+1}|}{|a_k|} = \frac{(k^m + bk^{m-1} + s(k)) - (k^m + ak^{m-1} + r(k))}{k^m + bk^{m-1} + s(k)} = \frac{(b-a)k^{m-1} + (r(k) - s(k)}{k^m + bk^{m-1} + s(k)}$$

and so

$$\begin{aligned}
\frac{|a_{k+1}|}{|a_k|} &= 1 - \frac{(b-a)k^{m-1} + (s(k) - r(k))}{k^m + bk^{m-1} + s(k)} \\
&= 1 - \frac{b-a}{k} + \left[\frac{b-a}{k} - \frac{(b-a)k^{m-1} + ((s(k) - r(k))}{k^m + bk^{m-1} + s(k)}\right] \\
&= 1 - \frac{b-a}{k} + \left[\frac{(b-a)(k^m + bk^{m-1} + s(k))}{k(k^m + bk^{m-1} + s(k))} - \frac{(b-a)k^m + k(s(k) - r(k))}{k(k^m + bk^{m-1} + s(k))}\right] \\
&= 1 - \frac{b-a}{k} + \frac{\phi(k)}{k(k^m + bk^{m-1} + s(k))}
\end{aligned}$$

where ϕ is a polynomial of degree $\leq m-1$. The last term is $O\left(\frac{1}{k^2}\right)$, so we have

$$\frac{|a_{k+1}|}{|a_k|} = 1 - \frac{b-a}{k} + O\left(\frac{1}{k^2}\right) .$$

Applying Gauss's Test, we obtain:

IV.2.6.8. THEOREM.

(i) If $b - a > 1$, the series converges absolutely.

(ii) If $b - a \leq 1$, the series does not converge absolutely.

This result completely settles the convergence question when the terms of the series eventually have constant sign. GAUSS went on to completely analyze the case where the terms alternate in sign. (We are not following GAUSS'S approach; see [Bre07] for a nice exposition of GAUSS'S arguments, including a proof of the next result.)

IV.2.6.9. THEOREM. If the terms of the series alternate in sign, then

(i) If $b < a$, the terms go to infinity in absolute value so the series diverges.

(ii) If $b = a$, the absolute values of the terms approach a nonzero finite limit and the series diverges.

(iii) If $b > a$, the series converges.

The convergence in (iii) is conditional if $a < b \leq a + 1$ and absolute if $b > a + 1$.

We can then obtain the convergence for the basic hypergeometric series:

IV.2.6.10. COROLLARY. Let $\sum_{k=1}^{\infty} a_k$ be the hypergeometric series of IV.2.6.2.(iii) with parameters α, β, γ, and $|x| = 1$. Then

(i) If $\alpha + \beta < \gamma$, the series converges absolutely.

(ii) If $\alpha + \beta - \gamma \geq 1$, the series diverges.

(iii) If $x = 1$ and $\alpha + \beta \geq \gamma$, the series diverges.

(iv) If $x = -1$ and $\gamma \leq \alpha + \beta < \gamma + 1$, the series converges conditionally.

PROOF: We have

$$\frac{|a_{k+1}|}{|a_k|} = \frac{k^2 + (\alpha + \beta)k + \alpha\beta}{k^2 + (1 + \gamma)k + \gamma}$$

by a simple calculation. The terms of the series are eventually constant in sign if $x > 0$ and eventually alternate if $x < 0$. 👍

IV.2.7. Scaling of Series

In this subsection, we examine what happens to the convergence of a series $\sum_{k=1}^{\infty} a_k$ when the terms a_k are scaled by factors c_k. Of course, when the c_k are constant, the scaling does not affect convergence, but when the c_k vary, convergence can be affected very subtly.

The most obvious example where convergence can be affected is where the c_k are unbounded. For bounded scaling, the simplest case where convergence changes is where $\sum_{k=1}^{\infty} a_k$ is a conditionally convergent alternating series and $c_k = (-1)^k$. But even if the c_k are all nonnegative and/or converge to a limit (even 0), convergence or divergence can change by the scaling:

IV.2.7.1. EXAMPLES. (i) Let $a_k = \dfrac{(-1)^k}{\sqrt{k}}$ for all k. Set $c_k = \dfrac{1}{\sqrt{k}}$ for k odd, and $c_k = \dfrac{1}{k}$ for k even. The c_k are nonnegative and converge to 0. Then $\sum_{k=1}^{\infty} a_k$ converges by the Alternating Series Test, but the series $\sum_{k=1}^{\infty} c_k a_k$ diverges since the series of positive terms (the even ones) converges, and the series of negative terms (the odd ones) diverges (IV.2.5.6.(ii)).

(ii) Let $a_k = c_k = \dfrac{(-1)^k}{\sqrt{k}}$ for all k. Then $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} c_k$ both converge by the Alternating Series Test, but $\sum_{k=1}^{\infty} c_k a_k$ is the harmonic series, which diverges.

The simplest general result is for absolutely convergent series:

IV.2.7.2. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent series, and (c_k) a bounded sequence of real numbers. Then $\sum_{k=1}^{\infty} c_k a_k$ also converges absolutely.

This result applies in particular if all the a_k and c_k are nonnegative.

PROOF: If $|c_k| \le m$ for all k, compare $\sum_{k=1}^{\infty} |c_k a_k|$ with $\sum_{k=1}^{\infty} m|a_k|$. 👍

IV.2.7.3. Conversely, if $\sum_{k=1}^{\infty} a_k$ is an infinite series and (d_k) a sequence of real numbers such that $\inf |d_k| > 0$, and $\sum_{k=1}^{\infty} d_k a_k$ is absolutely convergent, then $\sum_{k=1}^{\infty} a_k$ is also absolutely convergent, seen by applying IV.2.7.2. to scaling by $c_k = \frac{1}{d_k}$.

On the other hand, if (d_k) is a sequence of positive numbers with $\inf d_k = 0$, there is an infinite series $\sum_{k=1}^{\infty} a_k$ with positive terms such that $\sum_{k=1}^{\infty} a_k$ diverges but $\sum_{k=1}^{\infty} d_k a_k$ converges absolutely. And if (c_k) is an unbounded sequence of positive numbers, there is an absolutely convergent series $\sum_{k=1}^{\infty} a_k$ with positive terms such that $\sum_{k=1}^{\infty} c_k a_k$ diverges. See Exercise IV.2.16.20..

Abel's Partial Summation

To analyze the case where $\sum_{k=1}^{\infty} a_k$ is not absolutely convergent, it is convenient to use a summation formula due to ABEL in 1826. This formula is a discrete version of Integration by Parts, and will be used roughly to interchange the roles of the a_k and c_k.

IV.2.7.4. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, with partial sums $s_n = \sum_{k=1}^{n} a_k$. Set $s_0 = 0$. If (c_k) is a sequence of real numbers, then for any $n \geq 0$, $m \geq 1$,

$$\sum_{k=n+1}^{n+m} c_k a_k = \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) s_k - c_{n+1} s_n + c_{n+m+1} s_{n+m} \ .$$

PROOF: For any $k \geq 1$ we have

$$c_k a_k = c_k (s_k - s_{k-1}) = (c_k - c_{k+1}) s_k - c_k s_{k-1} + c_{k+1} s_k$$

and summing from $n+1$ to $n+m$ the last two terms telescope to $c_{n+m+1} s_{n+m} - c_{n+1} s_n$. 👍

Essentially the same proof shows the following technical strengthening of the formula:

IV.2.7.5. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series, with partial sums $s_n = \sum_{k=1}^{n} a_k$. Set $s_0 = 0$. If $0 \leq r \leq n$, set $s_{n,r} = s_n - s_r = \sum_{k=r+1}^{n} a_k$. If (c_k) is a sequence of real numbers, then for any $n \geq 0$, $m \geq 1$,

and any r, $0 \le r \le n$,

$$\sum_{k=n+1}^{n+m} c_k a_k = \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) s_{k,r} - c_{n+1} s_{n,r} + c_{n+m+1} s_{n+m,r} \ .$$

PROOF: For any $k \ge r+1$ we have

$$c_k a_k = c_k(s_{k,r} - s_{k-1,r}) = (c_k - c_{k+1}) s_{k,r} - c_k s_{k-1,r} + c_{k+1} s_{k,r}$$

and summing from $n+1$ to $n+m$ the last two terms telescope to $c_{n+m+1} s_{n+m,r} - c_{n+1} s_{n,r}$. 👍

Proposition IV.2.7.4. is the case $r = 0$. The case $r = n$ is also useful (cf. V.15.3.16.). Note that if $r = n$, one term disappears from the right side since $s_{n,n} = 0$.

The first consequence of Abel's summation formula is the following general test, which we will call the *Scaling Test*, a nonstandard name:

IV.2.7.6. THEOREM. [SCALING TEST] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with partial sums s_n, and (c_k) a sequence of numbers. If

(i) $\sum_{k=1}^{\infty} (c_k - c_{k+1}) s_k$ converges and

(ii) $\lim_{m\to\infty} c_{m+1} s_m$ exists,

Then the series $\sum_{k=1}^{\infty} c_k a_k$ converges.

PROOF: Abel's partial summation formula for $n = 0$ and any m gives

$$\sum_{k=1}^{m} c_k a_k = \sum_{k=1}^{m} (c_k - c_{k+1}) s_k + c_{m+1} s_m \ .$$

By hypothesis, the limit of the right side exists as $m \to \infty$. 👍

Abel's Test

The first special case of the Scaling Test we consider is due to ABEL:

IV.2.7.7. THEOREM. [ABEL'S TEST] Let $\sum_{k=1}^{\infty} a_k$ be a convergent infinite series, and (c_k) a bounded monotone sequence of numbers. Then $\sum_{k=1}^{\infty} c_k a_k$ converges.

PROOF: We have that (c_k) converges, and since (s_m) converges by hypothesis, $(c_{m+1}s_m)$ converges. We also have that $\sum_{k=1}^{\infty}(c_k - c_{k+1})$ converges, and hence converges absolutely since the terms are all of the same sign. Since the s_k are bounded, $\sum_{k=1}^{\infty}(c_k - c_{k+1})s_k$ converges (absolutely) by IV.2.7.2.; thus the hypotheses of the Scaling Test are satisfied. 👍

Dirichlet's Test

Another consequence of the Scaling Test is called *Dirichlet's Test*, although it has been suggested that it was probably also known to ABEL previously:

IV.2.7.8. THEOREM. [DIRICHLET'S TEST] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with bounded partial sums, and (c_k) a nonincreasing sequence of (nonnegative) numbers converging to zero. Then $\sum_{k=1}^{\infty} c_k a_k$ converges.

The Alternating Series Test is the special case where $a_k = (-1)^k$ or $(-1)^{k+1}$.

PROOF: As in the previous proof, the series $\sum_{k=1}^{\infty}(c_k - c_{k+1})s_k$ converges absolutely, and we have $\lim_{m\to\infty} c_{m+1}s_m = 0$ by (). Thus the hypotheses of the Scaling Test are satisfied. 👍

As a nice application, we have:

IV.2.7.9. EXAMPLE. If x is any fixed real number, then the series

$$\sum_{k=1}^{\infty} \frac{\sin kx}{k}$$

converges.

Note that for general x, $\sin kx$ jumps around through $[-1, 1]$ in an apparently chaotic manner as k increases, so the terms vary in size and sign in a seemingly unpredictable way. But the essence of the argument is that there is enough regularity to insure that the partial sums of the series

$$\sum_{k=1}^{\infty} \sin kx$$

are bounded. We can then take $c_k = \frac{1}{k}$ and use Dirichlet's Test.

To show that these partial sums are bounded, we use a clever argument due to DIRICHLET involving the sum formula for $\cos(\alpha + \beta)$ (). By periodicity, it suffices to prove the result for $0 < x < 2\pi$ (the case $x = 0$ is trivial). Fix $x \in (0, 2\pi)$ and set

$$s_n = \sum_{k=1}^{n} \sin kx \ .$$

For any k, we have

$$\cos\left(\left[k - \frac{1}{2}\right] x\right) - \cos\left(\left[k + \frac{1}{2}\right] x\right)$$
$$= \left[\cos kx \cos\frac{x}{2} + \sin kx \sin\frac{x}{2}\right] - \left[\cos kx \cos\frac{x}{2} - \sin kx \sin\frac{x}{2}\right]$$
$$= 2 \sin kx \sin\frac{x}{2}$$

so we have, by telescoping,

$$2s_n \sin\frac{x}{2} = \sum_{k=1}^{n} 2 \sin kx \sin\frac{x}{2}$$
$$= \sum_{k=1}^{n} \left[\cos\left(\left[k - \frac{1}{2}\right] x\right) - \cos\left(\left[k + \frac{1}{2}\right] x\right)\right]$$
$$= \cos\frac{x}{2} - \cos\left(\left[n + \frac{1}{2}\right] x\right) \ .$$

Thus we have, for all n,

$$|s_n| \leq \frac{\left|\cos\frac{x}{2} - \cos\left(\left[n + \frac{1}{2}\right] x\right)\right|}{2\left|\sin\frac{x}{2}\right|} \leq \frac{1}{\left|\sin\frac{x}{2}\right|} \ .$$

In fact, we get that

$$\sum_{k=1}^{\infty} \frac{\sin kx}{k^p}$$

converges for any $x \in \mathbb{R}$ and $p > 0$, and that

$$\sum_{k=2}^{\infty} \frac{\sin kx}{\log k}$$

converges for any $x \in \mathbb{R}$.

IV.2.8. Subseries and Rearrangements of Series

There are two natural procedures for manufacturing a new infinite series out of an old one:

IV.2.8.1. DEFINITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series.

(i) An infinite series $\sum_{k=1}^{\infty} b_k$ is a *subseries* of $\sum_{k=1}^{\infty} a_k$ if there is a strictly increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that $b_k = a_{\phi(k)}$ for all k.

(ii) An infinite series $\sum_{k=1}^{\infty} b_k$ is a *rearrangement* of $\sum_{k=1}^{\infty} a_k$ if there is a bijection $\phi : \mathbb{N} \to \mathbb{N}$ such that $b_k = a_{\phi(k)}$ for all k.

(iii) An infinite series $\sum_{k=1}^{\infty} b_k$ is a *subarrangement* of $\sum_{k=1}^{\infty} a_k$ if there is a one-to-one function $\phi : \mathbb{N} \to \mathbb{N}$ such that $b_k = a_{\phi(k)}$ for all k.

A subseries or a rearrangement is a special case of a subarrangement of a series. We have seen natural examples of subseries in IV.2.5.5.. A natural example of a rearrangement of a series is to write the terms of a strictly positive convergent series $\sum_{k=1}^{\infty} a_k$ in nonincreasing order (check that this is a rearrangement: in particular, for any k there can be only finitely many j for which $a_j \geq a_k$).

Note that in a subarrangement, each term of the original series occurs at most once (if there are repeated terms in the original series, the term can be repeated in the subarrangement up to as many times as in the original series). In a rearrangement, each term in the original series occurs exactly once in the rearrangement.

The ϕ in a subarrangement is uniquely determined. (If there are repeated terms in the original series, more than one subarrangement can give the same new series.) If $\sum_{k=1}^{\infty} b_k$ is a rearrangement of $\sum_{k=1}^{\infty} a_k$ via ϕ, then $\sum_{k=1}^{\infty} a_k$ is a rearrangement of $\sum_{k=1}^{\infty} b_k$ via ϕ^{-1}.

Some authors rather colorfully call a rearrangement of a series a *derangement* of the series.

We have seen in IV.2.5.5. that a subseries of a convergent series need not converge. We will see (IV.2.8.8.) that, in contrast with IV.1.12.4., a rearrangement of a convergent series need not converge, and even if it does it need not have the same sum. These difficulties do not occur for convergent nonnegative series (or for absolutely convergent series – see IV.2.8.5.):

IV.2.8.2. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be a convergent nonnegative infinite series. Then any subarrangement $\sum_{k=1}^{\infty} b_k$ of $\sum_{k=1}^{\infty} a_k$ also converges, and satisfies

$$\sum_{k=1}^{\infty} b_k \leq \sum_{k=1}^{\infty} a_k \ .$$

PROOF: Let $M = \sum_{k=1}^{\infty} a_k$. Any partial sum t_n of the nonnegative series $\sum_{k=1}^{\infty} b_k$ consists of some, but not necessarily all, of the terms in a corresponding larger partial sum s_m of $\sum_{k=1}^{\infty} a_k$ (where $m = \max\{\phi(1), \ldots, \phi(n)\}$), and hence $t_n \le s_m \le M$. Thus the partial sums of $\sum_{k=1}^{\infty} b_k$ are bounded by M, so $\sum_{k=1}^{\infty} b_k \le M$. 👍

IV.2.8.3. COROLLARY. Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent infinite series. Then any subarrangement $\sum_{k=1}^{\infty} b_k$ of $\sum_{k=1}^{\infty} a_k$ also converges absolutely.

PROOF: $\sum_{k=1}^{\infty} |b_k|$ is a subarrangement of $\sum_{k=1}^{\infty} |a_k|$. 👍

IV.2.8.4. COROLLARY. Let $\sum_{k=1}^{\infty} a_k$ be a convergent nonnegative infinite series. Then any rearrangement $\sum_{k=1}^{\infty} b_k$ of $\sum_{k=1}^{\infty} a_k$ also converges, and satisfies

$$\sum_{k=1}^{\infty} b_k = \sum_{k=1}^{\infty} a_k \ .$$

This result generalizes to absolutely convergent series:

IV.2.8.5. THEOREM. Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent infinite series. Then any rearrangement $\sum_{k=1}^{\infty} b_k$ of $\sum_{k=1}^{\infty} a_k$ also converges absolutely, and satisfies

$$\sum_{k=1}^{\infty} b_k = \sum_{k=1}^{\infty} a_k \ .$$

PROOF: $\sum_{k=1}^{\infty} b_k$ converges absolutely by IV.2.8.3..

Let $s = \sum_{k=1}^{\infty} a_k$, and write $s_n = \sum_{k=1}^{n} a_k$ and $t_n = \sum_{k=1}^{n} b_k$ for the partial sums of the series. We will show that $t_n \to s$, i.e. for every $\epsilon > 0$ there is an N such that $|t_n - s| < \epsilon$ for all $n > N$.

Let $\epsilon > 0$, and fix m such that

$$|s_m - s| < \frac{\epsilon}{2} \quad \text{and} \quad \sum_{k=m+1}^{\infty} |a_k| < \frac{\epsilon}{2}$$

(the second condition is possible since $\sum_{k=1}^{\infty} a_k$ converges absolutely; in fact, the first condition follows automatically from the second). Now let

$$N = \max\{\phi(k) : 1 \leq k \leq m\}$$

where $\phi : \mathbb{N} \to \mathbb{N}$ is the bijection of the rearrangement. Now suppose $n > N$. We have that

$$t_n - s_m = \sum_{k=1}^{n} b_k - \sum_{k=1}^{m} a_k$$

is a sum consisting of various a_k's and their negatives. Some a_k's occur in both sums and therefore cancel; this happens for all a_k with $k \leq m$ by the choice of N. Thus we have

$$t_n - s_m = \sum_{j=1}^{r} (\pm a_{k_j})$$

for some r and some k_j's, with $k_j > m$ for all j. So

$$|t_n - s_m| \leq \sum_{j=1}^{r} |a_{k_j}| \leq \sum_{k=m+1}^{\infty} |a_k| < \frac{\epsilon}{2}$$

and we get

$$|t_n - s| \leq |t_n - s_m| + |s_m - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

for all $n > N$. 👍

IV.2.8.6. This is in stark contrast to what happens in a conditionally convergent series. The sum of a conditionally convergent series can change if rearranged, and the rearranged series can even diverge.

> "It is easy to imagine the effect that the discovery of this apparent paradox must have had on the mathematicians of the eighteenth century, who were accustomed to operate with infinite series without regard to their convergence."
>
> *R. Courant*[2]

[2] [CJ99, p. 374].

If $\sum_{k=1}^{\infty} a_k$ is conditionally convergent, then the series $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ of positive and negative terms, respectively, both diverge. Thus if the series is rearranged to begin with many (or few) positive terms, then many (or few) negative ones, etc., the partial sums of the rearranged series can be made to oscillate wildly or to converge to any limit at all. Precisely:

IV.2.8.7. THEOREM. Let $\sum_{k=1}^{\infty} a_k$ be a conditionally convergent series, and let (u_n) and (v_n) be sequences of real numbers with $v_n \le u_n$ and $v_n \le u_{n+1}$ for all n. Then there is a rearrangement $\sum_{k=1}^{\infty} d_k$ of $\sum_{k=1}^{\infty} a_k$ and natural numbers $n_1 < m_1 < n_2 < m_2 < \cdots$ such that

(i) The partial sums t_n of $\sum_{k=1}^{\infty} d_k$ increase for $1 \le n \le n_1$ and for $m_j \le n \le n_{j+1}$ for all j, and decrease for $n_j \le n \le m_j$ for all j.

(ii) $t_{n_j} \ge u_j$ and $t_{m_j} \le v_j$ for all j.

(iii) $\limsup_{n\to\infty} t_n = \limsup_{j\to\infty} u_j$ and $\liminf_{n\to\infty} t_n = \liminf_{j\to\infty} v_j$.

PROOF: Set $d_1 = b_1$, and $d_k = b_k$ until the first index n_1 is reached with $t_{n_1} = \sum_{k=1}^{n_1} b_k \ge u_1$ (there is such an n_1 since $\sum_{k=1}^{\infty} b_k$ diverges). Then $u_1 \le t_{n_1} \le u_1 + b_{n_1}$. Set $d_{n_1+1} = c_1$, and $d_{n_1+k} = c_k$ until the first index $m_1 > n_1$ is reached with $t_{m_1} \le v_1$ (there is such an m_1 since $\sum_{k=1}^{\infty} c_k$ diverges). Then $v_1 + c_{m_1} \le t_{m_1} \le v_1$. Set $d_{m_1+1} = b_{n_1+1}$ and $d_{m_1+k} = b_{n_1+k}$ until the first index $n_2 > m_1$ is reached with $t_{n_2} \ge u_2$. Continue the process inductively to define d_k, n_j, and m_j for all j; the construction is possible at each step since all tails of $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ diverge. We have $u_j \le t_{n_j} \le u_j + b_{r_j}$, where $r_j = n_1 + \sum_{i=1}^{j-1}(n_{i+1} - m_i)$ is the number of b_k used in the first n_j steps, and similarly $v_j + c_{s_j} \le t_{m_j} \le v_j$, where $s_j = \sum_{i=1}^{j}(m_i - n_i)$ is the number of c_k used. Then r_j, s_j are strictly increasing, so $b_{r_j} \to 0$ and $c_{s_j} \to 0$ since $a_k \to 0$.

By construction, (i) and (ii) are satisfied for the rearrangement $\sum_{k=1}^{\infty} d_k$ of $\sum_{k=1}^{\infty} a_k$. We need only check (iii). It is clear that

$$\limsup_{n\to\infty} t_n = \limsup_{j\to\infty} t_{n_j}$$

and we have

$$u := \limsup_{j\to\infty} u_j \le \limsup_{j\to\infty} t_{n_j} \le \limsup_{j\to\infty} (u_j + b_{r_j}) = u$$

by () since $b_{r_j} \to 0$. Thus $\limsup_{n\to\infty} t_n = u$. The argument for lim inf is similar. 👍

IV.2.8.8. Corollary. Let $\sum_{k=1}^{\infty} a_k$ be a conditionally convergent series.

(i) For any $t \in \mathbb{R}$ there is a rearrangement $\sum_{k=1}^{\infty} d_k$ of $\sum_{k=1}^{\infty} a_k$ which converges with sum t.

(ii) For any u and v, $-\infty \leq v \leq u \leq +\infty$, there is a rearrangement $\sum_{k=1}^{\infty} d_k$ of $\sum_{k=1}^{\infty} a_k$ whose partial sums t_n satisfy

$$\limsup_{n\to\infty} t_n = u \quad \text{and} \quad \liminf_{n\to\infty} t_n = v \ .$$

Proof: Part (i) is the special case of (ii) with $u = v = t$. For (ii), if u and/or v is in $\mathbb{R}$ take $u_j = u$ and/or $v_j = v$ for all j; if u and/or v is $\pm\infty$, take u_j and/or v_j to be $\pm j$. 👍

This result was obtained by Riemann in (), although examples were known earlier. The argument applies more generally to any series $\sum_{k=1}^{\infty} a_k$ in which $a_k \to 0$ for which both the series of positive terms and of negative terms diverge.

> "The bottom line is that the analysts taught us to be wary of conditionally convergent series. Convergence is good, but not good enough. For an infinite series to behave like a finite sum in all respects, it needs much tighter constraints than conditional convergence can provide. Insisting on absolute convergence yields the behavior we'd expect intuitively."
>
> *S. Strogatz*[3]

This principle is one reason that Lebesgue integration is an analog (and generalization) of the theory of *absolutely* convergent series. Infinite series of functions, e.g. Fourier series, are generally only well behaved when they converge absolutely. Fortunately, for power series the notions of convergence and absolute convergence are effectively the same (cf. V.15.3.5.).

Revisions of Infinite Series

IV.2.8.9. Definition. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series. A *revision* of the series is a contraction (IV.2.1.20.) of a rearrangement (IV.2.8.1.) of the series.

IV.2.8.10. A revision of the series is thus a new series

$$(a_{k_{1,1}} + a_{k_{1,2}} + \cdots + a_{k_{1,n_1}}) + (a_{k_{2,1}} + a_{k_{2,2}} + \cdots + a_{k_{2,n_2}}) + \cdots$$

where each k occurs as exactly one $k_{i,j}$, i.e. each group consists of a finite number of terms of the series, with each term occurring exactly once (in exactly one group). Then by IV.2.1.21. and IV.2.8.5. we obtain:

[3] [Str13, p. 304].

IV.2.8.11. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent infinite series. Then any revision of the series converges, with the same limit.

IV.2.8.12. In this result, each group is allowed to include only finitely many terms of the series. In IV.2.10.16. we will show the result holds also for more drastic revisions where the groups are allowed to be infinite.

IV.2.9. Unordered Summations

IV.2.9.1. Let J be a set, possibly infinite, possibly even uncountable, and let $\{a_j : j \in J\}$ be an indexed collection of real (or complex) numbers. We want to make sense out of the expression $\sum_{j\in J} a_j$. If J is finite, there is an unambiguous meaning, but if J is infinite, there is no direct meaning. We have made sense of such a sum if $J = \mathbb{N}$, but the theory of infinite series depends critically on having a fixed ordering on $\mathbb{N}$. A general set J has no natural ordering, and even if a particular J has an ordering (e.g. if $J = \mathbb{N}$) we do not want our definition to depend on the ordering. It is to be expected from the case $J = \mathbb{N}$ that we can only make sense of the sum under certain conditions, not in general.

Infinite series give a clue as to how to proceed: we will take a limit of sums over finite subsets of J.

IV.2.9.2. DEFINITION. Let J be a set, and let $\{a_j : j \in J\}$ be an indexed collection of real numbers. We say the *unordered sum* $\sum_{j\in J} a_j$ *converges with sum* s if for every $\epsilon > 0$ there is a finite subset F of J such that, for every finite subset E of J containing F,

$$\left| s - \sum_{j\in E} a_j \right| < \epsilon \ .$$

When we write $\sum_{j\in J} a_j = s$, we mean that the unordered sum converges to s.

IV.2.9.3. Note that if $J = \mathbb{N}$, this definition is not the same as the usual definition of convergence of the sum as an infinite series, since for an infinite series to converge only the partial sums, which are sums of initial segments, must converge, not, for example, finite sums made up mostly of even-numbered terms. If (a_k) is a sequence of real numbers, and the unordered sum $\sum_{k\in\mathbb{N}} a_k$ converges with sum s, then the infinite series

$$\sum_{k=1}^{\infty} a_k$$

also converges, with sum s. But we will see that the converse is false in general; in fact, it follows easily from IV.2.8.5. and IV.2.8.8. that the unordered sum converges if and only if the infinite series converges absolutely (cf. IV.2.9.9.).

IV.2.9.4. PROPOSITION. Let (a_k) be a sequence of real numbers.

(i) If the unordered sum $\sum_{k\in\mathbb{N}} a_k$ converges, then the infinite series $\sum_{k=1}^{\infty} a_n$ converges, and the sums are the same.

(ii) If all the a_k are nonnegative, then the unordered sum $\sum_{k\in\mathbb{N}} a_k$ converges if and only if the infinite series $\sum_{k=1}^{\infty} a_n$ converges.

PROOF: (i) Let s_n be the n'th partial sum of the infinite series. If the unordered sum converges to s, and $\epsilon > 0$, choose F so that

$$\left| s - \sum_{k\in E} a_k \right| < \epsilon$$

for all finite E containing F. If n is large enough that $F \subseteq \{1, \ldots, n\}$, then $|s - s_n| < \epsilon$. Thus the infinite series converges with sum s.

(ii) Suppose all a_k are nonnegative, and that the infinite series converges to s. If $\epsilon > 0$, choose n so that $|s - s_n| < \epsilon$. If E is any finite subset of $\mathbb{N}$ containing $F = \{1, \ldots, n\}$, and m is the largest number in E, then

$$s_n \leq \sum_{k\in E} a_k \leq s_m \leq s$$

and so

$$\left| s - \sum_{k\in E} a_k \right| < \epsilon$$

and $\sum_{k\in\mathbb{N}} a_k$ converges with sum s. 👍

The first observation is an analog of the Divergence Criterion:

IV.2.9.5. PROPOSITION. Let J be a set, and let $\{a_j : j \in J\}$ be an indexed collection of real numbers. If $\sum_{j\in J} a_j$ converges, then for any $\epsilon > 0$ there are only finitely many j for which $|a_j| > \epsilon$.

PROOF: Let s be the sum of the series, and $\epsilon > 0$. Then there is a finite subset F of J such that

$$\left| s - \sum_{j\in E} a_j \right| < \frac{\epsilon}{2}$$

for all finite subsets E of J containing F. Let $j_0 \in J \setminus F$, and $E = F \cup \{j_0\}$. Then

$$|a_{j_0}| = \left| \sum_{j\in E} a_j - \sum_{j\in F} a_j \right| \leq \left| \sum_{j\in E} a_j - s \right| + \left| s - \sum_{j\in F} a_j \right| < \epsilon \ .$$

This is true for all $j_0 \notin F$. 👍

This has an interesting consequence. The index set J could very well be uncountable, but we can effectively reduce to the case where J is countable by the following result:

IV.2.9.6. PROPOSITION. Let J be a set, and let $\{a_j : j \in J\}$ be an indexed collection of real numbers. If $\sum_{j\in J} a_j$ converges, then the set $K = \{j \in J : a_j \neq 0\}$ is countable and $\sum_{j\in K} a_j$ converges to $\sum_{j\in J} a_j$.

PROOF: By IV.2.9.5., the set $K_n = \left\{j \in J : |a_j| > \frac{1}{n}\right\}$ is finite for each n, and $K = \cup_n K_n$. If F is any finite subset of J, then $\sum_{j\in F} a_j = \sum_{j\in F\cap K} a_j$. 👍

As with infinite series, we can extend IV.2.9.5. to obtain a Cauchy criterion:

IV.2.9.7. PROPOSITION. Let J be a set, and let $\{a_j : j \in J\}$ be an indexed collection of real numbers. Then $\sum_{j\in J} a_j$ converges if and only if, for any $\epsilon > 0$, there is a finite subset F of J such that

$$\left|\sum_{j\in D} a_j\right| < \epsilon$$

for every finite subset D of J disjoint from F.

PROOF: Suppose the unordered sum converges to s, and let $\epsilon > 0$. Then there is a finite subset F of J such that

$$\left|s - \sum_{j\in E} a_j\right| < \frac{\epsilon}{2}$$

for all finite subsets E of J containing F. Let D be a finite subset of J disjoint from F. Then

$$\left|\sum_{j\in D} a_j\right| = \left|\sum_{j\in D\cup F} a_j - \sum_{j\in F} a_j\right| \leq \left|\sum_{j\in D\cup F} a_j - s\right| + \left|s - \sum_{j\in F} a_j\right| < \epsilon\ .$$

Conversely, suppose the other condition holds. Choose an increasing sequence (F_n) of finite subsets of J such that

$$\left|\sum_{j\in D} a_j\right| < \frac{1}{n}$$

for all finite D disjoint from F_n. Then, for any finite E containing F_n, we have

$$\left|\sum_{j\in E} a_j - \sum_{j\in F_n} a_j\right| = \left|\sum_{j\in E\setminus F_n} a_j\right| < \frac{1}{n}\ .$$

set

$$\alpha_n = \inf\left\{\sum_{j\in E} a_j : F_n \subseteq E \text{ finite}\right\} \geq \sum_{j\in F_n} a_j - \frac{1}{n}$$

$$\beta_n = \sup\left\{\sum_{j\in E} a_j : F_n \subseteq E \text{ finite}\right\} \leq \sum_{j\in F_n} a_j + \frac{1}{n}$$

and set $I_n = [\alpha_n, \beta_n]$. We have $I_{n+1} \subseteq I_n$ since $F_n \subseteq F_{n+1}$. Then $\cap_n I_n = \{s\}$ for some $s \in \mathbb{R}$, and if E is any finite set containing F_n we have

$$\left| s - \sum_{j\in E} a_j \right| \leq \frac{2}{n}$$

(both numbers are in I_n), which can be made arbitrarily small. Thus $\sum_{j\in J} a_j$ converges to s. 👍

IV.2.9.8. We can extend the notion of unordered sum to arbitrary nonnegative sums. Let J be a set, and let $\{a_j : j \in J\}$ be an indexed collection of real numbers. If all the a_j are nonnegative, we can always make sense of the unordered sum as

$$\sum_{j\in J} a_j = \sup \sum_{j\in F} a_j$$

where the supremum is over all finite subsets F of J. If the supremum is $s < +\infty$, this definition agrees with the previous one: given $\epsilon > 0$, choose a finite F with

$$\left| s - \sum_{j\in F} a_j \right| < \epsilon$$

and this F works in the definition of convergence. If the supremum is $+\infty$, we will not say the unordered sum converges, but we will write $\sum_{j\in J} a_j = +\infty$.

We can now state the main result of this subsection:

IV.2.9.9. THEOREM. Let J be a set, and let $\{a_j : j \in J\}$ be an indexed collection of real numbers. Then $\sum_{j\in J} a_j$ converges if and only if $\sum_{j\in J} |a_j|$ converges (i.e. $\sum_{j\in J} |a_j| < +\infty$).

PROOF: If $\sum_{j\in J} |a_j|$ converges, then for any $\epsilon > 0$ there is a finite set F such that

$$\sum_{j\in D} |a_j| < \epsilon$$

for every finite subset D of J disjoint from F by IV.2.9.7.. But then

$$\left|\sum_{j\in D} a_j\right| \leq \sum_{j\in D} |a_j| < \epsilon$$

for every finite subset D of J disjoint from F, so $\sum\limits_{j\in J} a_j$ converges by IV.2.9.7..

Conversely, suppose $\sum\limits_{j\in J} a_j$ converges. Let F be a finite subset of J such that

$$\left|\sum_{j\in D} a_j\right| < 1$$

for every finite subset D of J disjoint from F. Suppose $\sum\limits_{j\in J} |a_j| = +\infty$; then there are finite subsets D of J disjoint from F with $\sum\limits_{j\in D} |a_j|$ arbitrarily large. In particular, there is such a D with $\sum\limits_{j\in D} |a_j| > 2$. For this D, let $D_+ = \{j \in D : a_j > 0\}$, $D_- = \{j \in D : a_j < 0\}$. Then either $\left|\sum\limits_{j\in D_+} a_j\right| = \sum\limits_{j\in D_+} |a_j| > 1$ or $\left|\sum\limits_{j\in D_-} a_j\right| = \sum\limits_{j\in D_-} |a_j| > 1$, a contradiction. Thus $\sum\limits_{j\in J} |a_j| < +\infty$. 👍

Thus there is no difference between convergence and "absolute convergence" for unordered sums.

IV.2.9.10. Corollary. Let J be a countable index set, and $\{a_j : j \in J\}$ an indexed set of numbers. Then the following are equivalent:

(i) The unordered sum $\sum\limits_{j\in J} a_j$ converges.

(ii) The infinite series $\sum\limits_{k=1}^{\infty} a_{\phi(k)}$ converges absolutely for some bijection $\phi : \mathbb{N} \to J$.

(iii) The infinite series $\sum\limits_{k=1}^{\infty} a_{\phi(k)}$ converges absolutely for every bijection $\phi : \mathbb{N} \to J$.

If the series converge, the sums are all equal.

Thus a convergent unordered sum can be converted to an ordinary infinite series by ordering the terms in an arbitrary way; the resulting infinite series will converge absolutely and have the same sum independently of how the terms are ordered.

IV.2.9.11. Corollary. Let (a_k) be a sequence of real numbers. Then the unordered series $\sum\limits_{k\in\mathbb{N}} a_k$ converges if and only if the infinite series $\sum\limits_{k=1}^{\infty} a_k$ converges absolutely. If the series converge, the sums are equal.

IV.2.10. Double Summations

Suppose we have a double array

$$\begin{array}{cccc} a_{11} & a_{12} & a_{13} & \cdots \\ a_{21} & a_{22} & a_{23} & \cdots \\ a_{31} & a_{32} & a_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

of numbers, and we want to "add them all up." There are many potentially reasonable ways to proceed, and we might hope (or even expect from commutativity of addition) that all these ways give the same answer. Experience with rearranging conditionally convergent series might make us suspicious that this works, and these suspicions are well founded: the various methods do not always give the same answer. But if the sums are all "absolutely convergent" in an appropriate sense, the methods all give the same sum.

IV.2.10.1. A first, somewhat naive, approach is to simply write all the elements down in some order to make an ordinary infinite series, and take the sum of this series (if it exists). More precisely, take the infinite series

$$\sum_{k=1}^{\infty} a_{\phi(k)}$$

where ϕ is a bijection from $\mathbb{N}$ to $\mathbb{N}\times\mathbb{N}$. Of course, there are many ways to do this, and the series obtained are rearrangements of one another (and any rearrangement of one of the series is another such series). If one such series converges absolutely, so does any other and all have the same sum (IV.2.8.5.), so the procedure seems to be satisfactory; but if the series do not converge absolutely, the answers will differ drastically (IV.2.8.8.), and not all will even converge.

IV.2.10.2. As a variation, we could consider the unordered sum (IV.2.9.2.)

$$\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} \ .$$

This unordered sum will converge if and only if the sums in IV.2.10.1. converge absolutely (IV.2.9.11.), with the same sum. This unordered sum can be regarded as the *double sum* of the array.

IV.2.10.3. A somewhat different approach is to consider sums of the form

$$\sum_{j,k=1}^{n} a_{jk} = \sum_{j=1}^{n} \left[\sum_{k=1}^{n} a_{jk} \right]$$

and let $n \to \infty$. These are finite sums, so each sum can be computed in any order; in particular, we have

$$\sum_{k=1}^{n}\sum_{j=1}^{n} a_{jk} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk} := \sum_{j=1}^{n} \left[\sum_{k=1}^{n} a_{jk} \right]$$

(such double sums are usually written without the brackets). These sums are often called *square sums* since the terms in a square in the upper left-hand corner of the array are added. As a variation, we could take

rectangular sums

$$\sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk}$$

and take a limit as $m, n \to \infty$.

The rectangular sum procedure is less likely to work (i.e. to have a limit), but if it does the square sum procedure will work and give the same sum. If the unordered sum procedure converges (i.e. the procedure of IV.2.10.1. gives absolutely convergent series), these rectangular and square sum procedures will also work and give the same sum, since they are limits of partial sums of contractions of certain of the series from IV.2.10.1..

IV.2.10.4. Another variation is to consider

$$\sum_{n=2}^{\infty}\sum_{j+k=n} a_{jk} \ .$$

This is called a *triangular sum*, since the terms on each northeast-southwest diagonal are first added (a finite sum) and then the diagonal sums are added. This procedure is particularly natural when products of power series are considered (IV.2.11.4.). If the unordered sum converges, this procedure will also converge and give the same sum, for the same reason as for square and rectangular sums.

IV.2.10.5. As a quite different procedure, we can sum along each row, and then add the row sums, or sum along each column and then add the column sums. Specifically, for each fixed j we can set

$$r_j = \sum_{k=1}^{\infty} a_{jk}$$

if the series converges; r_j is the sum of the terms in the j'th row. We then take

$$\sum_{j=1}^{\infty} r_j$$

which is usually written

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}$$

(when an expression like this is written, it is implicitly assumed that the inside sum converges for each j).

We can similarly first add each column: for each fixed k set

$$c_k = \sum_{j=1}^{\infty} a_{jk}$$

if the series converges; c_k is the sum of the terms in the k'th column. We then take

$$\sum_{k=1}^{\infty} c_k = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{jk} \ .$$

It is not immediately clear whether these two iterated summation procedures are well defined (i.e. all the infinite series will converge), or give the same answer if they are; and the relation between the convergence of these procedures and the previous ones is also not obvious. The main results of this section will establish the relationship in good cases.

Note that it is not quite correct to call such iterated summations infinite series; they are a different (but closely related) type of infinite sum which involves many infinite series.

IV.2.10.6. Example. Here is the standard example illustrating some of what can go wrong. Set

$$a_{jk} = \begin{cases} 1 & \text{if} \quad j = k \\ -1 & \text{if} \quad j = k+1 \\ 0 & \text{otherwise} \end{cases}$$

i.e. the terms on the main diagonal are all 1, the terms on the first subdiagonal are -1, and all other terms are 0. The row and column sums each have only finitely many nonzero terms (no more than 2), hence converge absolutely. But we have $r_1 = 1$, $r_j = 0$ for $j > 1$, and $c_k = 0$ for all k, so

$$1 = \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk} \neq \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{jk} = 0 \ .$$

The square sum procedure gives a sum of 1; but the rectangular and triangular procedures do not converge since the rectangular or triangular sums can be either 0 or 1. The unordered sum diverges in this case; the procedure of IV.2.10.1. always gives a divergent series since the terms do not approach 0.

To examine the convergence of such iterated summations, we first consider the case where all the terms are nonnegative. In this case we can make sense of any iterated sum by taking the convention that an infinite series with nonnegative terms and at least one term $+\infty$ has sum $+\infty$. The next result is a special case of Tonelli's Theorem for integrals on product spaces ().

IV.2.10.7. Theorem. [Tonelli's Theorem for Double Summations] Let $\{a_{jk} : j, k \in \mathbb{N}\}$ be an array with all $a_{jk} \geq 0$. If one of the three sums

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}, \quad \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{jk}, \quad \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk}$$

converges (i.e. is finite), then all three converge and have the same sum.

Proof: First note that for any $m, n \in \mathbb{N}$,

$$\sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} \leq \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} = \sup_{F\subseteq\mathbb{N}\times\mathbb{N} \text{ finite}} \sum_{(j,k)\in F} a_{jk}$$

since $\{1, \ldots, m\} \times \{1, \ldots, n\}$ is a finite subset of $\mathbb{N} \times \mathbb{N}$. Conversely, for any finite $F \subseteq \mathbb{N} \times \mathbb{N}$, we have

$$\sum_{(j,k)\in F} a_{jk} \leq \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk}$$

for sufficiently large m, n (large enough that $F \subseteq \{1, \dots, m\} \times \{1, \dots, n\}$); thus

$$\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} = \sup_{m,n\in\mathbb{N}} \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} \ .$$

Also, for any m, n,

$$\sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} \leq \sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} \leq \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}$$

so we get

$$\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} = \sup_{m,n\in\mathbb{N}} \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} \leq \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk} \ .$$

For the reverse inequality, suppose

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk} = \sup_{m\in\mathbb{N}} \sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} = s \leq +\infty$$

and fix $r < s$. Then there is an m such that

$$r < \sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} \ .$$

But by repeated applications of IV.2.1.16., we have

$$\sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} = \sum_{k=1}^{\infty}\sum_{j=1}^{m} a_{jk}$$

so

$$\sum_{k=1}^{\infty}\sum_{j=1}^{m} a_{jk} = \sup_{n\in\mathbb{N}} \sum_{k=1}^{n}\sum_{j=1}^{m} a_{jk} > r$$

and thus there is an n such that

$$\sum_{k=1}^{n}\sum_{j=1}^{m} a_{jk} > r$$

and we conclude that

$$\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} = \sup_{m,n\in\mathbb{N}} \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} \geq s \ .$$

Thus

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk} = \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} \ .$$

A nearly identical argument shows that

$$\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{jk} = \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} \ .$$

We now turn to the general absolutely convergent case. This result is a special case of Fubini's Theorem for integrals on product spaces ().

IV.2.10.8. THEOREM. [FUBINI'S THEOREM FOR DOUBLE SUMMATIONS] Let $\{a_{jk} : j, k \in \mathbb{N}\}$ be an array of numbers. If one (hence each by IV.2.10.7.) of the three sums

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}|a_{jk}|, \quad \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}|a_{jk}|, \quad \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}}|a_{jk}|$$

converges (i.e. is finite), then each of the three summations

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}a_{jk}, \quad \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}a_{jk}, \quad \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}}a_{jk}$$

converges absolutely, and the sums are equal. In particular, the iterated summations are equal.

PROOF: The argument is similar to the proof of IV.2.8.5.. First note that the three sums in the conclusion (and all the infinite series in the iterated summations) converge absolutely, hence converge, so we need only show the three sums are equal. We will show

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}a_{jk} = s := \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}}a_{jk}$$

and the proof that the other iterated sum equals s is almost identical.

Let $\epsilon > 0$. Then, since $\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}}|a_{jk}| < \infty$, there is a finite set $F \subseteq \mathbb{N}\times\mathbb{N}$ such that

$$\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}\setminus F}|a_{jk}| < \frac{\epsilon}{3}$$

and $M, N \in \mathbb{N}$ such that $F \subseteq \{1,\ldots,M\}\times\{1,\ldots,N\}$. Thus if $m \geq M$, $n \geq N$,

$$\left|s - \sum_{j=1}^{m}\sum_{k=1}^{n}a_{jk}\right| < \frac{\epsilon}{3}\ .$$

We have

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}a_{jk} = \lim_{m\to\infty}\sum_{j=1}^{m}\sum_{k=1}^{\infty}a_{jk}\ .$$

Fix $m \geq M$ such that

$$\left|\sum_{j=1}^{m}\sum_{k=1}^{\infty}a_{jk} - \sum_{j=1}^{\infty}\sum_{k=1}^{\infty}a_{jk}\right| < \frac{\epsilon}{3}\ .$$

Since m is fixed, we have

$$\sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} = \lim_{n\to\infty} \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk}$$

by (), so there is an $n \geq N$ such that

$$\left| \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} - \sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} \right| < \frac{\epsilon}{3} \ .$$

We then have

$$\left| s - \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk} \right|$$

$$\leq \left| s - \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} \right| + \left| \sum_{j=1}^{m}\sum_{k=1}^{n} a_{jk} - \sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} \right| + \left| \sum_{j=1}^{m}\sum_{k=1}^{\infty} a_{jk} - \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk} \right| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$$

and since $\epsilon > 0$ is arbitrary, the iterated sum equals s. 

IV.2.10.9. We have retained the commonly-used names "Tonelli's Theorem" and "Fubini's Theorem" for IV.2.10.7. and IV.2.10.8. since they are special cases of the general results commonly called "Tonelli's Theorem" () and "Fubini's Theorem" (), even though it is likely that the double sum results predated FUBINI and TONELLI (at least IV.2.10.7. was undoubtedly "known" to EULER, although it was probably only really proved later). I have been unable to discover who first explicitly stated and proved the double sum versions (very likely sometime in the nineteenth century). See IX.3.4.34. for additional historical comments.

IV.2.10.10. Actually, we must be careful in saying what it means for an iterated summation 

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}$$

to converge absolutely. The strong (and proper) interpretation is that the iterated summation

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} |a_{jk}|$$

should converge. Fubini's Theorem IV.2.10.8. says that if an iterated sum

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}$$

converges absolutely in this sense, then the iterated sum

$$\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{jk}$$

in the opposite order (which also converges absolutely by IV.2.10.7.) has the same sum, i.e. the order of summation can be reversed.

We could instead consider a weaker sense in which an iterated summation

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}$$

converges absolutely: we could just require

(i) For each j, the infinite series $\displaystyle\sum_{k=1}^{\infty} a_{jk}$ converges absolutely with sum b_j.

(ii) The infinite series $\displaystyle\sum_{j=1}^{\infty} b_j$ converges absolutely, i.e.

$$\sum_{j=1}^{\infty}\left|\sum_{k=1}^{\infty} a_{jk}\right|$$

converges.

But example IV.2.10.6. shows that even if both iterated summations

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}, \quad \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{jk}$$

converge absolutely in this weak sense, the sums may not be equal. In fact, if the iterated summation in one order converges absolutely in the weak sense, the iterated sum in the other order need not, and need not even make sense: consider, for example, the case where for all j $a_{j1} = 1$, $a_{j2} = -1$, and $a_{jk} = 0$ for all $k > 2$.

Combining IV.2.10.8. with the previous observations, we obtain:

IV.2.10.11. Corollary. Let $\{a_{jk} : j, k \in \mathbb{N}\}$ be an array of numbers. If one (hence each) of the three sums

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} |a_{jk}|, \quad \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} |a_{jk}|, \quad \sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} |a_{jk}|$$

converges (i.e. is finite), then each of the following summations converges absolutely, and the sums are equal:

(i) $\displaystyle\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} a_{jk}$.

(ii) $\displaystyle\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} a_{jk}$.

(iii) $\displaystyle\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk}$.

(iv) The rectangular sum $\lim_{m,n\to\infty} \sum_{j=1}^{m} \sum_{k=1}^{n} a_{jk}$.

(v) The square sum $\lim_{n\to\infty} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk}$.

(iv) The triangular sum $\sum_{n=2}^{\infty} \sum_{j+k=n} a_{jk}$.

(vii) Any sum of the form $\sum_{k=1}^{\infty} a_{\phi(k)}$ where ϕ is a bijection from $\mathbb{N}$ to $\mathbb{N} \times \mathbb{N}$.

IV.2.10.12. These results are stated for double summations, but by repeated applications of the double summation result the same conclusions hold for triple or n-fold summations for any n.

Drastic Revisions of an Infinite Series

IV.2.10.13. A *drastic revision* of an infinite series

$$\sum_{k=1}^{\infty} a_k$$

is a sum of the form

$$(a_{k_{1,1}} + a_{k_{1,2}} + \cdots) + (a_{k_{2,1}} + a_{k_{2,2}} + \cdots) + \cdots$$

which resembles a revision (IV.2.8.9.) of the series except that the groups are allowed to be infinite. More precisely:

IV.2.10.14. Definition. A *drastic revision* of the infinite series $\sum_{k=1}^{\infty} a_k$ is an iterated summation of the form

$$\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_{\phi_j(k)}$$

where $\phi_1, \phi_2, \ldots$ is a sequence of one-to-one functions from $\mathbb{N}$ to $\mathbb{N}$ with disjoint ranges whose union is all of $\mathbb{N}$.

IV.2.10.15. We can also allow some of the groups to be finite; such a sum is technically a drastic revision of a series obtained by inserting an infinite number of 0's into the original series, which makes no essential change to the series. Thus an ordinary revision is a drastic revision.

The main result about drastic revisions is:

IV.2.10.16. THEOREM. Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent infinite series. Then any drastic revision of the series converges absolutely, and has the same sum as the original series.

PROOF: This is an almost immediate corollary of Fubini's Theorem. Set $a_{jk} = a_{\phi_j(k)}$. Since the original series converges absolutely, we have that the unordered sum

$$\sum_{(j,k)\in\mathbb{N}\times\mathbb{N}} a_{jk} = \sum_{k\in\mathbb{N}} a_k$$

converges absolutely to the sum of the original series by IV.2.9.11. and IV.2.9.4.(i), and hence the iterated sum does too by IV.2.10.8.. 👍

IV.2.10.17. Fubini's Theorem is usually thought of as a result justifying interchanging the order of summation or integration in an iterated sum or integral. But IV.2.10.16. is an interesting and useful application not involving interchange of order; the relevant part of Fubini's Theorem is that the iterated sum is equal to the unordered sum (double sum).

IV.2.10.18. We could more generally consider drastic revisions of drastic revisions of a series, etc. If the series converges absolutely, all these will converge absolutely to the same sum. The process can be taken to any finite level, or to an infinite or even transfinite level. Details are left to the interested reader.

Drastic Revisions of Unordered Sums

IV.2.10.19. Suppose $\sum_{j\in J} a_j$ is an unordered summation, and we have a decomposition $J = \cup_{k\in K} J_k$ of J as a disjoint union of subsets. For each k we may form the unordered sum

$$\sum_{j\in J_k} a_j$$

and, if these sums all converge, we can form the unordered sum

$$\sum_{k\in K} \left(\sum_{j\in J_k} a_j \right) .$$

Such a sum is called a *drastic revision* of the unordered sum $\sum_{j\in J} a_j$.

IV.2.10.20. THEOREM. Let $\sum_{j\in J} a_j$ be a convergent unordered sum, and $J = \cup_{k\in K} J_k$ a decomposition of J. Then

(i) The unordered sum $\sum_{j\in J_k} a_j$ converges for each k.

(ii) The unordered sum (drastic revision)

$$\sum_{k\in K}\left(\sum_{j\in J_k} a_j\right) \tag{IV.1}$$

converges, and

$$\sum_{k\in K}\left(\sum_{j\in J_k} a_j\right) = \sum_{j\in J} a_j \ .$$

PROOF: We have $a_j \neq 0$ for only countably many J (IV.2.9.6.), and there are only countably many k for which J_k contains a nonzero term. Thus there is no loss of generality in assuming J and K are countable.

Fix bijections from $\mathbb{N}$ to K and to J_k for each k (the Countable AC is needed to do this; zero terms may have to be added to the unordered sum if some J_k are finite). The unordered sums thus become ordinary infinite series using these orderings, and the infinite series corresponding to IV.1 is a drastic revision of the original sum as an infinite series. The result then follows from IV.2.10.16. and IV.2.9.11.. 👍

IV.2.10.21. The converse of this theorem is true if all the terms of the unordered sum are nonnegative. But the converse is not true in general: if a drastic revision of an unordered sum converges, the original unordered sum need not converge, and if two drastic revisions of an unordered sum both converge, the sums may not be equal. Example IV.2.10.6. gives a counterexample if the series are regarded as unordered sums. The situation is essentially the same as the problem described in IV.2.10.10.. ✋

IV.2.10.22. Just as with ordinary series, we can take such drastic revisions to multiple levels. If the original unordered sum converges, so will all multiple-level drastic revisions, with the same sum.

IV.2.11. Products of Infinite Series

IV.2.11.1. Suppose we have two convergent infinite series

$$\sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k$$

with sums s and t respectively (it is notationally convenient to start the sums at 0 instead of 1). What should we mean by the product of the two series

$$\left(\sum_{k=0}^{\infty} a_k\right)\left(\sum_{k=0}^{\infty} b_k\right)$$

and what should its sum be? However we define the product its "sum" should be st.

IV.2.11.2. A natural approach is to consider the double array (c_{jk}), where $c_{jk} = a_j b_k$; this array would be obtained by formally multiplying out the series

$$(a_0 + a_1 + a_2 + \cdots)(b_0 + b_1 + b_2 + \cdots)$$

as though they were finite sums. Thus any of the sums considered in the last section would be a potential interpretation of the product of the series. Several of these automatically converge:

IV.2.11.3. PROPOSITION. Using the above notation, the following sums converge with sum st:

(i) The iterated sum $\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a_j b_k$.

(ii) The iterated sum $\sum_{k=0}^{\infty}\sum_{j=0}^{\infty} a_j b_k$.

(iii) The rectangular sum $\lim_{m,n\to\infty}\sum_{j=0}^{m}\sum_{k=0}^{n} a_j b_k$.

(iv) The square sum $\lim_{n\to\infty}\sum_{j=0}^{n}\sum_{k=0}^{n} a_j b_k$.

PROOF: (i) and (ii) each follow directly from two applications of IV.2.1.18.. For (iii) and (iv), set

$$s_m = \sum_{j=0}^{m} a_j \quad \text{and} \quad t_n = \sum_{k=0}^{n} b_k$$

and note that

$$\sum_{j=0}^{m}\sum_{k=0}^{n} a_j b_k = s_m t_n$$

and that

$$\lim_{m,n\to\infty} s_m t_n = \lim_{n\to\infty} s_n t_n = st \ .$$



IV.2.11.4. However, the triangular sum is more problematic. And it is the triangular sum which is the most natural definition of the product of the two series. For one thing, it is the only systematically obtained sum of the types previously considered which is actually a single infinite series. More importantly, it is especially natural when power series are considered (cf. III.5.3.2., V.16.3.11.): if

$$\sum_{k=0}^{\infty} a_k x^k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k x^k$$

are power series, if we collect together like powers of x in the array $c_{jk} = a_j b_k x^{j+k}$ we get the triangular sum

$$\sum_{n=0}^{\infty}\sum_{j+k=n} a_j b_k x^n = \sum_{n=0}^{\infty}\left[\left(\sum_{k=0}^{n} a_k b_{n-k}\right) x^n\right] \ .$$

We thus make the definition:

IV.2.11.5. DEFINITION. Let

$$\sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k$$

be infinite series. Then the *product*, or *Cauchy product*, of the series is the triangular sum

$$\left(\sum_{k=0}^{\infty} a_k\right)\left(\sum_{k=0}^{\infty} b_k\right) = \sum_{n=0}^{\infty} \sum_{j+k=n} a_j b_k = \sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k} \ .$$

IV.2.11.6. EXAMPLE. The product of two convergent infinite series can diverge. Set

$$a_k = b_k = \frac{(-1)^k}{\sqrt{k}}$$

for all $k \geq 1$. Then

$$\sum_{k=1}^{\infty} a_k \quad \text{and} \quad \sum_{k=1}^{\infty} b_k$$

converge by the Alternating Series Test. But the product series

$$\sum_{n=2}^{\infty} (-1)^n \left(\sum_{k=1}^{n-1} \frac{1}{\sqrt{k(n-k)}}\right)$$

diverges since the terms do not approach 0: $\frac{1}{\sqrt{k(n-k)}} \geq \frac{2}{n}$ for $1 \leq k \leq n-1$.

IV.2.11.7. If the series

$$\sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k$$

both converge absolutely to s and t respectively, then the iterated sum

$$\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} |a_j b_k|$$

converges, and thus by IV.2.10.11. all the associated sums, and in particular the triangular sum, converge absolutely to the same sum, which must be st. Thus we obtain:

IV.2.11.8. THEOREM. Let

$$\sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k$$

be absolutely convergent infinite series with sums s and t respectively. Then the product series

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k}$$

converges absolutely, with sum st.

While the absolutely convergent case is usually the most important one, the convergence result can be extended to the case where only one of the series converges absolutely. Note, however, that absolute convergence of the product series is not guaranteed in this case. This theorem is due to MERTENS.

IV.2.11.9. THEOREM. Let

$$\sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k$$

be convergent infinite series with sums s and t respectively. Suppose at least one of the series converges absolutely. Then the product series

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k}$$

converges, with sum st.

Actually, convergence of the product series is the only issue: if the product series converges, its sum is automatically the right thing. This result is due to ABEL.

IV.2.11.10. THEOREM. Let

$$\sum_{k=0}^{\infty} a_k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k$$

be convergent infinite series with sums s and t respectively. If the product series

$$\sum_{n=0}^{\infty} \sum_{k=0}^{n} a_k b_{n-k}$$

converges, its sum is st.

IV.2.11.11. We have considered products of two series. But we may iterate to obtain products of any finite number of series, and the extension of IV.2.11.8. applies: if each of the series converges absolutely, so does the product series. We should note the following technicality, which is proved by a straightforward calculation left to the reader:

IV.2.11.12. PROPOSITION. The product of infinite series is associative and commutative: if

$$\sum_{k=0}^{\infty} a_k \ , \quad \sum_{k=0}^{\infty} b_k \ , \quad \sum_{k=0}^{\infty} c_k$$

are infinite series, then

$$\left[\left(\sum_{k=0}^{\infty} a_k\right)\left(\sum_{k=0}^{\infty} b_k\right)\right]\left(\sum_{k=0}^{\infty} c_k\right) = \left(\sum_{k=0}^{\infty} a_k\right)\left[\left(\sum_{k=0}^{\infty} b_k\right)\left(\sum_{k=0}^{\infty} c_k\right)\right]$$

$$\left(\sum_{k=0}^{\infty} a_k\right)\left(\sum_{k=0}^{\infty} b_k\right) = \left(\sum_{k=0}^{\infty} b_k\right)\left(\sum_{k=0}^{\infty} a_k\right) \ .$$

IV.2.12. Infinite Products

IV.2.12.1. An *infinite product* is an expression of the form

$$\prod_{k=1}^{\infty} u_k$$

where the u_k are numbers (or later functions). The u_k are called the *factors* of the product.

Infinite products have a long history, and there are some famous examples such as the following expressions involving π due respectively to Viéte (1646) and Wallis (1653):

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \cdot \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} \cdot \ \cdots$$

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \ \cdots \ \cdot \frac{2m}{2m+1} \cdot \frac{2m+2}{2m+1} \cdot \ \cdots$$

In this section, we will make sense out of such expressions and discuss convergence.

IV.2.12.2. In analogy with the definition for infinite series, we could define *partial products*

$$p_n = \prod_{k=1}^{n} u_k$$

and say that the infinite product converges, with product p, if the sequence (p_n) converges to p. But there are two serious difficulties with this approach in general. First, if one of the u_k is zero, then all the partial products from that point on will be zero no matter what the rest of the factors are. An secondly, even if all the u_k are nonzero, so all the p_k are nonzero, we can easily have $p = 0$. In either of these cases it is not reasonable to say that the infinite product converges. Thus we make the following definition:

IV.2.12.3. Definition. Let $\prod_{k=1}^{\infty} u_k$ be an infinite product. Then the product *converges*, with product p, if $p \neq 0$ and the sequence (p_n) of partial products converges to p. If there is no such p, the infinite product *diverges*.

The factors of a convergent infinite product are necessarily nonzero. We can extend the definition by allowing a finite number of zero factors and determining convergence by the tail products after the last zero factor (this is useful when the factors are functions, to allow the product to be defined when one factor is zero).

The next fundamental observation is analogous to the Divergence Criterion for series:

IV.2.12.4. PROPOSITION. Let $\prod_{k=1}^{\infty} u_k$ be a convergent infinite product. Then $\lim_{k\to\infty} u_k = 1$.

PROOF: We have $\lim_{k\to\infty} p_k = p \neq 0$ and $\lim_{k\to\infty} p_{k+1} = p$, and hence by ()

$$\lim_{k\to\infty} u_k = \lim_{k\to\infty} \frac{p_{k+1}}{p_k} = \frac{\lim_{k\to\infty} p_{k+1}}{\lim_{k\to\infty} p_k} = \frac{p}{p} = 1 \ .$$



Note that $\lim_{k\to\infty} u_k = 1$ is a *necessary* condition for convergence of the infinite product, but not a *sufficient* one, just as with infinite series. 

IV.2.12.5. Two important consequences of IV.2.12.4. are:

(i) Since the factors of a convergent infinite product are eventually close to 1, they are eventually positive, i.e. a convergent infinite product can have only finitely many negative factors. For convenience, we will usually only consider infinite products with positive factors; ones with finitely many negative (or zero) factors can be reduced to this case by considering a tail product (in obvious analogy with IV.2.1.6., an infinite product converges if and only if some (every) tail product converges).

(ii) In light of IV.2.12.4., it is customary to write the factor u_k as $(1 + a_k)$ and to call the a_k the *terms* of the infinite product. IV.2.12.4. then says that the terms of a convergent infinite product tend to zero. We will eventually see that the infinite product $\prod_{k=1}^{\infty}(1 + a_k)$ converges roughly (but not exactly) if and only if $\sum_{k=1}^{\infty} a_k$ converges.

As a strengthening of (i), for convenience we often only consider infinite products with $|a_k| < 1$ for all k; this will be true in some tail of any convergent infinite product.

The next result in a sense completely reduces the convergence question for infinite products to a question about infinite series:

IV.2.12.6. PROPOSITION. Let $\prod_{k=1}^{\infty} u_k$ be an infinite product with $u_k > 0$ for all k. Then the infinite product converges if and only if the infinite series $\sum_{k=1}^{\infty} \log u_k$ converges.

PROOF: If we have $p_n \to p > 0$, we have

$$\log p_n = \sum_{k=1}^{n} \log u_k \to \log p$$

by continuity of the logarithm function, so $\sum_{k=1}^{\infty} \log u_k$ converges. Conversely, if $s_n = \sum_{k=1}^{n} \log u_k$ converges to s, we have $p_n = e^{s_n} \to e^s$ by continuity of the exponential function, so the infinite product converges. 👍

IV.2.12.7. Thus, if we want to analyze the convergence of the infinite product $\prod_{k=1}^{\infty}(1 + a_k)$ with all $|a_k| < 1$, we can just consider convergence of the infinite series $\sum_{k=1}^{\infty} \log(1 + a_k)$. But it would be preferable to have a criterion in terms of the closely related but simpler series $\sum_{k=1}^{\infty} a_k$. We can do this using the following lemma:

IV.2.12.8. LEMMA. There is a $\delta > 0$ such that

$$x - x^2 < \log(1 + x) < x - \frac{1}{3}x^2$$

whenever $|x| < \delta$.

PROOF: It suffices to show that

$$-1 < \lim_{x \to 0} \frac{\log(1 + x) - x}{x^2} < -\frac{1}{3} .$$

But this limit is $-\frac{1}{2}$ by l'Hôpital's Rule. 👍

IV.2.12.9. PROPOSITION. Let (a_n) be a sequence of numbers with $|a_n| < 1$ for all n. If $\sum_{k=1}^{\infty} a_k^2$ converges absolutely, then $\sum_{k=1}^{\infty} a_k$ converges if and only if $\sum_{k=1}^{\infty} \log(1 + a_k)$ converges.

PROOF: We have $a_k \to 0$. For each k, write $\log(1+a_k) = a_k + \theta_k a_k^2$. Then (θ_k) is a bounded sequence since $-1 < \theta_k < -\frac{1}{3}$ for sufficiently large k by IV.2.12.8. (actually $\theta_k \to -\frac{1}{2}$). Since $\sum_{k=1}^{\infty} a_k^2$ converges absolutely, we have that $\sum_{k=1}^{\infty} \theta_k a_k^2$ converges (absolutely) by IV.2.7.2.. Hence $\sum_{k=1}^{\infty} \log(1+a_k)$ and $\sum_{k=1}^{\infty} a_k$ either both converge or both diverge by IV.2.1.16.. 👍

IV.2.12.10. COROLLARY. Let $\prod_{k=1}^{\infty}(1+a_k)$ be an infinite product. If $\sum_{k=1}^{\infty} a_k^2$ converges absolutely, then the infinite product converges if and only if $\sum_{k=1}^{\infty} a_k$ converges.

IV.2.12.11. If the a_k are real numbers, then $\sum_{k=1}^{\infty} a_k^2$ is a nonnegative series, so convergence and absolute convergence are the same thing. But these results also hold for complex series and products, where absolute convergence of $\sum_{k=1}^{\infty} a_k^2$ does not necessarily follow from convergence.

IV.2.12.12. The convergence of $\sum_{k=1}^{\infty} a_k^2$ is not necessary for the convergence of the infinite product, however; see the example in Exercise IV.2.16.17.. See also Exercise IV.2.16.18. for an example of a divergent infinite product where $\sum_{k=1}^{\infty} a_k$ converges. ✋

However, if all terms have the same sign, convergence of $\sum_{k=1}^{\infty} a_k$ is both necessary and sufficient for convergence of the infinite product:

IV.2.12.13. THEOREM. Let $\prod_{k=1}^{\infty}(1+a_k)$ be an infinite product. If all a_k have the same sign, then the infinite product converges if and only if $\sum_{k=1}^{\infty} a_k$ converges.

PROOF: One direction follows from IV.2.12.10.: if $\sum_{k=1}^{\infty} a_k$ converges, it converges absolutely, and hence $\sum_{k=1}^{\infty} a_k^2$ converges (absolutely) by the Comparison Test. The converse does not follow from IV.2.12.10., however.

To prove the converse, suppose first that all $a_k > 0$. If $s_n = \sum_{k=1}^{n} a_k$, then we have

$$p_n = (1+a_1)(1+a_2)\cdots(1+a_n) = 1 + a_1 + a_2 + \cdots + a_n + (\text{higher-order terms}) \geq s_n$$

since the higher-order terms are all nonnegative. If the infinite product converges, the p_n are bounded, so the s_n are bounded and $\sum_{k=1}^{\infty} a_k$ converges.

Now assume the a_k are all negative. By passing to a tail, we may assume $|a_k| < 1$ for all k. Assume the infinite product converges with product p; then $p_n \geq p$ for all n. We have

$$1 - x \leq \frac{1}{1+x}$$

for any $x > -1$, since $(1+x)(1-x) = 1 - x^2$. Thus we have

$$1 - a_k \leq \frac{1}{1+a_k}$$

for all k, and thus

$$\prod_{k=1}^{n}(1-a_k) \leq \frac{1}{p_n} \leq \frac{1}{p}\ .$$

Thus the infinite product $\prod_{k=1}^{\infty}(1-a_k)$ converges, and hence $\sum_{k=1}^{\infty}(-a_k)$ converges by the previous argument. 👍

Absolutely Convergent Infinite Products

IV.2.12.14. DEFINITION. An infinite product $\prod_{k=1}^{\infty}(1+a_k)$ *converges absolutely* if $\prod_{k=1}^{\infty}(1+|a_k|)$ converges.

Just as in the sum case (IV.2.5.2.), we have:

IV.2.12.15. PROPOSITION. An absolutely convergent infinite product converges.

The proof is left to the reader (Exercise IV.2.16.19.).

The next result is an immediate corollary of IV.2.12.13., and along with the previous examples shows that there are many convergent infinite products which are not absolutely convergent.

IV.2.12.16. COROLLARY. An infinite product $\prod_{k=1}^{\infty}(1+a_k)$ converges absolutely if and only if $\sum_{k=1}^{\infty} a_k$ converges absolutely.

Rearrangement

Just as with infinite series, we may rearrange the order of factors in an absolutely convergent infinite product without affecting convergence or changing the product; however, rearrangement of order in a non-absolutely convergent infinite product can affect convergence and/or change the product.

IV.2.13. Infinite Series of Complex Numbers

Infinite series of complex numbers behave almost identically to series of real numbers. Almost all results about real infinite series carry over to the complex case, with identical proofs. We state the essentials here.

IV.2.13.1. DEFINITION. The infinite series $\sum_{k=1}^{\infty} a_k$ of complex numbers *converges*, with *sum* $s \in \mathbb{R}$, if the sequence (s_n) of partial sums converges to s in $\mathbb{C}$.

The infinite series $\sum_{k=1}^{\infty} a_k$ of complex numbers *converges absolutely* if $\sum_{k=1}^{\infty} |a_k|$ converges. The series *converges conditionally* if it converges but does not converge absolutely.

The next observation allows many aspects of the theory of complex infinite series to be reduced to the real case.

IV.2.13.2. PROPOSITION. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series of complex numbers. For each k, write $a_k = b_k + ic_k$ with $b_k, c_k \in \mathbb{R}$. Then $\sum_{k=1}^{\infty} a_k$ converges if and only if both $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ converge, and

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} b_k + i \sum_{k=1}^{\infty} c_k \ .$$

The series $\sum_{k=1}^{\infty} a_k$ converges absolutely if and only if both $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ converge absolutely.

PROOF: Let $r_n = \sum_{k=1}^{n} a_k$, $s_n = \sum_{k=1}^{n} b_k$, $t_n = \sum_{k=1}^{n} c_k$ be the partial sums. Then $r_n = s_n + it_n$ for each n, so the first result follows from (). The second result follows immediately from the Comparison Test and IV.2.1.16. since, for every k,

$$|b_k|, |c_k| \leq |a_k| \leq |b_k| + |c_k| \ .$$



IV.2.13.3. PROPOSITION. Every absolutely convergent series of complex numbers converges.

The proof is identical to the proof of IV.2.5.2.. Alternately, the result follows immediately from IV.2.13.2..

IV.2.13.4. Since $\sum_{k=1}^{\infty} |a_k|$ is an infinite series with nonnegative real terms, all the tests for convergence of nonnegative real series are tests for absolute convergence of complex series as well.

IV.2.13.5. The theorems about termwise addition and scalar multiplication (IV.2.1.16., IV.2.1.18.), including multiplication by complex scalars, hold also for complex series. The theorems about rearrangements and subarrangements (IV.2.8.3., IV.2.8.5.), on unordered and iterated summations, and the theorem on products of series (IV.2.11.8.), hold for absolutely convergent complex series. All the proofs are the same as in the real case, and/or the results follow immediately from the real case using IV.2.13.2.. (Riemann's Theorem on rearranging conditionally convergent series (IV.2.8.7.) is more subtle, however: see Exercise ().)

IV.2.13.6. Similar statements hold about infinite products of complex numbers. In particular, IV.2.12.10. and IV.2.12.16. hold for infinite products of complex numbers.

IV.2.14. Numerical Evaluation and Estimation of Series

IV.2.15. Summability Methods for Divergent Series

> "An infinite series does not have a sum by nature; rather, it acquires a sum only if we associate one to it by a procedure."
>
> *F. Waismann*[4]

IV.2.15.1. While we may quibble about whether a definite "procedure" is necessary, it is true that an infinite series does not have an intrinsic well-defined sum (this is the Fundamental Principle) but that the "sum" must be defined. There are in fact several more-or-less reasonable ways this can be done; we used what by general agreement is the most reasonable way to define convergent series and associated the limit of the partial sums, but this is not the only possibility.

We could in principle define the sum of an infinite series arbitrarily. But for the definition to be reasonable and useful, the "sum" so defined must have at least some of the properties of finite sums, e.g. analogs of IV.2.1.16. and IV.2.1.18. (linearity) and order compatibility (for example, the "sum" of a nonzero nonnegative series should be positive), and should agree with the algebraic sum for series with only finitely many nonzero terms. Additional properties such as grouping and contraction (cf. IV.2.1.21.) and rearrangement (IV.2.8.5.) would be desirable, but might be hard to obtain (even rearrangement does not hold for all convergent series). Various arguments can be given () that any reasonable definition of sum for infinite series must necessarily assign to any convergent, or at least absolutely convergent, series its sum in the usual sense.

IV.2.15.2. Divergent series are another matter. There are a number of quite reasonable ways of giving a definition of "sum" for certain divergent series, all of which extend the usual definition for convergent series (i.e. assign to each convergent series its usual sum) but which do not always agree even for all the divergent series for which they are defined.

[4] [?, p. 139].

IV.2.16. Exercises

IV.2.16.1. Prove that

$$\sum_{k=1}^{\infty} \frac{k}{2^k} = 2 \ .$$

This was one of the first infinite series to be summed, by Richard Suiseth in the fourteenth century.

IV.2.16.2. [?] Let $\sum_{k=1}^{\infty} a_k$ be a convergent infinite series with strictly positive terms, and for each n let $t_n = \sum_{k=n}^{\infty} a_k$ be the n'th tail sum. Set $b_k = \sqrt{t_k} - \sqrt{t_{k+1}}$ for each k. Show that $\sum_{k=1}^{\infty} b_k$ also converges, and that $\lim_{k\to\infty} \frac{b_k}{a_k} = +\infty$ (note that $a_k = t_k - t_{k+1}$), i.e. (b_k) converges to zero "more slowly" than (a_k).

IV.2.16.3. Let $\sum_{k=1}^{\infty} a_k$ be a divergent infinite series with strictly positive terms, and for each n let $s_n = \sum_{k=1}^{n} a_k$ be the n'th partial sum. Set $b_k = \frac{a_k}{\sqrt{s_k}}$ for each k. Show that $\sum_{k=1}^{\infty} b_k$ also diverges (its n'th partial sum is at least $\sqrt{s_n}$), and that $\lim_{k\to\infty} \frac{b_k}{a_k} = 0$, i.e. (b_k) converges to zero "more rapidly" than (a_k).

IV.2.16.4. Let $\sum_{k=1}^{\infty} a_k^{(1)}$, $\sum_{k=1}^{\infty} a_k^{(2)}$, ... be a sequence of convergent positive infinite series.

(a) Inductively find for each m a convergent positive series $\sum_{k=1}^{\infty} b_k^{(m)}$ such that $\lim_{k\to\infty} \frac{a_k^{(j)}}{b_k^{(m)}} = 0$ for $1 \le j \le m$ and $\lim_{k\to\infty} \frac{b_k^{(j)}}{b_k^{(m)}} = 0$ for $j < m$. [Apply the argument of Exercise IV.2.16.2. to the series $\sum_{k=1}^{\infty} (b_k^{(m-1)} + a_k^{(m)})$.]

(b) Construct a convergent positive infinite series $\sum_{k=1}^{\infty} b_k$ such that $\lim_{k\to\infty} \frac{a_k^{(m)}}{b_k} = 0$ for all m, i.e. such that the b_k go to zero "more slowly" than $a_k^{(m)}$ for every m. [Let $b_k = b_k^{(m)}$ for $n_{m-1} + 1 \le k \le n_m$, for a suitably chosen increasing sequence (n_m).]

(c) If each $\sum_{k=1}^{\infty} a_k^{(m)}$ diverges, similarly construct a divergent positive infinite series $\sum_{k=1}^{\infty} b_k$ such that $\lim_{k\to\infty} \frac{b_k}{a_k^{(m)}} = 0$ for all m, i.e. such that the b_k go to zero "more rapidly" than $a_k^{(m)}$ for every m.

IV.2.16.5. Choose an increasing sequence (k_n) such that $\sum_{k=k_n+1}^{\infty} \frac{1}{k^2} < 4^{-n}$, and for $k_n < k \le k_{n+1}$ let $a_k = \frac{2^n}{k^2}$. Set $a_k = \frac{1}{k^2}$ for $k \le k_1$.

(a) Show that $\limsup_{k\to\infty} \frac{a_{k+1}}{a_k} = 2$ and $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} = 1$.

(b) Show that $\sum_{k=1}^{\infty} a_k$ converges.

(c) Construct a similar convergent positive series $\sum_{k=1}^{\infty} b_k$ with $\limsup_{k\to\infty} \frac{b_{k+1}}{b_k} = +\infty$ and $\liminf_{k\to\infty} \frac{b_{k+1}}{b_k} = 1$.

IV.2.16.6. Choose an increasing sequence (k_n) such that $\sum_{k=1}^{k_n} \frac{1}{k} > 4^n$, and for $k_n < k \le k_{n+1}$ let $a_k = \frac{2^{-n}}{k}$. For $k \le k_1$ set $a_k = \frac{1}{k}$.

(a) Show that $\limsup_{k\to\infty} \frac{a_{k+1}}{a_k} = 1$ and $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} = \frac{1}{2}$.

(b) Show that $\sum_{k=1}^{\infty} a_k$ diverges.

(c) Construct a similar divergent positive series $\sum_{k=1}^{\infty} b_k$ with $\limsup_{k\to\infty} \frac{b_{k+1}}{b_k} = 1$ and $\liminf_{k\to\infty} \frac{b_{k+1}}{b_k} = 0$.

IV.2.16.7. Let $p > 0$ be fixed, and set

$$a_k = \begin{cases} \frac{1}{k^p} & \text{if} \quad k \text{ odd} \\ \\ \frac{2}{k^p} & \text{if} \quad k \text{ even} \end{cases} .$$

(a) Show that $\limsup_{k\to\infty} \frac{a_{k+1}}{a_k} = 2$ and $\liminf_{k\to\infty} \frac{a_{k+1}}{a_k} = \frac{1}{2}$.

(b) Show that $\sum_{k=1}^{\infty} a_k$ converges if $p > 1$ and diverges if $p \le 1$.

IV.2.16.8. [**?**] Fix an r, $0 < r < 1$, and consider the infinite series

$$\sum_{k=1}^{\infty} r^k |\sin k|$$

with strictly positive terms a_k.

(a) Show that the Root Test applies to yield that the series converges (this can be concluded more easily by simple comparison with the geometric series $\sum r^k$).

(b) Show that the numbers $\frac{a_{k+1}}{a_k}$ are dense in $[0, +\infty)$, and thus the Ratio Test fails spectacularly.

IV.2.16.9. (a) Prove the following extension of the Root Test (due to Cauchy):

Theorem. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms.

(i) If there is an $r > 1$ such that

$$\frac{\log\left(\frac{1}{a_k}\right)}{\log k} \geq r$$

for all sufficiently large k, then the series converges.
(ii) If

$$\frac{\log\left(\frac{1}{a_k}\right)}{\log k} \leq 1$$

for all sufficiently large k (or if $a_k \not\to 0$), then the series diverges.

COROLLARY. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Suppose

$$\lim_{k\to\infty}\left[\frac{\log\left(\frac{1}{a_k}\right)}{\log k}\right]$$

exists and equals L. Then
(i) If $L > 1$, the series converges.
(ii) If $L < 1$, the series diverges.
(iii) If $L = 1$, the series may or may not converge, and no conclusion may be drawn (the test fails).

(b) Show that the Theorem and the Corollary generalize the Basic Root Test and Ordinary Root Test respectively.
(c) Show that this test does not fail for p-series, and thus is strictly stronger than the Root Test.
(d) Prove even more delicate versions by replacing $\frac{\log\left(\frac{1}{a_k}\right)}{\log k}$ by

$$\frac{\log\left(\frac{1}{ka_k}\right)}{\log\log k}$$

and then by

$$\frac{\log\left(\frac{1}{ka_k\log k}\right)}{\log\log\log k}$$

etc. Apply these to the series

$$\sum_{k=2}^{\infty}\frac{1}{k(\log k)^p}\ ,\quad \sum_{k=3}^{\infty}\frac{1}{k\log k(\log\log k)^p}\ ,\ \cdots$$

which converge for $p > 1$ and diverge for $p \leq 1$ (the Integral Test also works on these series).

IV.2.16.10. (a) Prove the following extension of the Ratio Test:

THEOREM. Let $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} b_k$ be infinite series with strictly positive terms.

(i) If

$$\frac{a_{k+1}}{a_k} \leq \frac{b_{k+1}}{b_k}$$

for all sufficiently large k, and if $\sum_{k=1}^{\infty} b_k$ converges, then $\sum_{k=1}^{\infty} a_k$ converges.

(ii) If

$$\frac{a_{k+1}}{a_k} \geq \frac{b_{k+1}}{b_k}$$

for all sufficiently large k, and if $\sum_{k=1}^{\infty} b_k$ diverges, then $\sum_{k=1}^{\infty} a_k$ diverges.

(b) Show that the Ratio Test is the special case where $\sum_{k=1}^{\infty} b_k$ is a geometric series.

(c) Show that if $\sum_{k=1}^{\infty} a_k$ is an infinite series with strictly positive terms, and

$$\frac{a_{k+1}}{a_k} \geq \frac{k}{k+1}$$

for all sufficiently large k, then the series diverges.

IV.2.16.11. [**?**] Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with strictly positive terms. Let s_n be the n'th partial sum, and if the series converges, let s be the sum and

$$t_n = s - s_n = \sum_{k=n+1}^{\infty} a_k$$

the n'th tail sum.

(a) Show that if the series converges, then the sequence $p_k = \frac{t_k}{a_k}$ works in part (i) of Kummer's Test with $r = 1$.

(b) Show that if the series diverges, then the sequence $p_k = \frac{s_k}{a_k}$ works in part (ii) of Kummer's Test. [To show that $\sum_{k=1}^{\infty} \frac{1}{p_k}$ diverges, for each n choose m_n such that

$$\sum_{k=n+1}^{m_n} a_k \geq \sum_{k=1}^{n} a_k \ ;$$

show that $\sum_{k=n+1}^{m_n} \frac{1}{p_k} \geq \frac{1}{2}$ for all n and apply the Cauchy Criterion.]

IV.2.16.12. Let (x_k) be a sequence of positive numbers. Show that

$$\liminf_{k\to\infty}\left[k\left(x_k-1\right)\right]=\liminf_{k\to\infty}\left[k\left(1-\frac{1}{x_k}\right)\right] \quad \text{and} \quad \limsup_{k\to\infty}\left[k\left(x_k-1\right)\right]=\limsup_{k\to\infty}\left[k\left(1-\frac{1}{x_k}\right)\right]$$

as follows:

(a) Suppose $L=\liminf_{k\to\infty}\left[k\left(x_k-1\right)\right]\neq-\infty$, and let $r\in\mathbb{R}$, $r<L$. Then $k(x_k-1)>r$ for all sufficiently large k. Show that

$$\liminf_{k\to\infty}\left[k\left(1-\frac{1}{x_k}\right)\right]\geq\liminf_{k\to\infty}\frac{kr}{k+r}=r\ .$$

(b) Suppose $L'=\liminf_{k\to\infty}\left[k\left(1-\frac{1}{x_k}\right)\right]\neq-\infty$, and let $r\in\mathbb{R}$, $r<L'$. Then $k\left(1-\frac{1}{x_k}\right)>r$ for all sufficiently large k. Show that

$$\liminf_{k\to\infty}\left[k\left(x_k-1\right)\right]\geq\liminf_{k\to\infty}\frac{kr}{k-r}=r\ .$$

(c) Prove the lim inf equality in the case where one side is $-\infty$.

(d) Replace x_k by $\frac{1}{x_k}$ in the lim inf equality to obtain the lim sup equality.

IV.2.16.13. (a) Use Exercise IV.2.16.12. to prove the version IV.2.4.5. of Raabe's Test from IV.2.4.4., and the converse implication.

(b) Similarly prove the equivalence of IV.2.4.8. and IV.2.4.9..

IV.2.16.14. (a) Show that Raabe's Test can be equivalently written as follows:

THEOREM. Let $\sum_{k=1}^{\infty}a_k$ be an infinite series with strictly positive terms. Write

$$\frac{a_k}{a_{k+1}}=\frac{1}{1+\alpha_k}$$

for each k. Then

(i) If there is an $r>1$ such that $k\alpha_k\geq r$ for all sufficiently large k, then the series converges.

(ii) If $k\alpha_k\leq 1$ for all sufficiently large k, the series diverges.

COROLLARY. Let $\sum_{k=1}^{\infty}a_k$ be an infinite series with strictly positive terms. Write

$$\frac{a_k}{a_{k+1}}=\frac{1}{1+\alpha_k}$$

for each k. Suppose $\lim_{k\to\infty}k\alpha_k$ exists and equals L. Then

(i) If $L>1$, the series converges.

(ii) If $L<1$, the series diverges.

(iii) If $L = 1$ but $k\alpha_k > 1$ for infinitely many k, the series may or may not converge (the test fails).

(b) Deduce the Corollary from the Theorem.
(c) Give an alternate proof of part (ii) of the Theorem using IV.2.16.10. with the series $b_k = \frac{1}{k}$.
(d) Show using the Binomial Theorem () that if $1 < p < r$, then

$$\frac{1}{1+\frac{r}{k}} < \left[\frac{k}{k+1}\right]^p$$

for all sufficiently large k.
(e) Give an alternate proof of the theorem using (d) and IV.2.16.10. with $b_k = \frac{1}{k^p}$.

IV.2.16.15. Let $\sum_{k=1}^{\infty} a_k$ be an infinite series with nonincreasing nonnegative terms. Use the Cauchy Condensation Test to show that $\sum_{k=1}^{\infty} a_k$ converges if and only if $\sum_{k=1}^{\infty} (2k+1)a_{k^2}$ converges. Then show by the Limit Comparison Test that this series converges if and only if $\sum_{k=1}^{\infty} k a_{k^2}$ converges.

IV.2.16.16. This problem gives an alternate proof that an absolutely convergent series of real numbers converges (IV.2.5.2.).
Let $\sum_{k=1}^{\infty} a_k$ be an absolutely convergent infinite series. For each k set

$$b_k = \frac{1}{2}(|a_k| + a_k), \quad c_k = \frac{1}{2}(|a_k| - a_k) \ .$$

Note that $a_k = b_k - c_k$ and $|a_k| = b_k + c_k$.
(a) Show that $\sum_{k=1}^{\infty} b_k$ and $\sum_{k=1}^{\infty} c_k$ are nonnegative series which converge by the comparison test.
(b) Conclude from IV.2.1.16. and IV.2.1.18. that $\sum_{k=1}^{\infty} a_k$ also converges.
(c) Where was completeness of $\mathbb{R}$ used in this proof?
This proof is arguably simpler than the one in IV.2.5.2., but the one in IV.2.5.2. is more widely applicable since it works verbatim for infinite series in $\mathbb{C}$ and in Banach spaces ().

IV.2.16.17. For each k set $a_{2k-1} = \frac{1}{\sqrt{k}}$ and $a_{2k} = \frac{1}{k} - \frac{1}{\sqrt{k}}$.
(a) Show that $\sum_{k=1}^{\infty} a_k$ and $\sum_{k=1}^{\infty} a_k^2$ both diverge.
(b) Show that $\prod_{k=1}^{\infty} (1 + a_k)$ converges.

IV.2.16.18. For each k set $a_k = \frac{(-1)^k}{\sqrt{k}}$.

(a) Show that $\sum_{k=1}^{\infty} a_k$ converges and $\sum_{k=1}^{\infty} a_k^2$ diverges.

(b) Show that $\prod_{k=1}^{\infty}(1 + a_k)$ diverges.

IV.2.16.19. (a) Prove the following Cauchy Criterion for infinite products: An infinite product $\prod_{k=1}^{\infty} u_k$ converges if and only if, for every $\epsilon > 0$ there is an N such that

$$\left| \left[\prod_{k=n+1}^{m} u_k \right] - 1 \right| < \epsilon$$

whenever $N \leq n < m$.

(b) If $a_1, \ldots, a_n$ are any numbers, prove that

$$\left| \left[\prod_{k=1}^{n}(1 + a_k) \right] - 1 \right| \leq \left[\prod_{k=1}^{n}(1 + |a_k|) \right] - 1 \ .$$

(c) Prove that an absolutely convergent infinite product converges.

IV.2.16.20. (a) Let (c_k) be an unbounded sequence of positive numbers. Choose a strictly increasing sequence (k_j) such that $c_{k_j} \geq 2^j$. Set $a_{k_j} = 2^{-j}$, and $a_k = 2^{-k}$ for k not one of the k_j. Show that $\sum_{k=1}^{\infty} a_k$ converges and $\sum_{k=1}^{\infty} c_k a_k$ diverges.

(b) Let (d_k) be a sequence of positive numbers with $\inf_k d_k = 0$. Choose a strictly increasing sequence (k_j) such that $d_{k_j} \leq 2^{-j}$. Set $a_{k_j} = 1$; for k not one of the k_j set $a_k = 2^{-k}$ if $d_k \leq 1$ and $a_k = \frac{1}{2^k d_k}$ if $d_k > 1$. Show that $\sum_{k=1}^{\infty} a_k$ diverges but $\sum_{k=1}^{\infty} d_k a_k$ converges.

IV.2.16.21. (a) Let $\sum_{k=1}^{\infty} a_k$ be a divergent nonnegative infinite series, with partial sums $s_n > 0$. Show that $\sum_{k=1}^{\infty} \frac{a_k}{s_k}$ also diverges. [Show that if $n < m$, then

$$\sum_{k=n+1}^{m} \frac{a_k}{s_k} \geq 1 - \frac{s_n}{s_m}$$

and use that $s_n \to +\infty$.]

(b) Show that $\displaystyle\sum_{k=1}^{\infty} \frac{a_k}{s_k^p}$ converges if $p > 1$. (The case $p = 1/2$ is Exercise IV.2.16.3..)

This result was proved independently by Abel and Dini.

IV.2.16.22. (a) Let $\displaystyle\sum_{k=1}^{\infty} a_k$ be a convergent nonnegative infinite series, with tail sums $t_n > 0$. Show that $\displaystyle\sum_{k=1}^{\infty} \frac{a_k}{t_k}$ diverges.

(b) Show that $\displaystyle\sum_{k=1}^{\infty} \frac{a_k}{t_k^p}$ converges if $p < 1$.

This result was proved by Dini.

IV.2.16.23. (Euler) Let p_k be the k'th prime number (in increasing order). The goal of this problem is to prove that $\sum_k \frac{1}{p_k}$ diverges. (This provides another proof that there are infinitely many primes.)

(a) For each n, let m_n be the integer part of $\log_2 n$, i.e. 2^{m_n} is the largest power of 2 which is $\le n$. Show that, if $r \le n$ and $r = p_1^{e_1} p_2^{e_2} \cdots p_s^{e_s}$, then $e_j \le m_n$ for all j.

(b) Let $n \in \mathbb{N}$, and let r be the largest index for which $p_r \le n$. Show that

$$\left(\sum_{j=0}^{m_n} \frac{1}{p_1^j}\right)\left(\sum_{j=0}^{m_n} \frac{1}{p_2^j}\right)\cdots\left(\sum_{j=0}^{m_n} \frac{1}{p_r^j}\right) \ge \sum_{k=1}^{n} \frac{1}{k}\ .$$

[Multiply out the left side.]

(c) Show that the left side of the inequality in (b) is dominated by

$$\prod_{k=1}^{r} \frac{p_k}{p_k - 1} = \prod_{k=1}^{r} \frac{1}{1 - \frac{1}{p_k}}\ .$$

[The factors are partial sums of geometric series.]

(d) Conclude that there are infinitely many p_k, and that the infinite product

$$\prod_{k=1}^{\infty} \frac{p_k}{p_k - 1} = \prod_{k=1}^{\infty} \frac{1}{1 - \frac{1}{p_k}}$$

diverges, and hence that

$$\sum_{k=1}^{\infty} \frac{1}{p_k}$$

also diverges. [Use IV.2.12.16..]

IV.2.16.24. (a) Show that a revision of a revision of an infinite series $\sum_{k=1}^{\infty} a_k$ is a revision of $\sum_{k=1}^{\infty} a_k$.

(b) Say that two infinite series are *associates* if they are both revisions of the same infinite series. Find necessary and sufficient conditions that two given series are associates.

(c) Say that two absolutely convergent infinite series are *strong associates* if they are both revisions of the same absolutely convergent infinite series. Find necessary and sufficient conditions that two given absolutely convergent series are strong associates.

(d) Is the property of being associates [resp. strong associates] an equivalence relation on infinite series [resp. absolutely convergent infinite series]?

(e) Answer the same questions if "revisions" is replaced by "drastic revisions."

IV.2.16.25. What is wrong with the following argument? Let $x \in \mathbb{R}$, $x \neq 1$. Then

$$\sum_{k=-\infty}^{\infty} x^k = \sum_{k=0}^{\infty} x^k + \sum_{k=1}^{\infty} x^{-k}$$

$$= \sum_{k=0}^{\infty} x^k + \sum_{k=1}^{\infty} \left(\frac{1}{x}\right)^k = \frac{1}{1-x} + \frac{\frac{1}{x}}{1-\frac{1}{x}} = \frac{1}{1-x} + \frac{1}{x-1} = 0 \ .$$

(Note that if $x > 0$, all the terms in the sum are positive.)

IV.2.16.26. Show that

$$\lim_{n\to\infty} \left[-\log n + \sum_{k=1}^{n} \frac{1}{k} \right]$$

exists and is finite and nonzero. This limit is the *Euler-Mascheroni constant* γ. This constant appears repeatedly throughout mathematics, particularly in connection with the Γ function (e.g. V.13.4.3.). The approximate value is

$$\gamma = 0.57721566490153286060651209008240243104215933593992\ldots$$

It is not known whether γ is rational or irrational, algebraic or transcendental (from known results it is transcendental with probability 1).

IV.3. Convergence of Sequences of Functions

If (x_n) is a sequence of numbers, the meaning of an expression like $x_n \to x$ is unambiguous: there is only one notion of convergence for sequences of numbers (except in rare instances where an exotic topology is being considered, which need not concern us here). Similarly, there is one standard and universal notion of convergence for sequences of points in Euclidean space.

The situation with functions is dramatically different. If (f_n) is a sequence of functions, there is no unambiguous meaning for an expression like $f_n \to f$. There are many possible meanings (some dependent on context); some of the most common ones are pointwise convergence, uniform convergence, u.c. convergence, a.e. convergence, convergence in measure, mean convergence, weak convergence, ... The reader should form the habit of clearly specifying, or at least clearly understanding, what kind of convergence is being considered in any situation, to avoid writing "meaningless" expressions like $f_n \to f$. The conceptual understanding that sequences of functions can converge in some senses but not others is one of the most important and basic mindsets an analyst must develop.

Some of the notions of convergence of sequences of functions are stronger (more restrictive) than others. Properties of functions (e.g. continuity) are sometimes automatically preserved under certain types of convergence, but not others; the stronger the notion of convergence, the more properties are preserved, and the larger the number of desirable consequences of the convergence. There are a number of useful results in analysis and topology which allow one to deduce a stronger type of convergence from a weaker one under certain circumstances.

In this section, we will discuss a few of the most important and widely-used notions of convergence of sequences of functions; other notions will appear in certain contexts in other sections. We will emphasize the setting of functions from $\mathbb{R}$ to $\mathbb{R}$, but point out how the notions can be extended to more general situations.

IV.3.1. Pointwise and Uniform Convergence

IV.3.1.1. More than once Cauchy asserted that a (pointwise) limit of a sequence (or sum of an infinite series) of continuous functions is continuous, even after Abel had observed in 1826 that the statement in Cauchy's 1821 text *Cours d'analyse* was false; Cauchy still claimed in 1833 that the sum of a pointwise convergent Fourier series is continuous, and even in 1853 when he in effect added uniform convergence to the statement in the *Cours d'analyse* (and to its proof), he pointedly declined to retract his previous assertions, representing the change as a "clarification" rather than a "correction."

A. Robinson [Rob66], I. Lakatos [Lak78], and D. Laugwitz [Lau87]–[Lau89] have tried to claim that Cauchy could not have just made a mistake about such a thing and that he was thinking either about transfinite "sequences" of functions or the nonstandard real line (or both). This is no more than a combination of wishful thinking and hero worship. Cauchy *was* simply wrong (cf. [Dau88], [God04, p. 211]).

> "Now it is clearly quite implausible to ascribe to Cauchy or Leibniz even an inchoate understanding of a first-order transfer principle since ... the distinction between first- and second-order properties was made available only by the development of logic in the late 19th century, and even then it did not straightaway seem especially important. Even now very few mathematicians who are not logicians have any firm grip on which of the concepts they make use of are of first- and which second-order."
>
> *Michael Potter*[5]

[5] [Pot04, p. 145].

Cauchy was a great mathematician (and a strong proponent of increased rigor in mathematics), but he was somewhat prone to errors in his writing (e.g. his "proof" of the Mean Value Theorem; cf. [Bre07]), and his understanding of the foundations of analysis, while far ahead of his predecessors, was still fairly murky compared to today; the mistake in this case was a lack of understanding of the distinction between pointwise and uniform convergence, which was not definitively clarified (by Weierstrass) until considerably later. Cauchy's concept of limit was still based on infinitesimals, but to Cauchy an infinitesimal was a dynamic concept, a variable real number converging to zero. See [Gra00], [Jah03], [Koe91], and [Bre07] for a discussion.

IV.3.2. U.C. Convergence

Uniform convergence is very nice and has important consequences, but it is in a sense too nice – in many natural situations we do not have uniform convergence. But it turns out that a slightly weaker form of convergence is almost as good, and much more common: uniform convergence on compact sets, or u.c. convergence.

IV.3.2.1. Definition. Let X be a topological space, and (f_n) a sequence of functions from X to $\mathbb{R}$ (or a metric space Y). If $f : X \to \mathbb{R}$ (or Y), then $f_n \to f$ *uniformly on compact sets (u.c.)* on X if $f_n|_K \to f|_K$ uniformly for every compact subset K of X.

Note that since singletons are compact, u.c. convergence implies pointwise convergence (but usually turns out to be considerably stronger). Uniform convergence obviously implies u.c. convergence. If X itself is compact, u.c. convergence on X is the same as uniform convergence.

IV.3.2.2. Perhaps the most common and important situation is when $X = \mathbb{R}^n$ or a (usually open) subset. Since a subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded, $f_n \to f$ u.c. on X if and only if $f_n|_K \to f_K$ uniformly for every closed bounded subset K of X (note that "closed" means closed in $\mathbb{R}^n$, not just relatively closed in X).

The name "u.c. convergence" is not as standard as it should be; it is a very convenient and appropriate abbreviation which we will use frequently (and not just because the author got his Ph.D. at Berkeley!) In Complex Analysis, u.c. convergence of sequences of holomorphic functions on an open set in $\mathbb{C}$ is sometimes called *normal convergence* (XII.4.1.1.), but this name seems to be rarely used in a broader context (and it is also somewhat ambiguous, cf. V.15.1.4.). Some authors also use the term *compact convergence*.

IV.3.2.3. By considering partial sums, we can also say what is meant for an infinite series of functions to converge u.c.

Here are some examples showing that u.c. convergence is common and considerably weaker than uniform convergence.

IV.3.2.4. Examples. (i) Let $f_n : \mathbb{R} \to \mathbb{R}$ be given by $f_n(x) = x/n$. Then $f_n \to 0$ pointwise, but not uniformly. But restricted to any bounded set, $f_n \to 0$ uniformly, i.e. $f_n \to 0$ u.c.

(ii) On $(0, +\infty)$ (or any subset, e.g. $(0, 1)$), let $f_n(x) = \frac{1}{nx}$. Again the sequence converges pointwise to 0 but not uniformly. But any compact subset of $(0, +\infty)$ is contained in an interval $[\epsilon, M]$ for some ϵ, M, $0 < \epsilon < M < +\infty$, and on any such interval the sequence converges uniformly to 0. Thus $f_n \to 0$ u.c.

(iii) Consider the infinite series $\sum_{k=0}^{\infty} x^k$ where the terms are regarded as real-valued functions on $(-1,1)$ (which is the set of x for which the series converges). The infinite series converges pointwise to $f(x) = \frac{1}{1-x}$. More precisely,

$$s_n = \sum_{k=0}^{n} x^k = \frac{1}{1-x} - \frac{x^n}{1-x}$$

so, if $0 < r < 1$, for $x \in [-r, r]$, we have

$$\left| \frac{1}{1-x} - s_n \right| = \left| \frac{x^n}{1-x} \right| \leq \frac{r^n}{1-r}$$

which can be made uniformly small by taking n large enough (depending on r). Since every closed (in $\mathbb{R}$) subset of $(-1,1)$ is contained in $[-r,r]$ for some $r < 1$, the infinite series converges u.c. on $(-1,1)$. But since the partial sums are bounded on $(-1,1)$ and the limit function is unbounded, the convergence cannot be uniform on all of $(-1,1)$.

More generally, any power series converges u.c. on its interval (or open disk) of convergence ().

(iii) Let (f_n) be the sequence of Figure IV.1 of functions from $[0,1]$ to $\mathbb{R}$. Then $f_n \to 0$ pointwise but not u.c. (i.e. not uniformly since $[0,1]$ is compact). Note that the sequence is not uniformly bounded. If the height of the peaks is taken to be 1 for all n, the sequence is uniformly bounded but still does not converge u.c.

IV.3.2.5. PROPOSITION. Let X be a locally compact topological space, and $f_n, f : X \to \mathbb{R}$ (or a metric space). Then $f_n \to f$ u.c. if and only if $f_n \to f$ uniformly on some neighborhood of each point of X.

PROOF: One direction is obvious, since every point of a locally compact space has a compact neighborhood. Conversely, suppose $f_n \to f$ on some neighborhood of each point of X. If K is a compact subset of X, for each $x \in K$ there is an open neighborhood U_x of x on which $f_n \to f$ uniformly. Finitely many $U_{x_1}, \ldots, U_{x_m}$ of the U_x cover K, and $f_n \to f$ uniformly on $U_{x_1} \cup \cdots \cup U_{x_m}$ and hence on K. 👍

IV.3.2.6. A uniform limit of continuous functions is continuous (). The same is not true for u.c. convergence in complete generality (XIII.17.4.1.(b)); the problem is that a general X may not have enough compact subsets, illustrated by the fact that there are nondiscrete topological spaces in which every compact subset is finite, and on such a space u.c. convergence is just pointwise convergence, which does not necessarily preserve continuity. But under a mild restriction, u.c. convergence does preserve continuity:

IV.3.2.7. THEOREM. Let X be a topological space, and (f_n) a sequence of functions from X to $\mathbb{R}$ (or a metric space) converging u.c. to a function f. If each f_n is continuous on compact sets, then f is continuous on compact sets.

PROOF: This is an immediate corollary of (). 👍

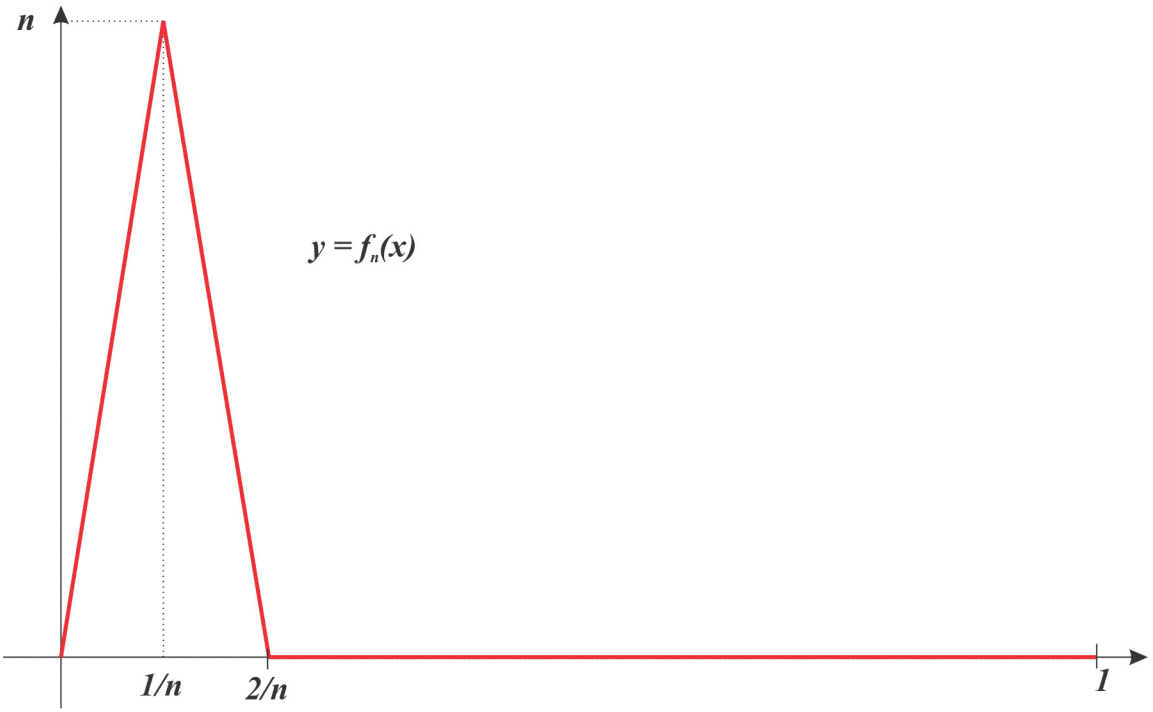


Figure IV.1: A Sequence of Functions

IV.3.2.8. Corollary. Let X be a compactly generated (e.g. first countable) topological space (XIII.11.11.1.), and (f_n) a sequence of functions from X to $\mathbb{R}$ (or a metric space) converging u.c. to a function f. If each f_n is continuous, then f is continuous.

IV.3.2.9. The corollary applies if X is any subset of $\mathbb{R}^n$, or more generally any subset of a metrizable space. The conclusion is obvious from () if X is locally compact, since continuity is a local property.

IV.3.3. Mean and Mean-Square Convergence

IV.3.4. Exercises

IV.3.4.1. [Dug78, XII.7.5] Let X and Y be topological spaces, and f, f_n $(n \in \mathbb{N})$ functions from X to Y. Say $f_n \to f$ *continuously* if, whenever (x_n) is a sequence in X and $x_n \to x$, then $f_n(x_n) \to f(x)$ in Y. Show that if X is first countable and Y is a metric space, and f and the f_n are continuous, then $f_n \to f$ continuously if and only if $f_n \to f$ u.c.

IV.3.4.2. Let X be a topological space, say first countable, and (Y, ρ) a metric space, and $f, \{f_n : n \in \mathbb{N}\}$ functions from X to Y. Then

$f_n \to f$ *locally approximately uniformly* if for every $a \in X$ and $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that, for every $n \geq N$, there is a neighborhood U of a in X such that $\rho(f_n(x), f(x)) < \epsilon$ for all $x \in U$.

$f_n \to f$ *locally uniformly* if for every $a \in X$ there is a neighborhood U of a in X such that $f_n \to f$ uniformly on U.

(a) Show that if each f_n is continuous, then $f_n \to f$ locally approximately uniformly if and only if $f_n \to f$ pointwise and f is continuous.

(b) Show that if X is locally compact, then $f_n \to f$ locally uniformly if and only if $f_n \to f$ u.c.

(c) Show that local uniform convergence implies local approximate uniform convergence, but the converse is false in general, even if $X = [0, 1]$ and each f_n is continuous.

There is nonuniformity of terminology concerning these types of convergence: some references such as [God04] call local approximate uniform convergence "locally uniform convergence." Locally uniform convergence should mean what we have defined it to be.

Chapter V

Calculus

"The calculus was the first achievement of modern mathematics, and it is difficult to overestimate its importance."

John von Neumann[1]

"Today, more than three centuries after its appearance, calculus continues to warrant such praise. It is the bridge that carries students from the basics of elementary mathematics to the challenges of higher mathematics and, as such, provides a dazzling transition from the finite to the infinite, from the discrete to the continuous, from the superficial to the profound. So esteemed is calculus that its name is often preceded by 'the,' as in von Neumann's observation above. This gives '*the* calculus' a status akin to '*the* law' – that is, a subject vast, self-contained, and awesome."

William Dunham[2]

[1] *Collected Works*, p. 3.
[2] [Dun18, p. 1].

V.1. Limits of Functions

The limit of a function at a point is a variation of the notion of limit of a sequence, and is one of the most fundamental ideas in calculus and analysis. In this section, we consider only limits of functions from $\mathbb{R}$ to $\mathbb{R}$; in later sections, the situation will be generalized in a number of ways.

V.1.1. Ordinary Limits

V.1.1.1. Suppose f is a real-valued function defined at least on an interval around a real number a, except possibly at a itself (i.e. defined at least on a deleted neighborhood of a ()). If we don't know the value $f(a)$ or even whether this value exists, but can calculate $f(x)$ for all x near a but different from a, we want to determine what $f(a)$ "ought to be" (it will be impossible to determine whether this prediction is correct without knowledge about the value precisely at a). Geometrically, we consider the graph $y = f(x)$ and cover up the vertical line $x = a$, and try to determine where the point of the graph on this line should be from what can be seen:

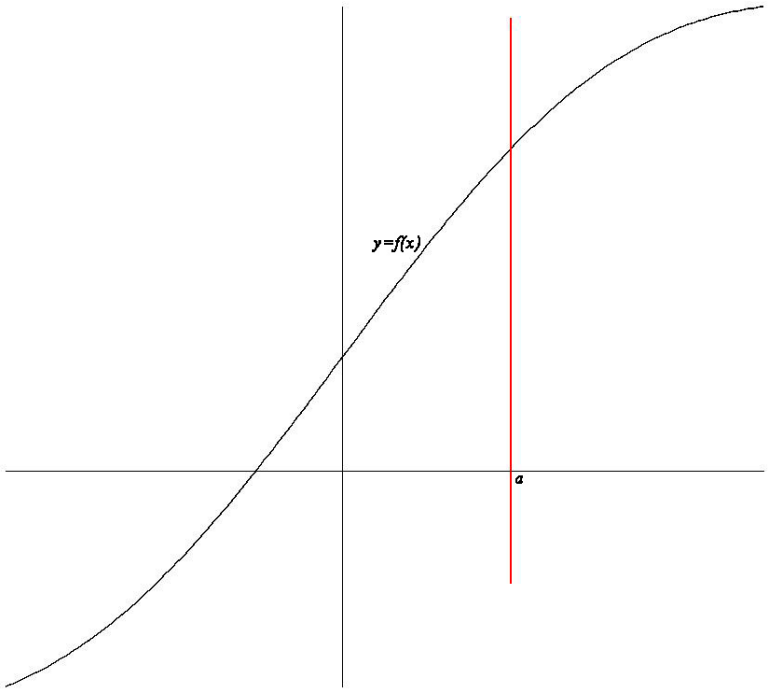


Figure V.1: Limit of a function

When I teach limits in Calculus, I take a piece of string to class and cover up the line $x = a$ on various graphs, and ask the students where the point, if any, under the string *is* and where it *should be*. Note that it will be impossible to determine from what can be seen whether there is in fact a point of the graph on this line, or, if so, where it is, but it can in many (but not all) cases be determined where it *ought* to be. Informally, we will say the limit of the function as x approaches a will be the y-coordinate of the point on

this vertical line (i.e. with x-coordinate a) which "should" be on the graph, if it can in fact be determined, and otherwise that the limit does not exist.

V.1.1.2. More formally, as in the case of limits of sequences, we will say that the limit of $f(x)$ as x approaches a is L, written

$$\lim_{x \to a} f(x) = L$$

or $f(x) \to L$ as $x \to a$, if $f(x)$ can be made arbitrarily close to L by making x close enough to a, but never equal to a. This can be phrased similarly to the definition of limit of a sequence, with a slight variation:

V.1.1.3. DEFINITION. Let V be a deleted neighborhood of $a \in \mathbb{R}$, $F : V \to \mathbb{R}$ a function, and $L \in \mathbb{R}$. Then the *limit* of $f(x)$ as x approaches a is L if, for every $\epsilon > 0$, there is a $\delta > 0$ such that

$$|f(x) - L| < \epsilon$$

for all $x \in V$ for which $0 < |x - a| < \delta$.

Note that the inequality "$0 < |x - a| < \delta$" is just a slick way of writing "$a - \delta < x < a + \delta$ and $x \neq a$."

V.1.1.4. Geometrically, $\lim_{x \to a} f(x) = L$ if, for any $\epsilon > 0$, no matter how small, if we draw the horizontal lines $y = L + \epsilon$ and $y = L - \epsilon$ above and below $y = L$, then there are vertical lines $x = a - \delta$ and $x = a + \delta$ to the left and right of a, such that the part of the graph $y = f(x)$ between these two vertical lines lies between the two horizontal lines, except possibly when $x = a$ (Figure V.2).

V.1.1.5. As with sequences (IV.1.3.9.), the definition of limit of a function can be described as a game or bet. If I claim that $\lim_{x \to a} f(x) = L$ and you challenge me, you give me an $\epsilon > 0$ and I must produce a $\delta > 0$ that works in the definition for your ϵ. If I can, I win; if I can't, you win. The smaller your ϵ is, the harder it is for me to find a δ, and the smaller it will have to be in general. But I don't have to choose my δ until you give me your ϵ; my δ can (and in general will) depend on your ϵ. If one δ works for a given ϵ, then any smaller δ also works for the same ϵ, so I don't have to find the largest possible δ for your ϵ; a smaller one which is easier to find will work just as well. Similarly, a δ which works for one ϵ will also work for any larger ϵ. A limit proof will be a demonstration that I can always win the bet no matter what ϵ you give me.

V.1.1.6. To apply the definition of limit directly, the limit L must be "guessed." Thus, to just show that a limit exists, and especially to show that a limit does not exist, one must work somewhat indirectly in a manner similar to the one in ().

V.1.1.7. Note that in the definition of limit, the x is allowed to be either to the right of a ($x > a$) or to the left of a ($x < a$). Thus the limit is often called the "two-sided limit" for emphasis. (But "limit" will always for us mean "two-sided limit" unless otherwise specified.) The "limit" as defined in V.1.1.3. is also sometimes called the "ordinary limit" to distinguish it from the extended notions of limit discussed in ().

V.1.1.8. Of course, $\lim_{x \to a} f(x)$ depends not only on the function f but also on the number a: for a fixed f, the limit can potentially be defined for any a, although in practice the limit is particularly important only for certain a's depending on f.

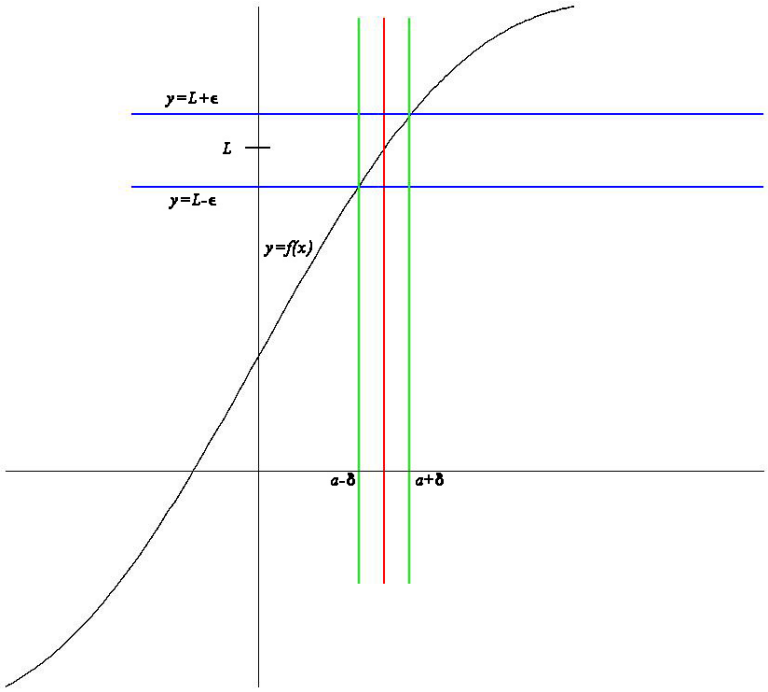


Figure V.2: Limit of a function

V.1.1.9. Note that there is a huge conceptual difference for a function f between the *value of* f *at* a and the *limit of* f *as* x *approaches* a. Either can exist without the other, and if they both exist they may or may not be numerically equal. It turns out that for "most" functions encountered naturally, and for "most" a, the two both exist and are numerically equal (the notion of *continuity* discussed in ()); this fact is ultimately very pleasant and significant, but can be confusing for beginners trying to distinguish between the concepts.

The following principle is basic in the study of limits:

V.1.1.10. PROPOSITION. Let f and g be functions defined in at least a deleted neighborhood of a. If $f(x) = g(x)$ for all x in a deleted neighborhood of a, and $\lim_{x\to a} f(x)$ exists, then $\lim_{x\to a} g(x)$ also exists and equals $\lim_{x\to a} f(x)$.

Note that we do *not* require that $f(a) = g(a)$, or even that either or both of these numbers are defined.

V.1.1.11. EXAMPLES. (i) Let $f(x) = x + 1$, $g(x) = \frac{x^2-1}{x-1}$,

$$h(x) = \begin{cases} x+1 & \text{if} \quad x \neq 1 \\ 3 & \text{if} \quad x = 1 \end{cases} .$$

We have that $f(1) = 2$, $g(1)$ does not exist, and $h(1) = 3$. The graph of f is a straight line with slope 1 and y-intercept 1; the graph of g is the same line with the point $(1, 2)$ removed, and the graph of h is the line with $(1, 2)$ removed and $(1, 3)$ added. If the vertical line $x = 1$ is covered, the graphs appear identical.

We guess that $\lim_{x\to 1} f(x) = 2$. To prove this, let $\epsilon > 0$. We must find a $\delta > 0$ such that

$$|f(x) - 2| = |x + 1 - 2| = |x - 1| < \epsilon$$

whenever $0 < |x - 1| < \delta$. We can clearly take $\delta = \epsilon$, proving the result. This simple formula makes sense since the graph is a straight line with slope 1 (if the line had slope m we would take $\delta = \frac{\epsilon}{m}$; if the function were nonlinear the formula would generally be more complicated).

Since $g(x) = h(x) = f(x)$ for $x \neq 1$, we also have

$$\lim_{x\to 1} g(x) = \lim_{x\to 1} h(x) = 2$$

by V.1.1.10.; these can be proved directly with exactly the same argument as for f. Note that g has a limit but no value at 1, while h has both a value and a limit at 1 but they are not numerically equal.

(ii) Set

$$f(x) = \frac{|x|}{x} = \begin{cases} 1 & \text{if} \quad x > 0 \\ -1 & \text{if} \quad x < 0 \end{cases}$$

$$g(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ -1 & \text{if} \quad x < 0 \end{cases} .$$

Then neither $\lim_{x\to 0} f(x)$ nor $\lim_{x\to 0} g(x)$ exists, by an argument almost identical to the one in (). (These functions have one-sided limits at 0 ().) Although $f(0)$ does not exist, we have that $g(0) = 1$. Thus g has a value but no limit at 0.

(iii) Let $f(x) = \sin\frac{1}{x}$, $g(x) = x\sin\frac{1}{x}$ for $x \neq 0$. Neither f nor g has a value at 0. Also, $\lim_{x\to 0} f(x)$ does not exist; there are not even one-sided limits. See Figure (). But $\lim_{x\to 0} g(x)$ does exist: it equals 0. For let $\epsilon > 0$. Set $\delta = \epsilon$. If $0 < |x - 0| = |x| < \delta$, we have

$$|g(x) - 0| = \left| x \sin\frac{1}{x} \right| \leq |x| < \delta = \epsilon$$

since $\left|\sin\frac{1}{x}\right| \leq 1$ for all $x \neq 0$. See Figure ().

(iv) Let $f(x) = \sqrt{x}$. If $a > 0$, $\lim_{x\to a} f(x)$ exists and equals $\sqrt{a}$ (). But $\lim_{x\to 0} f(x)$ cannot exist since f is not defined in any entire deleted neighborhood of 0. We do have a one-sided limit of 0 from the right in this case ().

V.1.2. The Sequential Criterion

There is a very useful reformulation of the definition of limit of a function using sequences:

V.1.2.1. THEOREM. [SEQUENTIAL CRITERION FOR LIMITS] Let f be a real-valued function defined on at least a deleted neighborhood I of a. Then $\lim_{x\to a} f(x) = L$ if and only if, whenever (x_n) is a sequence in I with $x_n \neq a$ for all n and $x_n \to a$, we have $f(x_n) \to L$.

V.1.2.2. Note that the theorem (reverse direction) requires that $f(x_n) \to L$ for *every* sequence (x_n) in I satisfying the conditions. It is not enough that there is one such sequence (). The condition $x_n \neq a$ for all n is actually automatic since $a \notin I$, but is explicitly stated for emphasis; we cannot allow any x_n to be a in general since $f(a)$ need not be defined, and may be unequal to L even if it is defined.

PROOF: Suppose $\lim_{x\to a} f(x) = L$, and (x_n) is a sequence in I, $x_n \neq a$ for all n, and $x_n \to a$. Let $\epsilon > 0$. Then there is a $\delta > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - a| < \delta$. Since $x_n \to a$, there is an N such that $0 < |x_n - a| < \delta$ for all $n \geq N$. Thus $|f(x_n) - L| < \epsilon$ for all $n \geq N$. Since $\epsilon > 0$ is arbitrary, $f(x_n) \to L$. Conversely, suppose $\lim_{x\to a} f(x) = L$ is false. Then there is an $\epsilon > 0$ such that, for every $\delta > 0$, there is an $x \in I$ with $0 < |x - a| < \delta$ and $|f(x) - L| \geq \epsilon$. For this ϵ, for each n choose an $x_n \in I$ with $0 < |x_n - a| < \frac{1}{n}$ and $|f(x_n) - L| \geq \epsilon$. Thus $x_n \neq a$ for all n and $x_n \to a$, but $f(x_n) \not\to L$, so the sequential criterion does not hold.

V.1.2.3. This proof requires the Countable AC to choose the x_n. In fact, the reverse direction cannot be proved without some form of the Countable AC.

V.1.3. Limit Theorems

We can use the Sequential criterion to give quick proofs of the important basic properties of limits of functions, although these can also be proved directly by arguments similar to the proofs of the corresponding results for sequences (Exercise V.1.8.1.).

V.1.3.1. PROPOSITION. Function limits, when they exist, are unique: let f be a real-valued function defined on at least a deleted neighborhood I of a. If $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} f(x) = M$, then $L = M$.

PROOF: Let (x_n) be a sequence in I with $x_n \to a$ (and of course $x_n \neq a$ for all n). Then, by the sequential criterion, $\lim_{n\to\infty} f(x_n) = L$ and $\lim_{n\to\infty} f(x_n) = M$. Thus $L = M$ by IV.1.3.13..

V.1.3.2. Thus we may legitimately talk about *the* limit of a function (when it exists). Note that no Choice is needed here since the x_n can be chosen by a recipe, e.g. $x_n = a + \frac{1}{n}$ for all sufficiently large n.

Similarly, from IV.1.3.18. we have:

V.1.3.3. PROPOSITION. Let f be a real-valued function defined on at least a deleted neighborhood of a. If $c \leq f(x) \leq d$ for all x in a deleted neighborhood of a. and $\lim_{x\to a} f(x)$ exists, then

$$c \leq \lim_{x\to a} f(x) \leq d \ .$$

V.1.3.4. As in IV.1.3.19., one cannot replace the inequalities with strict inequalities.

V.1.3.5. THEOREM. Let f, g be real-valued functions defined at least on a deleted neighborhood I of a. Suppose $\lim_{x\to a} f(x) = L$ and $\lim_{x\to a} g(x) = M$. Then

(i) $\lim_{x\to a}[f(x) + g(x)]$ exists and equals $L + M$.

(ii) $\lim_{x\to a}[f(x)g(x)]$ exists and equals LM.

(iii) If $M \neq 0$, then $g \neq 0$ on a smaller deleted neighborhood of a, and $\lim_{x\to a} \dfrac{f(x)}{g(x)}$ exists and equals $\dfrac{L}{M}$.

PROOF: We give the proof of (i); the others are almost identical. Let (x_n) be a sequence in I with $x_n \to a$ (and of course $x_n \neq a$ for all n). Then, by the sequential criterion, $\lim_{n\to\infty} f(x_n) = L$ and $\lim_{n\to\infty} g(x_n) = M$. Thus, by IV.1.3.21.,

$$\lim_{n\to\infty}[f(x_n) + g(x_n)] = L + M \ .$$

Since (x_n) is an arbitrary sequence in I converging to a, and $[f + g](x_n) = f(x_n) + g(x_n)$ by definition, we have

$$\lim_{x\to a}[f(x) + g(x)] = L + M$$

by the sequential criterion. 👍

V.1.3.6. As in IV.1.3.22., The theorem is often informally stated as "the limit of a sum [product, quotient] is the sum [product, quotient] of the limits," and written

$$\lim_{x\to a}[f(x) + g(x)] = \left[\lim_{x\to a} f(x)\right] + \left[\lim_{x\to a} g(x)\right]$$

$$\lim_{x\to a}[f(x)g(x)] = \left[\lim_{x\to a} f(x)\right]\left[\lim_{x\to a} g(x)\right]$$

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)}$$

(provided $\lim_{x\to a} g(x) \neq 0$).

However, stating the result this way can be misleading if one is not careful. The results only hold under the assumption that the limits on the right side exist; existence of the limits on the left in these expressions does not mean the limits on the right exist and that the limits on the left can be calculated in the indicated way. So care must be exercised not to try to apply the converse of the statements of V.1.3.5.. ✋

V.1.3.7. Note that V.1.3.5.(iii) does not apply if $M = 0$, even if $g(x) \neq 0$ for all x near a. If $M = 0$ and $L \neq 0$, then $\lim_{x\to a} \dfrac{f(x)}{g(x)}$ cannot exist (since if it did, and equaled C, we would have

$$L = \lim_{x\to a}\left[g(x)\frac{f(x)}{g(x)}\right] = \left[\lim_{x\to a} g(x)\right]\left[\lim_{x\to a} \frac{f(x)}{g(x)}\right] = 0 \cdot C = 0$$

by (ii), which is impossible). But if $L = M = 0$, the limit of the quotient can exist, but cannot be computed simply in general. This is a type of *indeterminate form*. One of the important parts of calculus and analysis is the resolution of indeterminate forms of various types; see the next section.

V.1.4. Indeterminate Forms

Indeterminate forms or indeterminate expressions

In [CJ99] it is asserted that the term *indeterminate expression* is "misleading and vague" since the limit of such an expression is often not indeterminate (i.e. can be determined), and that the subject of indeterminate expressions, especially of the form $\frac{0}{0}$, is discussed "at exaggerated length" in many textbooks. I think the term *indeterminate form*, at least, is appropriate and useful, since the limit of such an expression cannot be determined or even guessed by inspection or by straightforward applications of algebraic limit theorems and requires an indirect procedure or argument to evaluate the limit or even determine whether it exists. And, in my experience, one of the most important skills a calculus student or user must learn is to be able to recognize indeterminate forms and know how to evaluate them properly. For example, the difference quotient limit appearing in the definition of the derivative of a function is always (for a continuous function) an indeterminate form of type $\frac{0}{0}$, and is the most important class of limits in the whole subject.

V.1.5. One-Sided Limits

There are one-sided versions of the Sequential criterion of V.1.2.1.:

V.1.5.1. Theorem. [Sequential Criterion for One-sided Limits] Let f be a real-valued function defined at least on an interval $I = (a, b)$, $a < b$.

(i) $\lim_{x \to a+} f(x) = L$ if and only if, whenever (x_n) is a sequence in I with $x_n > a$ for all n and $x_n \to a$, we have $f(x_n) \to L$.

(ii) $\lim_{x \to b-} f(x) = L$ if and only if, whenever (x_n) is a sequence in I with $x_n < b$ for all n and $x_n \to b$, we have $f(x_n) \to L$.

The proof is a simple modification of the proof of V.1.2.1., and is left to the reader.

V.1.6. Extended Limits

V.1.7. Limit Superior and Limit Inferior

V.1.8. Exercises

V.1.8.1. Give direct proofs of the limit theorems for functions along the lines of the corresponding results for sequences.

V.1.8.2. (a) Find $\lim_{x \to 0-} \dfrac{1}{1 - e^{1/x}}$ and $\lim_{x \to 0+} \dfrac{1}{1 - e^{1/x}}$.

(b) Find $\lim_{x \to 0} \dfrac{x}{1 - e^{1/x}}$, or show the limit does not exist.

V.1.8.3. ([dlVP46, 87, Ex.6], attributed to Cauchy) Let ϕ be a real-valued function on an interval $(a, +\infty)$. Suppose $\lim_{x \to +\infty} [\phi(x + 1) - \phi(x)]$ exists (in the usual or extended sense) and equals L.

(a) If ϕ is bounded on $[b, b+1]$ for all sufficiently large b (e.g. if ϕ is continuous), show that $\lim_{x\to+\infty} \dfrac{\phi(x)}{x}$ also exists and equals L.

(b) Show that $\lim_{x\to+\infty} \dfrac{\phi(x)}{x}$ need not exist if no boundedness assumption is made. [Consider a Hamel basis for $\mathbb{R}$ over $\mathbb{Q}$ which contains 1 and is not bounded above, and let $\phi(x)$ be the coefficient of 1 in the expansion of x with respect to this basis.] But if the limit does exist, it must equal L.

(c) Suppose ψ is a function from $(a, +\infty)$ to $(0, +\infty)$ for which $\lim_{x\to+\infty} \dfrac{\psi(x+1)}{\psi(x)}$ exists and equals L. Find similar conditions on ψ insuring that $\lim_{x\to+\infty} [\psi(x)]^{1/x}$ exists (and then necessarily equals L). [Consider $\phi = \log \psi$.] Compare with IV.2.3.11..

This is a discrete version of l'Hôpital's Rule.

V.2. Continuity on $\mathbb{R}$

> "[Mathematicians before the early 19th century] assumed that any sensible formula, such as a polynomial, automatically defined a continuous function. But they had never proved this. In fact, they couldn't prove it, because they hadn't defined 'continuous.' The whole area was awash with vague intuitions, most of which were wrong."
>
> *Ian Stewart*[3]

V.2.1. Continuity at a Point

V.2.2. Continuity on a Set

In discussions and uses of continuity, it is much more common to use the notion of continuity on a set rather than continuity at a point. The transition between the two is simple but involves a subtlety which must be understood. It should be noted that conventions vary in references as to the meaning of the phrase "continuity on a set A," and the convention in any reference should be noted carefully. We have picked the convention which seems the most common and useful.

The definition is a slight variation of ():

V.2.2.1. Definition. Let f be a real-valued function defined on a subset of $\mathbb{R}$, and let A be a subset of the domain of f. Then f is *continuous on* A if, for every $a \in A$ and $\epsilon > 0$, there is a $\delta > 0$ such that $|f(x) - f(a)| < \epsilon$ whenever $x \in A$ and $|x - a| < \delta$.

V.2.2.2. If f is continuous (in the sense of ()) at every point of A, then f is continuous on A. The converse is false in general, since in V.2.2.1. only values of the function at points of A are considered; if a point a of A is a limit point of A^c the behavior of f on A^c can force f to be discontinuous at a, but will not affect whether f is continuous on A.

V.2.2.3. Examples. (i) Let

$$f(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ 0 & \text{if} \quad x < 0 \end{cases} .$$

Then f is discontinuous at 0, but it is continuous on $[0, \infty)$.

(ii) More generally, if $f = \chi_A$ is the indicator function of a set $A \subseteq \mathbb{R}$, then f is continuous on A since it is a constant function on A. Similarly, f is continuous on A^c. But f can be discontinuous at many numbers: e.g. if $A = \mathbb{Q}$, f is discontinuous everywhere (i.e. at every $a \in \mathbb{R}$).

(iii) If $A = \{a\}$, then any real-valued function defined at a is continuous on $\{a\}$ since the condition in the definition is essentially vacuous. Thus "continuity *on* $\{a\}$" is quite different from "continuity *at* a." The similarity of the terminology is somewhat unfortunate, but such near-ambiguity is unavoidable.

V.2.2.4. Continuity on a set A is most frequently used when A is either an open set or a closed or half-open interval. In these cases, the ambiguities largely disappear:

[3] [?, p. 151].

V.2.2.5. PROPOSITION. Let f be a real-valued function on a subset of $\mathbb{R}$, and U an open set in $\mathbb{R}$ contained in the domain of f. Then f is continuous on U if and only if f is continuous (in the sense of ()) at a for every $a \in U$.

V.2.2.6. PROPOSITION. Let f be a real-valued function on a subset of $\mathbb{R}$, and I an interval in $\mathbb{R}$ contained in the domain of f. Then f is continuous on I if and only if f is continuous (in the sense of ()) at a for every a in the interior of I, and right continuous () at the left endpoint of I and left continuous () at the right endpoint of I when there are such endpoints in I.

In other words, f is continuous on $[a, b]$ if and only if it is continuous at every point of (a, b), right continuous at a, and left continuous at b; it is continuous on $[a, b)$ if and only if it is continuous at every point of (a, b) and right continuous at a, etc.

V.2.2.7. We say a function f defined on a subset of $\mathbb{R}$ is *continuous* if it is continuous on its domain.

Three Characterizations of Continuity

There is a sequential criterion for continuity on a set:

V.2.2.8. THEOREM. [SEQUENTIAL CRITERION] Let f be a real-valued function on a subset of $\mathbb{R}$, and A a subset of $\mathbb{R}$ contained in the domain of f. Then f is continuous on A if and only if, whenever (x_n) is a sequence in A, $x \in A$, and $x_n \to x$, we have $f(x_n) \to f(x)$.

Note that in this criterion, the terms x_n of the sequence *and the limit* x must all be in A.
PROOF: Suppose f is continuous on A, and let (x_n) be a sequence in A converging to $x \in A$. Let $\epsilon > 0$. Since f is continuous at x, there is a $\delta > 0$ such that $|f(y) - f(x)| < \epsilon$ for all $y \in A$ with $|y - x| < \delta$. There is an N such that $|x_n - x| < \delta$ for all $n \geq N$. Then $|f(x_n) - f(x)| < \epsilon$ for all $n \geq N$. Thus $f(x_n) \to f(x)$.
Conversely, suppose the sequential criterion is satisfied. If f is not continuous on A, then there is an $x \in A$ and $\epsilon > 0$ for which there is no corresponding δ with $|f(y) - f(x)| < \epsilon$ whenever $y \in A$ and $|y - x| < \delta$; thus, for this x and ϵ, for each $n \in \mathbb{N}$ there is an $x_n \in A$ with $|x_n - x| < \frac{1}{n}$ but $|f(x_n) - f(x)| \geq \epsilon$. So $x_n \to x$ but $f(x_n) \not\to f(x)$, a contradiction. 👌

Note that the Countable AC is used in this proof. (In fact, this theorem cannot be proved in ZF ().)

There is a third standard characterization of continuity in terms of preimages:

V.2.2.9. THEOREM. Let $A \subseteq \mathbb{R}$, and $f : A \to \mathbb{R}$ a function. Then the following are equivalent:

(i) f is continuous on A.

(ii) The preimage of every open set in $\mathbb{R}$ is (relatively) open in A.

(iii) The preimage of every open interval in $\mathbb{R}$ is (relatively) open in A.

(iv) The preimage of every closed set in $\mathbb{R}$ is (relatively) closed in A.

PROOF: (ii) and (iv) are transparently equivalent since preimages respect complements (). (ii) ⇒ (iii) is trivial.
(i) ⇒ (ii): Suppose f is continuous, and let U be an open set in $\mathbb{R}$, and $x \in f^{-1}(U)$. There is an $\epsilon > 0$ such that $\{y \in \mathbb{R} : |y - f(x)| < \epsilon\} \subseteq U$ since U is open, and there is a $\delta > 0$ such that $|f(z) - f(x)| < \epsilon$ whenever $z \in A$ and $|z - x| < \delta$ since f is continuous at x. Thus if $z \in A$, $|z - x| < \delta$, then $f(z) \in U$, i.e. $z \in f^{-1}(U)$. So $f^{-1}(U)$ is open.
(iii) ⇒ (i): Let $x \in A$ and $\epsilon > 0$. Let $U = \{y \in \mathbb{R} : |y - x| < \epsilon\}$. Then U is an open interval, so $f^{-1}(U)$ is relatively open in A by assumption. Since $x \in f^{-1}(U)$, there is a $\delta > 0$ such that $A \cap (x - \delta, x + \delta) \subseteq f^{-1}(U)$. If $z \in A$, $|z - x| < \delta$, then $z \in f^{-1}(U)$, i.e. $f(z) \in U$, $|f(z) - f(x)| < \epsilon$. So f is continuous at x. 👍

V.2.2.10. Thus we have three equivalent characterizations of continuous functions:

(i) The $\epsilon - \delta$ characterization (definition).

(ii) The sequential criterion.

(iii) The preimage characterization.

Each is useful in different contexts: some applications are much easier using one or the other characterization.

There are more general settings for continuity: metric spaces () and, even more generally, topological spaces (). Characterization (i) is special to metric spaces. The sequential criterion is valid more generally in topological spaces for which sequences determine the topology (first-countable topological spaces). Only the third is appropriate in general topological spaces, and is taken as the definition of continuity in general spaces. The three characterizations are equivalent in metric spaces (). The definition V.2.2.1. precisely says that a function from $\mathbb{R}$ (or a subset) to $\mathbb{R}$ is continuous on a subset A of its domain if and only if the restriction of f to A is continuous in this generalized sense when A is given the relative or subspace topology.

V.2.3. The Intermediate Value Theorem

One of the most important properties of a continuous function on an interval is that its range is also an interval, i.e. that it takes all intermediate values between any two of its values:

V.2.3.1. THEOREM. [INTERMEDIATE VALUE THEOREM (IVT)] Let I be an interval in $\mathbb{R}$, $f : I \to \mathbb{R}$ a continuous function, and $a, b \in I$. If d is any real number between $f(a)$ and $f(b)$, then there is at least one c between a and b with $f(c) = d$.

V.2.3.2. Geometrically, the graph of f gives an "unbroken" curve between $(a, f(a))$ and $(b, f(b))$, and thus any horizontal line of intermediate height must intersect it (Figure () is the case $f(a) < f(b)$; the other case is similar).

This picture, although informal, gives a clue how to prove the theorem from the Completeness Axiom (): take the supremum of the x-coordinates of all points on the graph below (or on) the line; this should be the x-coordinate of the "last" point of intersection (we could equally well take the infimum for points above the line, giving the "first" point of intersection; of course, there may be many other points of intersection too). In fact, the IVT is equivalent to the Completeness Axiom (Exercise ()).

PROOF: We may assume without loss of generality that $a < b$. Suppose $f(a) < f(b)$ (the case $f(a) > f(b)$ is similar, and if $f(a) = f(b)$ there is nothing to prove). Fix d with $f(a) < d < f(b)$, and set

$$S = \{x \in [a, b] : f(x) \leq d\} .$$

Then S is a nonempty subset of $[a, b]$ since $a \in S$. Set $c = \sup S$. We claim $f(c) = d$. If $f(c) \neq d$, then there is a $\delta > 0$ such that $|f(x) - f(c)| < |f(c) - d|$ for $x \in [a, b]$, $|x - c| < \delta$. If $f(c) > d$, then $c \neq a$ and $f(x) > d$ for $x \in [a, b]$ and $c - \delta < x < c$ (thus $c - \delta > a$ since $f(a) < d$), so $c - \delta$ is an upper bound for S, a contradiction. And if $f(c) < d$, then $c \neq b$ and if $x \in [a, b]$, $c < x < c + \delta$, then $f(x) < d$, contradicting that c is an upper bound for S. 👍

V.2.3.3. COROLLARY. Let f be a continuous real-valued function on an interval I. Then $f(I)$ is an interval (or single point if f is constant on I).

V.2.3.4. This result can be done in the more general context of connected subsets of topological spaces: a continuous image of a connected set is connected (XIII.13.2.1.). The connected subsets of $\mathbb{R}$ are precisely the intervals and singletons (degenerate intervals); cf. XIII.13.1.7..

V.2.3.5. The IVT has many applications. For example, we get a simple proof of III.1.12.2.: if $f(x) = x^n$, then $f(0) = 0$ and $f(1) = 1$, and f is continuous, so f takes all values between 0 and 1 on $[0, 1]$, i.e. every number between 0 and 1 has an n'th root between 0 and 1 (which is unique since f is strictly increasing). For numbers larger than 1, consider reciprocals.

We also get the following important algebraic fact:

V.2.3.6. THEOREM. Every polynomial of odd order with real coefficients has at least one real root.

PROOF: There is no loss of generality to consider monic polynomials. So let

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

be a monic polynomial with n odd and $a_k \in \mathbb{R}$. We have

$$\frac{f(x)}{x^n} = 1 + \frac{a_{n-1}}{x} + \cdots + \frac{a_0}{x^n}$$

for $x \neq 0$, so if

$$|x| > \max(1, n|a_0|, \ldots, n|a_{n-1}|)$$

we have $\frac{f(x)}{x^n} > 0$. Thus, if x is sufficiently large positive, $f(x) > 0$. And, since n is odd, $x^n < 0$ if $x < 0$, so $f(x) < 0$ if x is sufficiently large negative. Thus f takes both positive and negative values, so also takes the value 0 by the IVT since it is continuous. 👍

V.2.3.7. This can also be proved from the Fundamental Theorem of Algebra (XII.1.4.2.), since for polynomials with real coefficients the nonreal complex roots come in conjugate pairs, so there are an even number of nonreal roots. This argument gives the slightly stronger result that a polynomial of odd degree with real coefficients always has an odd number of real roots (counted with multiplicity); this can be proved directly from V.2.3.6. by dividing out a linear factor for each root. If an even number of factors is divided out, the resulting polynomial still has odd degree.

V.2.3.8. A similar argument can be efficiently used to locate real roots of a polynomial or any other continuous function: find a decreasing sequence of subintervals on which the function changes sign, say by successive bisection; the point of intersection of the intervals is a root.

V.2.4. The Max-Min Theorem

A second crucial result about continuous functions concerns attaining maximum and minimum values on closed bounded sets. The proof uses the Bolzano-Weierstrass Theorem (); in fact this result is also equivalent to the Completeness Axiom (Exercise ()).

V.2.4.1. Theorem. [Max-Min Theorem or Extreme Value Theorem] Let f be a continuous real-valued function on a closed bounded subset A of $\mathbb{R}$. Then f is bounded on A, and attains its maximum and minimum on A: there are $c, d \in A$ such that $f(d) \leq f(x) \leq f(c)$ for all $x \in A$.

Proof: We first show that f is bounded above. If not, then there is a sequence (x_n) in A with $f(x_n) \geq n$ for all n. By Bolzano-Weierstrass, there is a convergent subsequence (x_{k_n}) since A is bounded, say $x_{k_n} \to a$. We have $a \in A$ since A is closed. But then $f(x_{k_n}) \to f(a)$ by the Sequential Criterion, a contradiction since $(f(x_{k_n}))$ is unbounded (IV.1.3.16.).

Let $M = \sup_{x \in A} f(x)$. Then, for each $n \in \mathbb{N}$, there is an $x_n \in A$ with $f(x_n) > M - \frac{1}{n}$. Thus $f(x_n) \to M$. By B-W, there is a convergent subsequence (x_{k_n}). Set $c = \lim_{n\to\infty} x_{k_n}$. Then $c \in A$, and $f(x_{k_n}) \to f(c)$ by the Sequential Criterion. But $f(x_{k_n}) \to M$ also; thus $f(c) = M$.

The argument that f is bounded below and attains its minimum is almost identical. 👍

Note that this proof uses the Countable AC. (But the result does not require any form of choice; cf. XIII.11.1.8..) Also note that, as in the IVT, the c and d are not necessarily unique.

V.2.4.2. What is really going on here is compactness: a subset of $\mathbb{R}$ is compact if and only if it is closed and bounded (XIII.11.1.5.), and a continuous image of a compact set is compact (XIII.11.1.7.).

V.2.4.3. Examples. The result may fail if the set A is not closed or not bounded, or if f is not continuous on A.

(i) The function $f(x) = x$ is not bounded on $[0, +\infty)$. It is bounded on $A = [0, 1)$ but does not attain a maximum on A.

(ii) The function $g(x) = \frac{1}{x}$ is continuous on the bounded set $(0, 1]$ but is not bounded above. If we set

$$g(x) = \begin{cases} \frac{1}{x} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$$

then g is not bounded above on the closed bounded interval $[0,1]$ (of course, it is not continuous on $[0,1]$).
(iii) Set

$$f(x) = \begin{cases} x & \text{if} \quad x \neq 1 \\ 0 & \text{if} \quad x = 1 \end{cases}$$

Then f is bounded above on $[0,1]$, but does not attain a maximum on $[0,1]$ (again, it is not continuous on $[0,1]$).

V.2.5. Continuity and Oscillation

It is often fruitful to think of continuity of functions in terms of oscillations.

V.2.5.1. DEFINITION. Let f be a real-valued function defined on an interval I.
(i) If $\emptyset \neq A \subseteq I$, define the *oscillation of f on A* to be

$$\omega(f, A) = \sup\{|f(x) - f(y)| : x, y \in A\} \in [0, +\infty] \ .$$

(ii) If $x \in I$, define the *oscillation of f at x* to be

$$\omega(f, x) = \inf\{\omega(f, U) : U \text{ a neighborhood of } x\} \in [0, +\infty] \ .$$

The next proposition is obvious.

V.2.5.2. PROPOSITION. (i) If $\emptyset \neq A \subseteq B \subseteq I$, then $\omega(f, A) \leq \omega(f, B)$.
(ii) If $x \in I$, then $\omega(f, x) = \inf\{\omega(f, U) : U \text{ an open neighborhood of } x\}$.
(iii) f is continuous at $x \in I$ if and only if $\omega(f, x) = 0$.

Proof of (iii): If $\omega(f, x) = 0$, then for every $\epsilon > 0$ there is a neighborhood U of x such that $\omega(f, U) < \epsilon$. There is a $\delta > 0$ such that $(x - \delta, x + \delta) \subseteq U$. Then $\omega(f, (x - \delta, x + \delta)) < \epsilon$ by (i), so if $y \in (x - \delta, x + \delta)$, we have $|f(x) - f(y)| < \epsilon$.
Conversely, suppose f is continuous at x, and $\epsilon > 0$. There is a $\delta > 0$ such that if $y \in (x - \delta, x + \delta)$, $|f(x) - f(y)| < \frac{\epsilon}{2}$. Then, if $y, z \in (x - \delta, x + \delta)$,

$$|f(y) - f(z)| \leq |f(y) - f(x)| + |f(x) - f(z)| < \epsilon$$

so $\omega(f, (x - \delta, x + \delta)) \leq \epsilon$. Since $\epsilon > 0$ is arbitrary, $\omega(f, x) = 0$.

The next observation is important and useful:

V.2.5.3. PROPOSITION. Let f be a real-valued function on an open interval I, and $\epsilon > 0$. Then

$$\{x \in I : \omega(f, x) < \epsilon\}$$

is an open set.

PROOF: Call the set A. If $x \in A$, then there is an open neighborhood U of x with $\omega(f, U) < \epsilon$. If $y \in U$, then U is also an open neighborhood of y, so $\omega(f, y) \leq \omega(f, U) < \epsilon$. Thus $y \in A$, $U \subseteq A$. 👍

V.2.5.4. COROLLARY. Let f be a real-valued function on an interval I. Then the set of points where f is continuous is a G_δ (), i.e. the set of discontinuities of f is an F_σ.

PROOF: We may assume that I is open (why?) If A is the set of points where f is continuous, then

$$A = \{x \in I : \omega(f, x) = 0\} = \bigcap_{n=1}^{\infty} \left\{ x \in I : \omega(f, x) < \frac{1}{n} \right\} .$$

👍

V.3. Differentiation and Derivatives

> "The method of fluxions [derivatives] is probably one of the greatest, most subtle, and sublime discoveries of any age: it opens a new world to our view, and extends our knowledge, as it were, to infinity; carrying us beyond the bounds that seemed to have been prescribed to the human mind, at least infinitely beyond those to which the ancient geometry was confined."
>
> *Charles Hutton*[4]

> "The method of Fluxions is the general key by help whereof the modern mathematicians unlock the secrets of Geometry, and consequently of Nature. And, as it is that which hath enabled them so remarkably to outgo the ancients in discovering theorems and solving problems, the exercise and application thereof is become the main if not sole employment of all those who in this age pass for profound geometers."
>
> *George Berkeley*[5]

> "Every one versed in the matter will agree that even the elements of a scientific study of nature can be understood only by those who have a knowledge of at least the elements of the differential and integral calculus, as well as of analytic geometry – i.e. the lower part of the so-called higher mathematics."
>
> *F. Klein*[6]

> "Geometry may sometimes appear to take the lead over analysis, but in fact precedes it only as a servant goes before his master to clear the path and light him on his way. The interval between the two is as wide as between empiricism and science, as between the understanding and the reason, or as between the finite and the infinite."
>
> *J. J. Sylvester*[7]

The derivative of a function is of great importance, both theoretical and practical. It is the principal topic in any calculus course. The process of finding and using the derivative of a function is called *differentiation*.

[4] *A Philosophical and Mathematical Dictionary* (London, 1815); cf. [Mor93, 1911].
[5] *The Analyst*; cf. [Mor93, 1915].
[6] [Mor93, 1918].
[7] *Philosophic Magazine*, v. 31 (1866), p. 521; [Mor93, 1931].

V.3.1. The Derivative

The derivative of a function can be described in many ways: the instantaneous rate of change of the function, the slope of the tangent line to the graph of the function, and the best linear approximation to the function are three of the most common.

The Tangent Line to the Graph of a Function

V.3.1.1. There is an intuitive notion of "tangent line" to a curve at a point on the curve. Informally, the tangent line to the curve at the point P is the line which just "touches" the curve at P but does not cross it (indeed, the word "tangent" comes from the Latin *tangere*, which means "to touch"). This definition is adequate for some simple curves: for example, the tangent line to a circle at a point P is the line perpendicular to the diameter line of the circle through P (this is easily seen to be the only line through P which does not "cross" the curve). Indeed, this turns out to be the "generic" situation ().

V.3.1.2. But simple examples show that this intuitive definition is inadequate in general. For example, the x-axis is clearly the only reasonable candidate for the tangent line to the curve $y = x^3$ at the origin, but it "crosses" the curve there:

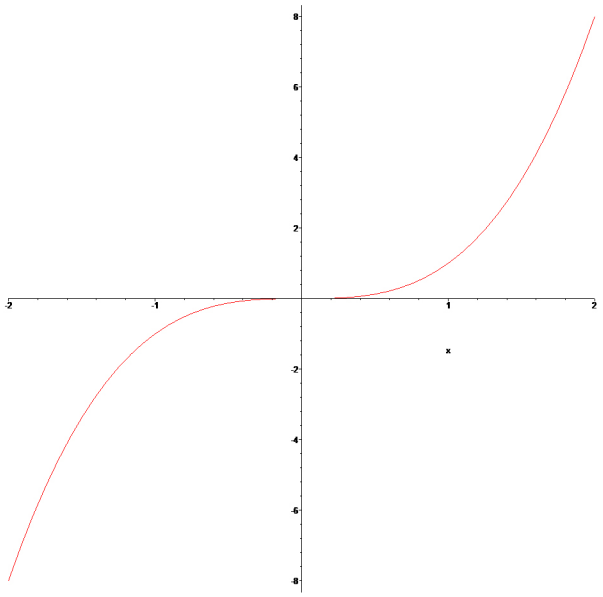


Figure V.3: Maple plot of graph of $f(x) = x^3$

V.3.1.3. For more complicated curves, such as the graph of $f(x) = x^2 \sin \frac{1}{x^2}$ (extended by setting $f(0) = 0$), this definition is even more inadequate (Figure V.4).

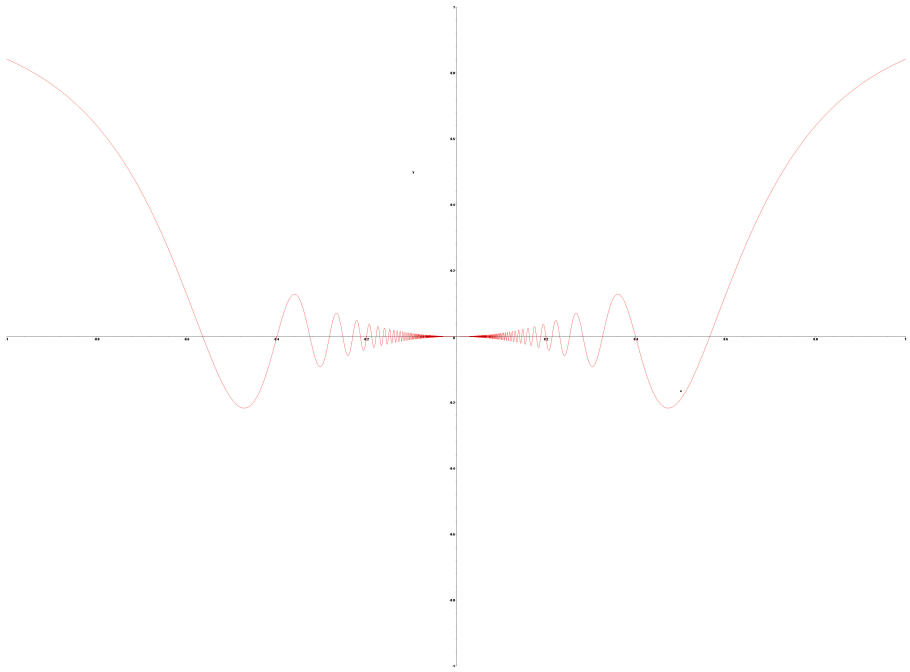


Figure V.4: Maple plot of graph of $f(x) = x^2 \sin\left(\frac{1}{x^2}\right)$

It is clear that the "tangent line" at the origin, if it exists, must again be the x-axis, but this line crosses the curve infinitely often in any interval around 0.

V.3.1.4. In addition, if there is a line through P which does not "cross" the curve, there may be more than one. For example, consider the curve $y = |x|$. Any line through the origin with slope strictly between -1 and 1 just "touches" the curve and does not cross it. Which, if any, of these lines should be considered the "tangent line"?

V.3.1.5. Finally, suppose the curve itself is a straight line. What could the "tangent line" be but the line itself? But how does this line just "touch" the original line?

The Difference Quotient

V.3.1.6. Mathematicians wrestled for a long time with the difficulties in making a proper definition of tangent line, and finally came up with a satisfactory dynamic definition, now recognized as a limit.

If we want to define the tangent line at P, we take an indirect approach. Suppose Q is another point on the curve. Then there is a unique line through P and Q, called the *secant line* to the curve through P and Q. (Note that, as with tangent lines, the term "secant line" has nothing to do with trigonometry; the word "secant" comes from the Latin *secare*, "to cut".) If P is held fixed and Q "approaches" P, the secant line should "approach" the tangent line at P.

V.3.1.7. To make this precise, suppose the curve is the graph of a function f, and that $P = (a, f(a))$. If $Q = (b, f(b))$, the line through P and Q has slope

$$\frac{f(b) - f(a)}{b - a}$$

as long as $P \neq Q$, i.e. $b \neq a$.

We want to regard a as fixed and let b vary, so we usually use a variable name like x or t instead of b.

V.3.1.8. Definition. Let f be a function defined on an open interval I around $a \in \mathbb{R}$. Then the *difference quotient* of f at a with variable x is

$$\frac{f(x) - f(a)}{x - a}$$

which is defined for $x \in I$, $x \neq a$.

Definition of the Derivative

We can then make the main definition:

V.3.1.9. Definition. Let f be defined in an open interval around $a \in \mathbb{R}$. Then f is *differentiable at a* if

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

exists. If it exists, it is called the *derivative* of f at a, denoted $f'(a)$.

Note that this is an *ordinary two-sided limit*, i.e. x can be either larger or smaller than a, and the limit must be finite, not $\pm\infty$.

Then we can make a careful definition of tangent line:

V.3.1.10. Definition. Let f be a function defined on an open interval I, and $a \in I$. If f is differentiable at a, then the *tangent line* to the graph $y = f(x)$ at a is the line through $(a, f(a))$ with slope $f'(a)$.

V.3.1.11. If f is not differentiable at a, we will generally say that the graph has no tangent line at $(a, f(a))$. (This may not always be accurate: the graph of a function can have a vertical tangent line at some points. We will discuss this more carefully in ().)

Differentiability vs. Continuity

We will see in the next few sections that "most" functions which are naturally encountered are differentiable wherever they are defined. However, there is one obvious situation which prevents differentiability: a discontinuity. Since a differentiable function is "approximately linear" near a, and linear functions are continuous, it is not surprising that differentiable functions are continuous:

V.3.1.12. PROPOSITION. Let f be a function defined on an open interval I, and $a \in I$. If f is differentiable at a, then f is continuous at a.

PROOF: If f is differentiable at a, then

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

exists. Since $\lim_{x \to a}(x - a) = 0$, we must also have $\lim_{x \to a}(f(x) - f(a)) = 0$ by (). But this is just the definition of continuity of f at a. 👍

V.3.1.13. However, continuity of f at a is not sufficient to give differentiability at a, as the following example shows. Thus differentiability at a is a strictly stronger condition than continuity.

V.3.1.14. EXAMPLE. Let $f(x) = |x|$. Then f is continuous at 0 (). But f is not differentiable at 0: the difference quotient

$$\frac{f(x) - f(0)}{x - 0} = \frac{|x|}{x}$$

is 1 if $x > 0$ and -1 if $x < 0$. Thus the limit as $x \to 0$ does not exist ().

Here is a dramatic example that shows that differentiability at a does not imply continuity anywhere except at a:

V.3.1.15. EXAMPLE. Let

$$f(x) = \begin{cases} x^2 & \text{if} \quad x \in \mathbb{Q} \\ 0 & \text{if} \quad x \notin \mathbb{Q} \end{cases} .$$

Then f is discontinuous everywhere except at 0 (). But f is differentiable at 0 since

$$\left| \frac{f(x) - f(0)}{x - 0} \right| = \left| \frac{f(x)}{x} \right| \leq |x|$$

for $x \neq 0$, and hence $f'(0)$ exists and equals 0 by the Squeeze Theorem ().

Geometric Interpretation of the Tangent Line

V.3.1.16. A somewhat informal but conceptually excellent way to interpret the tangent line to a curve is by magnification. If a curve has a tangent line at the point P, then when smaller and smaller pieces of the curve around P are magnified more and more, they become increasingly almost indistinguishable from the tangent line at P (see e.g. https://www.ima.umn.edu/~arnold/calculus/tangent/tangent-g.html, http://archives.math.utk.edu/visual.calculus/2/tangents.5/, or https://www.geogebra.org/material/simple/id/335793).

Conversely, if the curve becomes more and more indistinguishable from a straight line in small neighborhoods of P, then this line is the tangent line at P, and if the curve is the graph of a function and the tangent line is not vertical, its slope is the derivative of the function at that point.

Although this approach to the tangent line is not rigorous without additional work, it is much more precise and universally applicable than the idea of the tangent line just "touching" the curve at the point of tangency.

V.3.1.17. The magnification interpretation gives a satisfactory geometric explanation of why a function such as $f(x) = |x|$ is not differentiable. If an arbitrarily small segment of this graph around 0 is magnified arbitrarily much, it still looks exactly like the original curve and does not ever resemble a straight line segment.

V.3.1.18. Note, however, that even if a curve has a tangent line at a point P, and a small segment of the curve near P when magnified looks approximately like the tangent line, if an even smaller piece of the magnified segment *not* including the point P is further magnified, it does not need to bear any resemblance to the tangent line:

The magnification principle only applies to segments of the curve around P. In fact, the curve need not even have a tangent line at any other point than P, i.e. if it is the graph of a function, the function need not be differentiable except at P; it can even fail to be continuous except at P! For examples, see V.3.1.15. and V.3.9.6..

Rate of Change Interpretation

V.3.1.19. The difference quotient and derivative can be interpreted as rates of change. If f is a function defined on an interval containing a and b, then the *average rate of change of f from a to b* can be reasonably defined as the total change in f divided by the total change in x. The total change in f from a to b is $f(b) - f(a)$, and the total change in x from a to b is $b - a$. Thus the average rate of change in f from a to b is

$$\frac{f(b) - f(a)}{b - a}$$

provided $b \neq a$. (Here we could have $a < b$ or $a > b$.)

V.3.1.20. If we hold a fixed and let b vary, i.e. replace b by a variable x, then as $x \to a$ the average rate of change of f from a to x should approach the "instantaneous rate of change of f at a." Thus

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

can be interpreted as the *instantaneous rate of change of f at a*.

V.3.1.21. The most natural instance of this interpretation is if $f(t)$ denotes the position of an object at time t. Then

$$\frac{f(b)-f(a)}{b-a}$$

naturally denotes the *average velocity* of the object between $t=a$ and $t=b$, and

$$f'(a)=\lim_{t\to a}\frac{f(t)-f(a)}{t-a}$$

has a natural interpretation as the *instantaneous velocity* of the object at $t=a$.

An Alternate Approach: Overtaking

An alternate description of tangent lines and the derivative which does not use limits, at least not explicitly (it is more closely tied to the ancient Greek method of exhaustion), was nicely explained in [**?**].

The Derivative as a Function

Other Notations for the Derivative

V.3.1.22. The formula for the derivative $f'(a)$ in V.3.1.9. is frequently rewritten by writing $h=x-a$, so that $x=a+h$; then

$$f'(a)=\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}\ .$$

V.3.1.23. Another variation is that the $h=x-a$ is written Δx, thought of as the "change in x." Note that Δx is regarded as a single symbol and that the Δ and x have no independent meaning related to this symbol. Thus

$$f'(a)=\lim_{\Delta x\to 0}\frac{f(a+\Delta x)-f(a)}{\Delta x}\ .$$

This notation is more commonly used when the a is replaced by x, which is treated like a constant in the calculation of the limit, and then as a variable:

$$f'(x)=\lim_{\Delta x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}\ .$$

If another symbol such as t is used for the independent variable (e.g. when it is interpreted as time), the corresponding replacement is normally made in the change symbol:

$$f'(t)=\lim_{\Delta t\to 0}\frac{f(t+\Delta t)-f(t)}{\Delta t}\ .$$

V.3.1.24. Carrying this a step further, we often work with the equation $y=f(x)$ and define

$$\Delta y=\Delta f=f(x+\Delta x)-f(x)$$

where x and $x+\Delta x$ are in the domain of f. With this notation,

$$f'(x)=\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}=\lim_{\Delta x\to 0}\frac{\Delta f}{\Delta x}\ .$$

This notation nicely matches up with the "rate of change" interpretation: $\frac{\Delta y}{\Delta x}$ is the average rate of change of y with respect to x between x and $x+\Delta x$.

V.3.1.25. This leads to the notation originally used by LEIBNIZ in his development of calculus, not surprisingly called *Leibniz notation*. We use the symbol $\frac{dy}{dx}$ or $\frac{df}{dx}$ to denote the derivative:

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} \ .$$

Note that $\frac{dy}{dx}$ (pronounced "d-y-d-x") is regarded as a single symbol and is not a fraction, i.e. the dy and dx have no independent meaning (although meanings are sometimes attached to them in such a way as to make the quotient equal to $\frac{dy}{dx}$). The x and y can be replaced by any other variable names.

V.3.1.26. This notation arose because LEIBNIZ thought of dx as an "infinitesimal change" in x and dy the corresponding "infinitesimal change" in $y = f(x)$; the quotient $\frac{dy}{dx}$ should then (up to an "infinitesimal error") be the exact slope of the tangent line. (Actually functional notation did not exist at that time, and one thought only about independent and dependent variables related by some formula. We have also oversimplified what LEIBNIZ did; he was a very smart guy. NEWTON did something similar, but phrased it differently.) The exact meaning of the term "infinitesimal change" was unclear, and was the principal logical weak point of calculus until the nineteenth century when infinitesimals were effectively eliminated from the rigorous calculus developed then. They have reappeared in a rigorous way in "nonstandard analysis" (). Infinitesimals have never disappeared as a useful intuitive device in calculus and its applications, however.

> "The *difference quotient* $\frac{\Delta y}{\Delta x}$ absolutely must be formed with differences Δx which are not equal to 0. *After* the forming of this difference quotient we must imagine the passage to the limit carried out by means of a transformation or some other device. We have no right to suppose that *first* Δx goes through something like a limiting process and reaches a value which is infinitesimally small but still not 0, so that Δx and Δy are replaced by 'infinitely small quantities' or 'infinitesimals' dx and dy, and that the quotient of these quantities is then formed. Such a conception of the derivative is incompatible with the clarity of ideas demanded in mathematics, in fact, it is entirely meaningless. For a great many simple-minded people it undoubtedly has a certain charm, the charm of mystery which is always associated with the word 'infinite'; and in the early days of the differential calculus even Leibnitz [sic] himself was capable of combining these vague mystical ideas with a thoroughly clear understanding of the limiting process. It is true that this fog which hung round the foundations of the new science did not prevent Leibnitz or his great successors from finding the right path. But this does not release us from the duty of avoiding every such hazy idea in the building-up of the differential and integral calculus." [Emphasis original]
>
> *Richard Courant*[8]

V.3.1.27. In some ways, the Leibniz notation is inferior to the prime notation, mainly because in Leibniz notation it is not explicit where the derivative is being evaluated. However, for many purposes Leibniz notation is efficient and suggestive, so it continues to be commonly used. For example, if y is a function of t, then $\frac{dy}{dt}$ has a natural interpretation as the (instantaneous) rate of change of y with respect to t.

V.3.1.28. As a variation of Leibniz notation, we define an *operator* $\frac{d}{dx}$, which operates on (differentiable) functions and yields the derivative, i.e.

$$\frac{d}{dx}\left(f(x) \right) = f'(x) \ .$$

[8] [?, p. 121]. Note that this was written before the development of nonstandard analysis.

This notation is often convenient for expressing differentiation formulas; cf. V.3.2.1.. Again, $\frac{d}{dx}$ is regarded as a single symbol.

V.3.2. Elementary Rules of Differentiation

V.3.2.1. THEOREM. [POWER RULE, VERSION 1] Let $n \in \mathbb{N}$, and $f : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^n$. Then f is differentiable at any $a \in \mathbb{R}$, and

$$f'(a) = na^{n-1} \ .$$

As the Power Rule is often written,

$$\frac{d}{dx}\left(x^n\right) = nx^{n-1} \ .$$

PROOF: The following formula for any x and a in $\mathbb{R}$ can be quickly verified by multiplying out the right side and telescoping:

$$x^n - a^n = (x-a)\sum_{k=0}^{n-1} x^{n-1-k}a^k \ .$$

(This formula should technically be proved by induction.) Thus we have

$$f'(a) = \lim_{x\to a}\frac{f(x)-f(a)}{x-a} = \lim_{x\to a}\frac{x^n-a^n}{x-a} = \lim_{x\to a}\sum_{k=0}^{n-1} x^{n-1-k}a^k \ .$$

The last expression is a polynomial $p(x)$, hence by () the last limit exists and equals $p(a) = na^{n-1}$. 👍

V.3.2.2. This result will be extended to negative integer n in (), to fractional n in (), and to arbitrary real n in (). The same derivative formula holds in all cases.

The Power Rule can be proved for positive integer n more easily by induction using the Product Rule (cf. Exercise V.3.10.7.). It can also be proved using the Binomial Theorem (Exercise V.3.10.1.).

The Sum and Constant Multiple Rules

V.3.2.3. PROPOSITION. [SUM RULE] Let f and g be functions from an open interval I to $\mathbb{R}$, and $h = f + g : I \to \mathbb{R}$. If f and g are differentiable at $a \in I$, then h is differentiable at a and

$$h'(a) = f'(a) + g'(a) \ .$$

PROOF: By (), we have

$$\lim_{x\to a}\frac{[f(x)+g(x)]-[f(a)+g(a)]}{x-a} = \lim_{x\to a}\frac{[f(x)-f(a)]+[g(x)-g(a)]}{x-a}$$
$$= \lim_{x\to a}\left[\frac{f(x)-f(a)}{x-a}+\frac{g(x)-g(a)}{x-a}\right] = \lim_{x\to a}\frac{f(x)-f(a)}{x-a}+\lim_{x\to a}\frac{g(x)-g(a)}{x-a} \ .$$

👍

V.3.2.4. Proposition. [Constant Multiple Rule] Let f be a function from an open interval I to $\mathbb{R}$, $c \in \mathbb{R}$, and $h = cf : I \to \mathbb{R}$. If f is differentiable at $a \in I$, then h is differentiable at a and

$$h'(a) = cf'(a) \ .$$

Proof: By (), we have

$$\lim_{x \to a} \frac{cf(x) - cf(a)}{x - a} = \lim_{x \to a} \left[c \frac{f(x) - f(a)}{x - a} \right] = c \cdot \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \ .$$

V.3.2.5. Note that the Sum Rule and the Constant Multiple Rule are not merely formulas. They are (easy) theorems, with hypotheses and a conclusion with a theoretical part (that h is differentiable at a) and a formula part (the derivative of h at a).

V.3.3. The Product and Quotient Rules

V.3.3.1. It is not true that a derivative of a product of two functions is the product of the derivatives; almost any product where one is able to compute derivatives directly (e.g. a product of two power functions) is a counterexample. In fact, if it were true that the derivative of a product is the product of the derivatives, it would force the derivative of any differentiable function to be zero: if f is differentiable, we would have

$$f'(x) = \frac{d}{dx}\left(f(x)\right) = \frac{d}{dx}\left(1 \cdot f(x)\right) = \left[\frac{d}{dx}\left(1\right)\right]\left[\frac{d}{dx}\left(f(x)\right)\right] = 0 \cdot f'(x) = 0 \ .$$

Actually, the derivative of a *composition* of two functions is the product of the derivatives; see V.3.4.3. for a precise statement.

V.3.3.2. Instead, the derivative of a product of two differentiable functions is given by a more complicated formula with two terms, in each of which one of the factors is differentiated and the other left unchanged. Actually, as with the sum and constant multiple rules, the Product Rule is a theorem with hypotheses and a conclusion with a theoretical part (that a product of two differentiable functions is differentiable) and a formula for the derivative of the product:

V.3.3.3. Theorem. [Product Rule] Let f and g be functions from an open interval I to $\mathbb{R}$, and $h = fg : I \to \mathbb{R}$. If f and g are differentiable at $a \in I$, then h is differentiable at a and

$$h'(a) = f'(a)g(a) + f(a)g'(a) \ .$$

Proof: Write $u = f(a)$ and $v = g(a)$, and $y = h(a) = uv$. If $\Delta x \neq 0$, write $\Delta u = f(a + \Delta x) - f(a)$, $\Delta v = g(a + \Delta x) - g(a)$, $\Delta y = h(a + \Delta x) - h(a) = (u + \Delta u)(v + \Delta v) - uv$. Then we have

$$\frac{\Delta y}{\Delta x} = \frac{(u + \Delta u)(v + \Delta v) - uv}{\Delta x}$$

$$= \frac{v\cdot\Delta u + u\cdot\Delta v + \Delta u\cdot\Delta v}{\Delta x} = v\frac{\Delta u}{\Delta x} + u\frac{\Delta v}{\Delta x} + \Delta u\frac{\Delta v}{\Delta x}\ .$$

By assumption, $\lim_{\Delta x\to 0}\dfrac{\Delta u}{\Delta x}$ exists and equals $f'(a)$, and $\lim_{\Delta x\to 0}\dfrac{\Delta v}{\Delta x}$ exists and equals $g'(a)$. We also have $\lim_{\Delta x\to 0}\Delta u = 0$ since f is differentiable at a, hence continuous at a (VIII.3.1.14.). Thus, by () and (), since u and v are constants we have that $h'(a) = \lim_{\Delta x\to 0}\dfrac{\Delta y}{\Delta x}$ exists and equals

$$v\left[\lim_{\Delta x\to 0}\frac{\Delta u}{\Delta x}\right] + u\left[\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x}\right] + \left[\lim_{\Delta x\to 0}\Delta u\right]\left[\lim_{\Delta x\to 0}\frac{\Delta v}{\Delta x}\right] = f'(a)g(a) + f(a)g'(a) + 0\cdot g'(a)\ .$$



V.3.3.4. The above proof is a natural one, but requires introduction of some new notation. We can write essentially the same proof more compactly in straightforward notation, but when the proof is done in this way it involves the "trick" of adding and subtracting the same quantity, so does not seem as natural: PROOF: We have, for $x \neq a$,

$$\frac{h(x)-h(a)}{x-a} = \frac{f(x)g(x)-f(a)g(a)}{x-a} = \frac{[f(x)g(x)-f(a)g(x)]+[f(a)g(x)-f(a)g(a)]}{x-a}$$

$$= g(a)\frac{f(x)-f(a)}{x-a} + f(x)\frac{g(x)-g(a)}{x-a}\ .$$

We have that $\lim_{x\to a}\dfrac{f(x)-f(a)}{x-a}$ exists and equals $f'(a)$, and that $\lim_{x\to a}\dfrac{g(x)-g(a)}{x-a}$ exists and equals $g'(a)$. Also, $\lim_{x\to a} f(x)$ exists and equals $f(a)$ since f is differentiable at a, hence continuous at a. Thus by () and () we have that $h'(a) = \lim_{x\to a}\dfrac{h(x)-h(a)}{x-a}$ exists and equals

$$g(a)\left[\lim_{x\to a}\frac{f(x)-f(a)}{x-a}\right] + \left[\lim_{x\to a} f(x)\right]\left[\lim_{x\to a}\frac{g(x)-g(a)}{x-a}\right] = g(a)f'(a) + f(a)g'(a)\ .$$



V.3.3.5. COROLLARY. Let f and g be differentiable functions on an interval I. Then the product fg is also differentiable on I.

V.3.3.6. Now we turn to the Quotient Rule. Just as with products, the derivative of a quotient is not the quotient of the derivatives (try taking as denominator the constant function 1!) Again, the Quotient Rule is a theorem with hypotheses and a theoretical part to the conclusion as well as a formula for the derivative of the quotient. The one additional technical complication concerns the domain of the quotient function.

The quotient of the derivatives of numerator and denominator of a quotient of two functions does appear in l'Hôpital's Rule (), but there it does *not* represent the derivative of the quotient.

V.3.3.7. THEOREM. [QUOTIENT RULE] Let f and g be functions from an open interval I to $\mathbb{R}$. If f and g are differentiable at $a \in I$ and $g(a) \neq 0$, then $h = \frac{f}{g}$ is defined in an interval around a and differentiable at a, and

$$h'(a) = \frac{g(a)f'(a) - f(a)g'(a)}{[g(a)]^2} \ .$$

PROOF: Since g is differentiable at a and hence continuous at a, and $g(a) \neq 0$, there is an $\epsilon > 0$ such that, whenever $|x - a| < \epsilon$, $x \in I$ and $g(x) \neq 0$. Thus h is defined at least on the interval $(a - \epsilon, a + \epsilon)$.

Write $u = f(a)$ and $v = g(a)$, and $y = h(a) = \frac{u}{v}$. If $0 < |\Delta x| < \epsilon$, write $\Delta u = f(a + \Delta x) - f(a)$, $\Delta v = g(a + \Delta x) - g(a)$ (note that $v + \Delta v \neq 0$), $\Delta y = h(a + \Delta x) - h(a) = \frac{u+\Delta u}{v+\Delta v} - \frac{u}{v}$. Then, using a common denominator, we have

$$\frac{\Delta y}{\Delta x} = \frac{1}{\Delta x}\left[\frac{u + \Delta u}{v + \Delta v} - \frac{u}{v}\right]$$

$$= \frac{1}{\Delta x}\left[\frac{v(u + \Delta u) - u(v + \Delta v)}{v(v + \Delta v)}\right] = \frac{1}{\Delta x}\left[\frac{v \cdot \Delta u - u \cdot \Delta v}{v(v + \Delta v)}\right] = \frac{1}{v(v + \Delta v)}\left[v\frac{\Delta u}{\Delta x} - u\frac{\Delta v}{\Delta x}\right] \ .$$

By assumption, $\lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x}$ exists and equals $f'(a)$, and $\lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x}$ exists and equals $g'(a)$. We also have $\lim_{\Delta x \to 0} \Delta v = 0$ by continuity of g at a. Thus, by the limit laws, since u and v are constants we have that $h'(a) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$ exists and equals

$$\left(\lim_{\Delta x \to 0} \frac{1}{v(v + \Delta v)}\right)\left[v\left(\lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x}\right) - u\left(\lim_{\Delta x \to 0} \frac{\Delta v}{\Delta x}\right)\right] = \frac{1}{v^2}\left[g(a)f'(a) - f(a)g'(a)\right] \ .$$



V.3.3.8. COROLLARY. Let f and g be differentiable functions on an interval I, with $g(x) \neq 0$ for all $x \in I$. Then $\frac{f}{g}$ is differentiable on I.

V.3.3.9. COROLLARY. Every rational function () is differentiable on its domain.

V.3.3.10. The proof of V.3.3.7. can be rewritten in straightforward notation as in V.3.3.4.; cf. Exercise V.3.10.2.. We write this version only in a simplified special case, which can be combined with the Product Rule to give an alternate proof of the Quotient Rule (cf. V.3.10.3.):

V.3.3.11. Corollary. Let g be a function from an open interval I to $\mathbb{R}$ which is differentiable at $a \in I$, with $g(a) \neq 0$. Then $h = \frac{1}{g}$ is defined in an open interval around a and differentiable at a, and

$$h'(a) = -\frac{g'(a)}{[g(a)]^2] } \ .$$

Proof: This is the special case of V.3.3.7. with $f(x) = 1$. We give an alternative proof. Let $\epsilon > 0$ be as in the proof of V.3.3.7.; then h is defined at least on the interval $(a - \epsilon, a + \epsilon)$.
If $0 < |x - a| < \epsilon$, we have

$$\frac{h(x) - h(a)}{x - a} = \frac{1}{x - a}\left[\frac{1}{g(x)} - \frac{1}{g(a)}\right] = \frac{1}{x - a} \cdot \frac{g(a) - g(x)}{g(x)g(a)} = -\frac{1}{g(x)g(a)} \cdot \frac{g(x) - g(a)}{x - a} \ .$$

Since $\lim\limits_{x \to a} \dfrac{g(x) - g(a)}{x - a}$ exists and equals $g'(a)$, and $\lim_{x \to a} g(x) = g(a) \neq 0$ since g is continuous at a, we have that $h'(a) = \lim\limits_{x \to a} \dfrac{h(x) - h(a)}{x - a}$ exists and equals

$$-\left[\lim_{x \to a} \frac{1}{g(x)g(a)}\right]\left[\lim_{x \to a} \frac{g(x) - g(a)}{x - a}\right] = -\frac{1}{[g(a)]^2} g'(a) \ .$$

V.3.3.12. Corollary. [Power Rule, Version 2] Let n be a negative integer. Then the function $f(x) = x^n$ is differentiable on $\mathbb{R} \setminus \{0\}$, with derivative

$$f'(x) = nx^{n-1} \ .$$

Proof: Let $m = -n \in \mathbb{N}$, and let $g(x) = x^m$. Then $f(x) = \frac{1}{g(x)}$, so f is differentiable wherever g is nonzero and

$$f'(x) = -\frac{g'(x)}{[g(x)]^2} = -\frac{mx^{m-1}}{[x^m]^2} = -m\frac{x^{m-1}}{x^{2m}} = -mx^{m-1-2m} = nx^{n-1} \ .$$



V.3.3.13. **Caution:** The derivative of a power function is a multiple of a power function with exponent one smaller than the exponent in the original. But if the original exponent is negative, the exponent in the derivative is *more negative*. Thus, for example, 

$$\frac{d}{dx}(x^5) = 5x^4$$

but

$$\frac{d}{dx}(x^{-5}) = -5x^{-6} \ .$$

V.3.4. The Chain Rule

The Chain Rule is a result about differentiation of composite functions. As with the Product Rule and Quotient Rule, the Chain Rule is a theorem, not simply a formula; it has a theoretical part (that a composition of differentiable functions is differentiable) and a computational part (a formula for the derivative of the composite function).

V.3.4.1. The motivation and general idea of the Chain Rule is that if y is indirectly a function of x via an intermediate variable u, i.e. $u = f(x)$ and $y = g(u)$ for some functions f and g, and f and g are differentiable, we want to describe the rate of change $\frac{dy}{dx}$ in terms of $\frac{dy}{du} = g'(u)$ and $\frac{du}{dx} = f'(x)$. Roughly, if x is changed by a small amount Δx, then u is changed by an approximately proportional amount with constant of proportionality $f'(x)$, i.e. $\Delta u \approx f'(x) \cdot \Delta x$. Similarly, $\Delta y \approx g'(u) \cdot \Delta u$. So

$$\Delta y \approx g'(u) \cdot \Delta u \approx g'(u) f'(x) \cdot \Delta x \ .$$

Thus we should have that $y = g(f(x))$ is a differentiable function of x and

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} \ .$$

This formula is easily remembered by pretending that $\frac{dy}{du}$ and $\frac{du}{dx}$ are fractions (which they are not!) and "dividing out" the du's. This is a prime example of the utility of the Leibniz notation for derivatives, although this notation is somewhat imprecise as to exactly where the relevant functions are evaluated.

V.3.4.2. If f and g are linear functions, this all works out properly and quite simply. If $u = f(x) = mx+b$ and $y = g(u) = nu + c$, then $f'(x) = \frac{du}{dx}$ and $g'(u) = \frac{dy}{du}$ are the constant functions m and n respectively. If $h = g \circ f$, then $y = h(x) = g(f(x))$ is also a linear function with formula

$$y = h(x) = nf(x) + c = n(mx + b) + c = nmx + (nb + c)$$

and we have

$$\frac{dy}{dx} = h'(x) = nm = \frac{dy}{du}\frac{du}{dx} \ .$$

Thus the slope of a composition of two linear functions is the product of the slopes of the component functions.

The general Chain Rule asserts that if f and g are "approximately linear" (i.e. differentiable), then $g \circ f$ is also approximately linear, with slope equal to the product of the slopes of f and g. But one must be careful about where the derivatives are evaluated since they are not constant. Here is the precise statement.

V.3.4.3. Theorem. [Chain Rule] Let f be defined on a neighborhood $\mathcal{U}$ of $a \in \mathbb{R}$ and differentiable at a, with $f(a) = b \in \mathbb{R}$, and let g be defined on a neighborhood $\mathcal{V}$ of b and differentiable at b, with $f(\mathcal{U}) \subseteq \mathcal{V}$. Set $h = g \circ f$ on $\mathcal{U}$. Then h is differentiable at a, and

$$h'(a) = (g \circ f)'(a) = g'(b)f'(a) = g'(f(a))f'(a) \ .$$

A straightforward and essentially correct proof can be given which contains a subtle fallacy; see Exercise V.3.10.5.. A fully correct proof is not much harder if done right, but requires a slightly modified approach which is not so obvious. We first define a new function we will use in the proof:

V.3.4.4. DEFINITION. Let g and b be as in V.3.4.3.. Define a function G_b on $\mathcal{V}$ by

$$G_b(u) = \begin{cases} \dfrac{g(u)-g(b)}{u-b} & \text{if } u \neq b \\ \\ g'(b) & \text{if } u = b \end{cases} .$$

V.3.4.5. LEMMA. G_b is continuous at b.

PROOF: We need to check that $G_b(b) = \lim_{u\to b} G_b(u)$. But this is true by definition of $g'(b)$. 👍

V.3.4.6. LEMMA. Using the notation of V.3.4.3., if $x \in \mathcal{U}$, $x \neq a$, we have

$$\frac{h(x)-h(a)}{x-a} = G_b(f(x)) \cdot \frac{f(x)-f(a)}{x-a} .$$

PROOF: If $f(x) \neq b = f(a)$, we have

$$G_b(f(x)) \cdot \frac{f(x)-f(a)}{x-a} = \frac{g(f(x))-g(b)}{f(x)-b} \cdot \frac{f(x)-f(a)}{x-a}$$

and, since $b = f(a)$, $h(x) = g(f(x))$, and $h(a) = g(b)$, the formula holds. If $f(x) = b$, then both sides of the equation are zero, so the formula again holds. 👍

We now prove V.3.4.3..

PROOF: Since f is continuous at a by () and G_b is continuous at $b = f(a)$, $G_b \circ f$ is continuous at a by () and

$$\lim_{x\to a} G_b(f(x)) = G_b(f(a)) = G_b(b) = g'(b) .$$

Also, we have

$$\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = f'(a)$$

So by (), we have

$$\lim_{x\to a} \frac{h(x)-h(a)}{x-a} = \left[\lim_{x\to a} G_b(f(x))\right] \left[\lim_{x\to a} \frac{f(x)-f(a)}{x-a}\right] = g'(b)f'(a) .$$

👍

Here are some of the most important standard applications of the Chain Rule:

V.3.4.7. Corollary. [Chain Rule for Powers] Let f be differentiable at a, $f(a) > 0$, and $r \in \mathbb{R}$. Then $h(x) = [f(x)]^r$ is differentiable at a, and $h'(a) = r[f(a)]^{r-1} f'(a)$. Alternatively,

$$\frac{d}{dx}\left([f(x)]^r\right) = r[f(x)]^{r-1} f'(x) \ .$$

Proof: Apply V.3.4.3. with $g(u) = u^r$. 👍

V.3.4.8. Corollary. [Chain Rule for Exponentials] Let f be differentiable at a. Then $h(x) = e^{f(x)}$ is differentiable at a, and $h'(a) = e^{f(a)} f'(a)$. Alternatively,

$$\frac{d}{dx}\left(e^{f(x)}\right) = f'(x)e^{f(x)} \ .$$

V.3.4.9. Corollary. [Chain Rule for Logarithms] Let f be differentiable at a, with $f(a) \neq 0$. Then $h(x) = \log|f(x)|$ is differentiable at a, and $h'(a) = \frac{1}{f(a)} f'(a)$. Alternatively,

$$\frac{d}{dx}\left(\log|f(x)|\right) = \frac{f'(x)}{f(x)} \ .$$

The proofs of V.3.4.8. and V.3.4.9. are similar to the proof of V.3.4.7., and are left to the reader.

V.3.5. Derivatives of Inverse Functions

V.3.6. Critical Points and Local Extrema

One of the most important elementary applications of the derivative is the location and analysis of local extrema.

V.3.6.1. Definition. Let f be a function defined on an interval I, and a an interior point of I. The function f has a *local maximum*, or *relative maximum*, at a if there is a $\delta > 0$ such that $f(x) \leq f(a)$ for all x with $|x - a| < \delta$. The function f has a *local minimum*, or *relative minimum*, at a if there is a $\delta > 0$ such that $f(x) \geq f(a)$ for all x with $|x - a| < \delta$. A *local extremum*, or *relative extremum*, is a local maximum or local minimum.

V.3.6.2. A continuous function f on a closed bounded interval attains a maximum and minimum on the interval (). If the maximum is attained at an interior point a of the interval, then f has a local maximum at a. The converse is false: if f has a local maximum at a, its value there may not be the maximum value of f on the entire interval (it is only the maximum on some sufficiently small subinterval around a). Indeed, f may have two or more local maxima in the interval, where it may or may not attain the maximum on the whole interval, and the maximum on the interval can occur at an endpoint and not at a local maximum; in fact, f need not have any local maxima. Corresponding statements hold for local minima.

The crucial relation between local extrema and the derivative is:

V.3.6.3. THEOREM. Let f be a function defined on an interval I, and a an interior point of I. If f has a local extremum at a, and f is differentiable at a, then $f'(a) = 0$.

PROOF: We prove the statement for a local maximum; the proof for a local minimum is identical with a few inequalities reversed. Fix $\delta > 0$ such that $f(x) \leq f(a)$ for $|x - a| < \delta$. If $a < x < a + \delta$, then we have

$$\frac{f(x) - f(a)}{x - a} \leq 0$$

and thus by ()

$$\lim_{x \to a+} \frac{f(x) - f(a)}{x - a} \leq 0$$

(if the limit exists). Similarly, If $a - \delta < x < a$, then we have

$$\frac{f(x) - f(a)}{x - a} \geq 0$$

since numerator and denominator are both negative (nonpositive), and thus

$$\lim_{x \to a-} \frac{f(x) - f(a)}{x - a} \geq 0$$

(assuming the limit exists). So, if f is differentiable at a, both limits exist and are equal, hence 0; thus

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a+} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a-} \frac{f(x) - f(a)}{x - a} = 0 \ .$$



V.3.6.4. A continuous function need not be differentiable at a local extremum: consider $f(x) = |x|$ at 0 (cf. V.4.2.7.). And if f is differentiable at a and $f'(a) = 0$, then f need not have a local extremum at a: consider $f(x) = x^3$ at 0. However, the places where $f'(a) = 0$ (or does not exist) are the *candidates* for where f can have local extrema. Thus we make the following definition:



V.3.6.5. DEFINITION. Let f be a function defined on an interval I, and a an interior point of I. Then a is a *critical number* for f if either f is not differentiable at a, or f is differentiable at a and $f'(a) = 0$.

V.3.6.6. Thus every local extremum is at a critical number, but not conversely in general. The term *critical point* is often used as a synonym for *critical number*, but "critical point" is also sometimes used to refer to the corresponding point on the graph, so "critical number" is less ambiguous. Technically, the terms *local maximum*, etc., also refer to points on the graph; thus we say "f has a local extremum at a" rather than "a is a local extremum of f."

Increasing and Decreasing Functions at a Point

There is a sharper form of the contrapositive to V.3.6.3.. We first make a definition:

V.3.6.7. DEFINITION. Let f be a function defined on an interval I, and a an interior point of I. Then f is *increasing at* a if there is a $\delta > 0$ such that $f(x) < f(a)$ for $a-\delta < x < a$ and $f(x) > f(a)$ for $a < x < a+\delta$. Similarly, f is *decreasing at* a if there is a $\delta > 0$ such that $f(x) > f(a)$ for $a - \delta < x < a$ and $f(x) < f(a)$ for $a < x < a + \delta$.

Note that this is strictly a pointwise definition: we only compare $f(x)$ with $f(a)$ for x near a, we do not compare $f(x_1)$ and $f(x_2)$ for x_1, x_2 near a, which is a much more problematic matter (cf. VIII.5.7.2.).

V.3.6.8. COROLLARY. Let f be a function defined on an interval I, and a an interior point of I. Suppose f is differentiable at a. Then

(i) If $f'(a) > 0$, then f is increasing at a.

(ii) If $f'(a) < 0$, then f is decreasing at a.

PROOF: The proof of each part is essentially a piece of the proof of V.3.6.3.. For (i), suppose f is not increasing at a. Then for any $\delta > 0$ there is an x with $0 < |x - a| < \delta$ and

$$\frac{f(x) - f(a)}{x - a} \leq 0 \ .$$

Thus

$$f'(a) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \leq 0 \ .$$

The proof of (ii) is the same with inequalities reversed.

V.3.6.9. Thus, if f is differentiable on an interval I and $f'(x) > 0$ for all $x \in I$, then f is increasing at every point of I. It stands to reason that f should actually be strictly increasing on I, i.e. that $f(x_1) < f(x_2)$ if $x_1, x_2 \in I$, $x_1 < x_2$. This turns out to be true, but is subtle to prove (V.8.2.1., V.8.6.12.). A symmetric result holds if $f' < 0$ on I.

Classification of Critical Points

V.3.6.10. If a function f is continuous at a point a, increasing on an interval (r, a) for some $r < a$, and decreasing on an interval (a, s) for some $s > a$, it will have a local maximum at a. Thus, if f is differentiable on some deleted neighborhood of a, and $f' > 0$ on an interval to the left of a, and $f' < 0$ on an interval to the right of a, then f should have a local maximum at a. This also turns out to be true, but the proof is nontrivial (V.8.2.6.). Similarly, if f is decreasing on an interval (r, a) for some $r < a$, and increasing on an interval (a, s) for some $s > a$, it will have a local minimum at a. This is called the *First Derivative Test* for extrema. Note that the converse is not true in general, even if f is differentiable in an interval around a (V.3.10.10.).

Endpoint Extrema

We can analyze endpoint behavior in the same way.

V.3.6.11. DEFINITION. (i) Let f be a function defined on an interval with left endpoint a. Then f has an *endpoint maximum* at a if there is a $\delta > 0$ such that $f(x) \leq f(a)$ for all x with $a < x < a + \delta$. Similarly, f has an *endpoint minimum* at a if there is a $\delta > 0$ such that $f(x) \geq f(a)$ for $a < x < a + \delta$.

(ii) Let f be a function defined on an interval with right endpoint b. Then f has an *endpoint maximum* at b if there is a $\delta > 0$ such that $f(x) \leq f(b)$ for all x with $b - \delta < x < b$. Similarly, f has an *endpoint minimum* at b if there is a $\delta > 0$ such that $f(x) \geq f(b)$ for $b - \delta < x < b$.

An *endpoint extremum* is an endpoint maximum or endpoint minimum.

V.3.6.12. COROLLARY. Let f be a function on an interval I, and $[a, b]$ a closed bounded interval in the interior of I. Then

(i) If f has an endpoint maximum at a and f is differentiable at a, then $f'(a) \leq 0$.

(ii) If f has an endpoint minimum at a and f is differentiable at a, then $f'(a) \geq 0$.

(iii) If f has an endpoint maximum at b and f is differentiable at b, then $f'(b) \geq 0$.

(iv) If f has an endpoint minimum at b and f is differentiable at b, then $f'(b) \leq 0$.

The proof is again pieces of the proof of V.3.6.3..

V.3.7. Higher Derivatives

V.3.7.1. If f is a real-valued function defined on a subset A of $\mathbb{R}$, then the derivative f' is another function defined on a subset (proper subset in general) of A. Thus f' itself has a derivative on a subset of its domain, called the *second derivative* of f. The process can be repeated to define the third derivative and higher derivatives of all orders.

V.3.7.2. DEFINITION. Let $A \subseteq \mathbb{R}$, $a \in A$, and $f : A \to \mathbb{R}$ a function. If f' is differentiable at a, then f is *second-order differentiable*, or *twice differentiable*, at a and

$$f''(a) = (f')'(a) = \lim_{x \to a} \frac{f'(x) - f'(a)}{x - a}$$

is the *second derivative* of f at a, also denoted $f^{(2)}(a)$. (Also write $f^{(0)} = f$ and $f^{(1)} = f'$; f' is called the *first derivative* of f.)

Inductively, if $n \in \mathbb{N}$ and $f^{(n-1)}$ is differentiable at a, then f is *differentiable to order n* at a, or *n-times differentiable at a*, and $f^{(n)}(a) = (f^{(n-1)})'(a)$ is the *n'th derivative* of f at a.

If I is an interval contained in A, then f is *differentiable to order n* on I, or *n-times differentiable on I*, if $f^{(n)}(a)$ is defined at all $a \in I$, i.e. $f^{(n-1)}$ is differentiable on I. If f is n'th-order differentiable on I and $f^{(n)}$ is continuous on I, then f is $\mathcal{C}^{(n)}$ on I (*continuously n'th order differentiable on I*).

V.3.7.3. Note that for $f^{(n)}(a)$ to be defined, it is necessary for $f^{(n-1)}$ to be defined in an entire interval around a and continuous at a (VIII.3.1.14.). Thus if f is n'th order differentiable on I, then it is $\mathcal{C}^{(n-1)}$ on I (the converse is false in general). Note that "f is $\mathcal{C}^0$ on I" just means "f is continuous on I."

V.3.7.4. **Notation:** We can use additional primes to denote higher derivatives: f''' for the third derivative $f^{(3)}$, f'''' for $f^{(4)}$, etc. Three seems to be the conventional dividing line: the second derivative is usually denoted f'', the third derivative commonly denoted either f''' or $f^{(3)}$, and only rarely are primes used for $f^{(n)}$, $n > 3$.

Leibniz notation is also commonly used for higher derivatives: we write

$$\frac{d^2f}{dx^2} = \frac{d^2f}{dx^2}(x) = \frac{d^2}{dx^2}\big(f(x)\big) = \frac{d}{dx}\left(\frac{d}{dx}\big(f(x)\big)\right) = \left(\frac{d}{dx}\right)^2\big(f(x)\big) = f''(x)$$

and more generally

$$\frac{d^nf}{dx^n} = \frac{d^n}{dx^n}\big(f(x)\big) = f^{(n)}(x)$$

(note the position of the exponents: we write $\frac{d^nf}{dx^n}$, not $\frac{df^n}{dx^n}$ or $\frac{d^nf}{d^nx}$; think of $\frac{d^n}{dx^n} = \left(\frac{d}{dx}\right)^n$ as an operator). 

V.3.7.5. Second derivatives are especially important in physics: for example, acceleration is the second derivative of position. Higher derivatives beyond the second are not so commonly used in applications.

Geometric Interpretation

V.3.7.6. All the higher derivatives of a function f have geometric interpretations in the shape of the graph of f, but the interpretations of $f^{(n)}$ are increasingly subtle as n increases and are only easily described for $n = 2$ (V.8.3.).

The Formula of Faà di Bruno

V.3.7.7. Under the hypotheses of the Chain Rule V.3.4.3., if f and g are differentiable to order n at a and b respectively, it follows inductively from the Product Rule and Chain Rule that the composite function $g \circ f$ is also differentiable to n-th order at a. If $n = 2$, we have

$$(g \circ f)''(a) = g'(f(a))f''(a) + g''(f(a))[f'(a)]^2$$

and if $n = 3$,

$$(g \circ f)'''(a) = g'(f(a))f'''(a) + 3g''(f(a))f'(a)f''(a) + g'''(f(a))[f'(a)]^3 \ .$$

What is the general formula for $(g \circ f)^{(n)}$? It is a nontrivial exercise to even find a general expression, much less prove it. The most commonly written version of the formula is:

V.3.7.8. Theorem. [Faà di Bruno's Formula] Let f be defined on a neighborhood $\mathcal{U}$ of $a \in \mathbb{R}$ and n-times differentiable at a, with $f(a) = b \in \mathbb{R}$, and let g be defined on a neighborhood $\mathcal{V}$ of b and n-times differentiable at b, with $f(\mathcal{U}) \subseteq \mathcal{V}$. Set $h = g \circ f$ on $\mathcal{U}$. Then h is n-times differentiable at a, and

$$h^{(n)}(a) = (g \circ f)^{(n)}(a) = \sum \frac{n!}{k_1! \cdots k_n!} g^{(k)}(b) \left(\frac{f'(a)}{k_1!}\right)^{k_1} \cdots \left(\frac{f^{(n)}(a)}{k_n!}\right)^{k_n}$$

where the sum is over all sets of nonnegative integers $k_1, \ldots, k_n$ with $k_1 + 2k_2 + \cdots + nk_n = n$, and $k = k_1 + k_2 + \cdots + k_n$.

The formula is named after Francesco Faà di Bruno, a respected Italian mathematician of the mid-nineteenth century; however, the attribution is not historically accurate since he was not the first to prove or publish it (we may continue to use it just because we like the name). The formula is essentially combinatorial, and there are many forms and variations, arising in various contexts throughout mathematics. In principle it can be proved by induction, but this is very difficult in practice, and the standard proofs generally employ power series (cf. V.16.3.13.(iii)), Bell polynomials, or umbral calculus instead. The variations, proofs, and the complicated history of the formula are discussed in detail in [Joh02], which contains an extensive list of references.

See V.3.10.13. for a different version of the formula.

V.3.8. Continuous but Nondifferentiable Functions

It came as quite a shock to many mathematicians when in () Weierstrass gave an example of a function which is continuous everywhere but differentiable nowhere (Bolzano apparently had an earlier example, not published). There have since been many other explicit examples given (sometimes essentially the same examples have been rediscovered and republished), and it is now known (XIII.12.2.5.) that "most" continuous functions are not differentiable anywhere, quite contrary to our standard naive mental picture of a "typical" continuous function.

Van der Waerden's Example

V.3.8.1. Example. Here is one of the simplest examples of a continuous function which is not differentiable anywhere, due to B. L. van der Waerden [vdW30] (cf. [**?**,]). (The first example of this sort was given by T. Takagi in 1903 [**?**], and other similar examples were constructed by several other mathematicians.)

For each $k \in \mathbb{N} \cup \{0\}$, define $f_k : \mathbb{R} \to \mathbb{R}$ by letting $f_k(x)$ be the distance from x to the closest decadent rational number of the form $\frac{m}{10^k}$ for $m \in \mathbb{Z}$. See Figure V.5.

It is clear that f_k is continuous, periodic with period $\frac{1}{10^k}$, and that

$$0 \le f_k(x) \le \frac{1}{2 \cdot 10^k}$$

for all x. Then set

$$f(x) = \sum_{k=0}^{\infty} f_k(x) \ .$$

Since the series is uniformly convergent, f is continuous on all of $\mathbb{R}$. See Fig. V.6.

We claim that f is not differentiable anywhere. Fix $x \in \mathbb{R}$, and write

$$x = r.d_1 d_2 d_3 \cdots$$

in a decimal expansion. (If $x < 0$, take the decimal expansion of $x + s$ for some $s \in \mathbb{N}$, $s > |x|$ instead. The d_n are independent of s. There is the usual ambiguity if x is a decadent rational; in this case, either decimal expansion will do.) We will show that f is not differentiable at x.

For each $n \in \mathbb{N}$ set $x_n = x - \frac{1}{10^n}$ if d_n is 4 or 9, and $x_n = x + \frac{1}{10^n}$ otherwise. Then $x_n \to x$, so if f is differentiable at x we must have

$$\lim_{n\to\infty} \frac{f(x_n) - f(x)}{x_n - x} = f'(x) \in \mathbb{R}$$

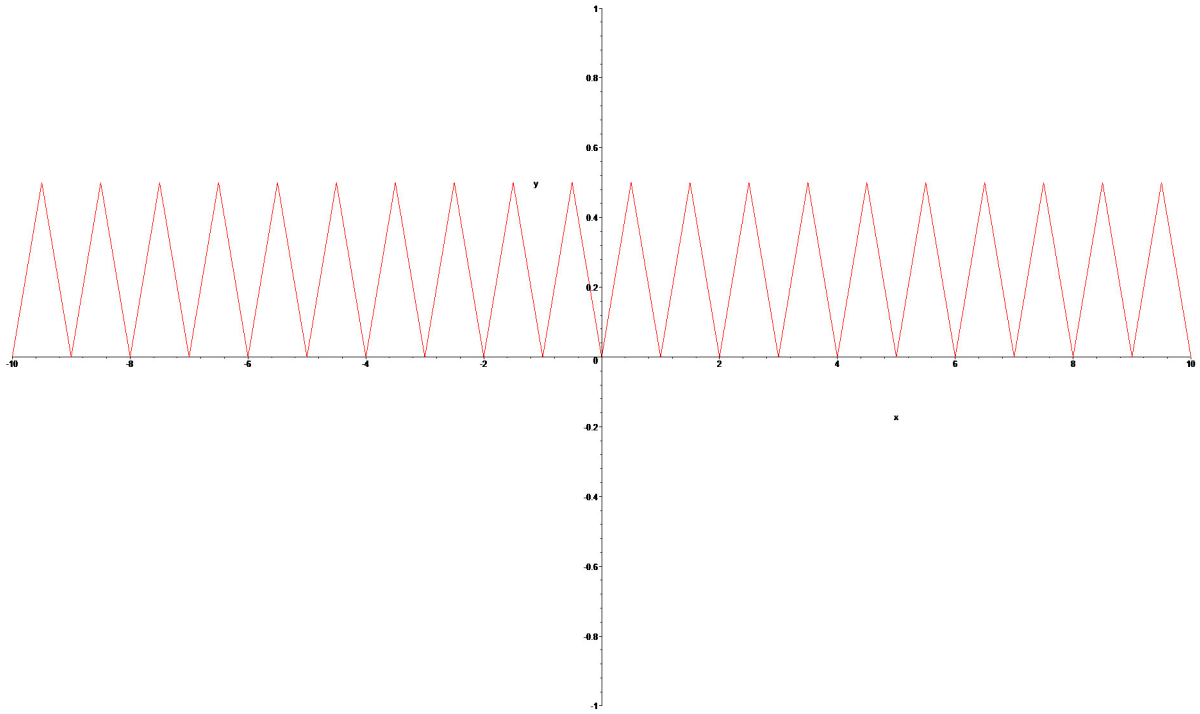


Figure V.5: Maple plot of graph of f_0

and thus the limit must exist (in the ordinary sense).

However, if $k < n$, we have that the closest number y of the form $\frac{m}{10^k}$ to x_n is also the closest one to x, and x and x_n are on the same side of y, so we have

$$|f_k(x_n) - f_k(x)| = |(x_n - y) - (x - y)| = |x_n - x| \ .$$

On the other hand, if $k \geq n$, then $\frac{1}{10^n} = |x_n - x|$ is an integer multiple of $\frac{1}{10^k}$ and $f_k(x_n) - f_k(x) = 0$ by periodicity of f_k. We thus have

$$f(x_n) - f(x) = \sum_{k=0}^{\infty} [f_k(x_n) - f_k(x)] = \sum_{k=0}^{n-1} [\pm(x_n - x)] = p_n(x_n - x)$$

where p_n is an integer of the same parity (odd or even) as n. Thus

$$\frac{f(x_n) - f(x)}{x_n - x} = p_n$$

cannot have a limit in $\mathbb{R}$ (i.e. be eventually constant) as $n \to \infty$.

V.3.8.2. Unlike Weierstrass's example, for this example there are uncountably many points where some or all of the Dini derivatives are finite (cf. Exercise V.3.10.14.), although such points form a set of measure zero by the Denjoy-Young-Saks Theorem.

V.3.9. The Bush-Wunderlich-Swift Example

We now describe a particularly simple example of a continuous nondifferentiable function. The example we give is from [Swi61], a special case of an example from [Bus52] and [Wun52].

V.3.9.1. Let x be a real number. Write the ternary (base 3) decimal expansion of the fractional part of x as

$$.t_1t_2t_3\cdots$$

where each t_k is 0, 1, or 2 (if x is a triadic rational number, there will be two such expansions, one ending in a string of 0's and one in a string of 2's). Define $f(x)$ to be the number in the unit interval with *binary* expansion

$$.d_1d_2d_3\cdots$$

($d_k = 0$ or 1) defined by the following inductive rule: $d_1 = 1$ if and only if $t_1 = 1$, and $d_{n+1} = d_n$ if and only if $t_{n+1} = t_n$.

V.3.9.2. Proposition. The function f is well defined.

Proof: The definition of f is via a definite procedure which gives a well-defined binary decimal for every ternary decimal. Thus we need only show that if x is a ternary rational, the two ternary decimal representations of x give binary decimals representing the same number. This is completely elementary, but a little ugly. Note that if a triadic expansion ends in a constant sequence, the corresponding dyadic expansion also ends in a constant sequence. Denote by $t_1t_2\cdots$ the ternary expansion of x ending in 2's, and $.t'_1t'_2\cdots$ the ternary expansion ending in 0's, and $.d_1d_2\cdots$ and $.d'_1d'_2\cdots$ the corresponding dyadic expansions. There are six cases to check. The first is

$$.t_1t_2\cdots = .\cdots 00222\cdots$$

$$.t'_1t'_2\cdots = .\cdots 01000\cdots$$

where the omitted initial segments are equal, and the displayed segments start with t_n and t'_n. Then either $d_n = d'_n = 0$, in which case

$$.d_1d_2\cdots = .\cdots 00111\cdots$$

$$.d'_1d'_2\cdots = .\cdots 01000\cdots$$

with the omitted initial segments equal, or $d_n = d'_n = 1$, in which case

$$.d_1d_2\cdots = .\cdots 11000\cdots$$

$$.d'_1d'_2\cdots = .\cdots 10111\cdots$$

and in either case the dyadic rational numbers represented are equal. The other five cases where $.t_1t_2\cdots$ is $.\cdots 10222\cdots$, $.\cdots 20222\cdots$, $.\cdots 01222\cdots$, $.\cdots 11222\cdots$, or $.\cdots 21222\cdots$ can be checked similarly. The cases $.0222\cdots$ and $.1222\cdots$ must be checked separately. 👍

V.3.9.3. PROPOSITION. The function f is continuous.

PROOF: Recall () that if $x_n, x \in \mathbb{R}$, then $x_n \to x$ if and only if, for each k there is an N such that, for all $n \geq N$, the triadic expansion of x_n (at least one expansion if x_n is triadic rational) agrees with the triadic expansion of x (at least one expansion if x is triadic rational) at least up to the k'th decimal place, and similarly for dyadic expansions. Since the first k decimal places of the dyadic expansion of $f(x)$ are completely determined by the first k decimal places of the triadic expansion of x, it follows that if $x_n \to x$, then $f(x_n) \to f(x)$. 👍

V.3.9.4. PROPOSITION. Let $x \in \mathbb{R}$. Then either $D^+f(x) = +\infty$ or $D_+f(x) = -\infty$, and either $D^-f(x) = +\infty$ or $D_-f(x) = -\infty$. In particular, f is not differentiable anywhere.

PROOF: Let $.t_1t_2\cdots$ be the ternary expansion for the fractional part of x not ending in a string of 2's, and let $.d_1d_2\cdots$ be the corresponding expansion of $f(x)$. There are infinitely many n such that $t_{n+1} \leq t_n$ and $t_{n+1} \neq 2$. Fix such an n. Let $t'_{n+1} = 2$ if $t_{n+1} = t_n < 2$ and $t'_{n+1} = t_n$ if $t_{n+1} < t_n$. Then successively choose t'_k for $k > n+1$ so that the number y with integer part the same as x and fractional part

$$.t_1 \cdots t_n t'_{n+1} t'_{n+2} \cdots$$

has image

$$f(y) = .d_1 \cdots d_n d'_{n+1} d_{n+2} \cdots$$

i.e. $d'_{n+1} \neq d_{n+1}$ but $d'_k = d_k$ for all $k \neq n+1$. We then have

$$0 < y - x \leq 3^{-n}$$

$$|f(y) - f(x)| = 2^{-(n+1)}$$

so

$$\left| \frac{f(y) - f(x)}{y - x} \right| \geq \frac{1}{2} \left(\frac{3}{2} \right)^n \quad .$$

Since there are infinitely many such n, the result for $D^+f(x)$ and $D_+f(x)$ follows. The proof for $D^-f(x)$ and $D_-f(x)$ is similar. 👍

V.3.9.5. The function takes values in $[0,1]$, and takes the value 0 at the integers; thus, if x is an integer, we must have $D^-f(x) \leq 0$ and $D_+f(x) \geq 0$, so $D^+f(x) = +\infty$ and $D_-f(x) = -\infty$. Similarly, $f\left(\frac{1}{2}\right) = 1$, so if x is $\frac{1}{2}$ or an integer translate, then $D_-f(x) \geq 0$ and $D^+f(x) \leq 0$, so $D^-f(x) = +\infty$ and $D_+f(x) = -\infty$.

Continuous Functions Differentiable at Only One Point

V.3.9.6. EXAMPLE. Let f be a continuous function from $\mathbb{R}$ to $\mathbb{R}$ which is not differentiable anywhere, e.g. one of the above examples. Set $g(x) = x^2 f(x)$. Then g is a continuous function from $\mathbb{R}$ to $\mathbb{R}$. We claim that g is differentiable at 0, and only at 0.

To see that g is differentiable at 0, note that f is bounded on the interval $[-1, 1]$ since it is continuous there; say $|f(x)| \leq M$ for $|x| \leq 1$. Then, for $0 < |x| \leq 1$, we have

$$\left| \frac{g(x) - g(0)}{x - 0} \right| = \left| \frac{x^2 f(x)}{x} \right| \leq M|x|$$

and thus $g'(0)$ exists and equals 0 by the Squeeze Theorem.

But g is not differentiable anywhere except at 0, since if it were $f(x) = \frac{g(x)}{x^2}$ would also be differentiable there by the Quotient Rule.

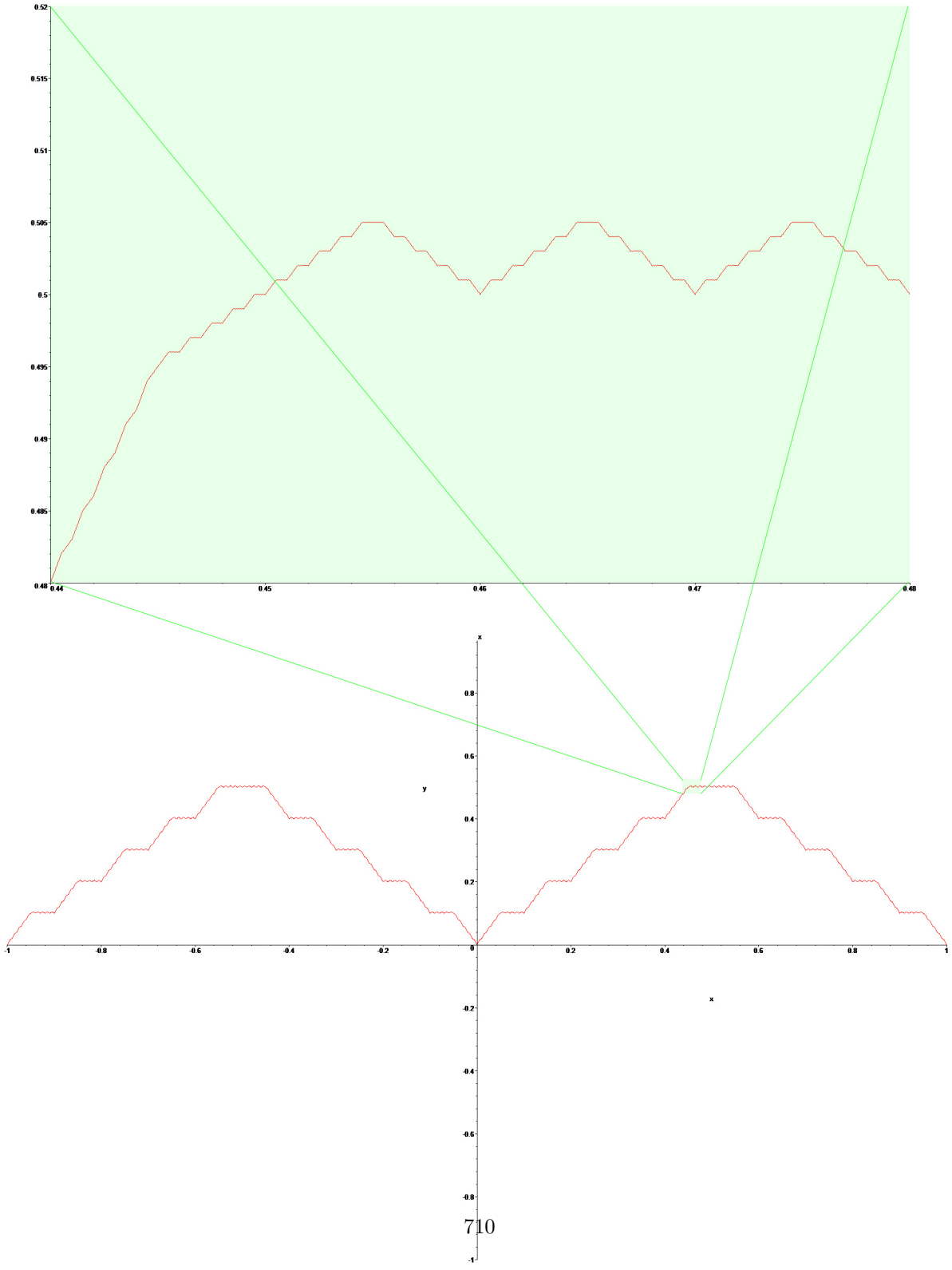


Figure V.6: Maple plot of graph of f

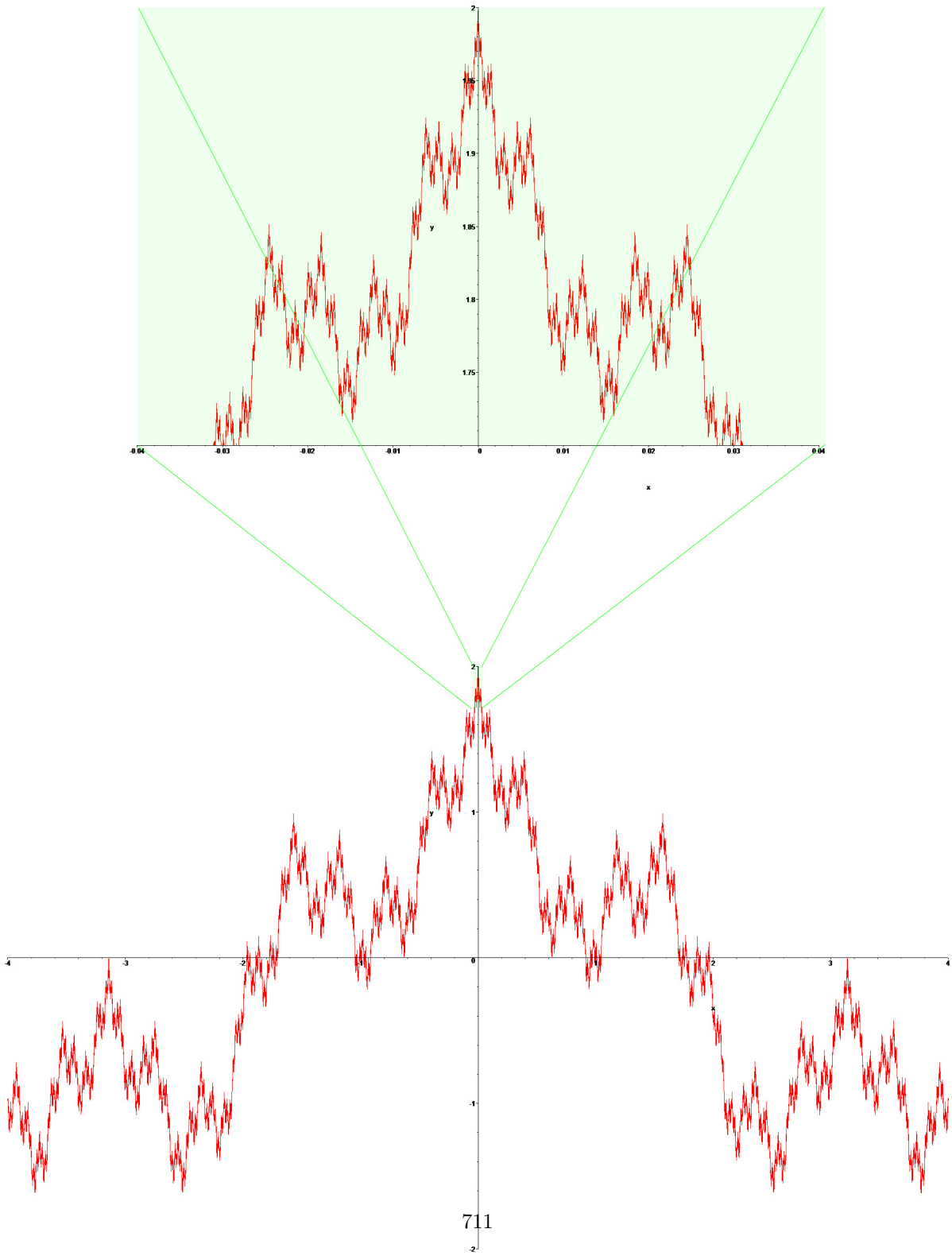


Figure V.7: Maple plot of $f(x) = \sum_{k=1}^{\infty} 2^{-k} \cos(4^k x)$

V.3.10. Exercises

V.3.10.1. Write out the proof of the power rule by writing, for $f(x) = x^n$ ($n \in \mathbb{N}$) and $a \in \mathbb{R}$,

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} = \lim_{h\to 0} \frac{(a+h)^n - a^n}{h}$$

and expanding $(a+h)^n$ using the Binomial Theorem (V.12.5.3.).

V.3.10.2. Write the proof of the full Quotient Rule along the lines of the proofs of V.3.3.4. and V.3.3.11..

V.3.10.3. Here is an alternate derivation of the quotient formula and "proof" of the Quotient Rule which has appeared in some textbooks.

Let f and g be defined in a neighborhood of a and differentiable at a, with $g(a) \neq 0$.
(a) Set $h = \frac{f}{g}$. Then, since g is continuous at a and $g(a) \neq 0$, h is defined in a neighborhood of a, and $f = gh$ on this neighborhood.
(b) By the Product Rule, $f'(a) = g'(a)h(a) + g(a)h'(a)$.
(c) Solve this equation for $h'(a)$ to obtain the Quotient Rule formula.
(d) What is wrong with this argument as a proof of the Quotient Rule? What does this argument actually prove?

V.3.10.4. (a) Show by direct calculation that if $a \neq 0$, then $f(x) = \frac{1}{x}$ is differentiable at a and $f'(a) = -\frac{1}{a^2}$.
(b) Use (a) and the Chain Rule to give an alternate proof of V.3.3.11..
(c) Conclude that the Quotient Rule is a consequence of (a), the Product Rule, and the Chain Rule.

V.3.10.5. What is wrong with the following "proof" of the Chain Rule? [Hint: the problem is very similar to the problem in ().]

Let f be defined on a neighborhood $\mathcal{U}$ of a and differentiable at a, with $f(a) = b$, and let g be defined on a neighborhood $\mathcal{V}$ of b and differentiable at b, with $f(\mathcal{U}) \subseteq \mathcal{V}$. Set $h = g \circ f$ on $\mathcal{U}$.
(a) For $x \in \mathcal{U}$, $x \neq a$, write

$$\frac{h(x) - h(a)}{x - a} = \frac{g(f(x)) - g(f(a))}{f(x) - f(a)} \cdot \frac{f(x) - f(a)}{x - a}$$

(b) For $x \in \mathcal{U}$, write $u = f(x)$. Since $\lim_{x\to a} \frac{f(x)-f(a)}{x-a}$ exists and $\lim_{x\to a}(x-a) = 0$, we have $\lim_{x\to a}(u-b) = 0$.
(c) Combining (a) and (b), we have

$$\lim_{x\to a} \frac{h(x) - h(a)}{x - a} = \left[\lim_{u\to b} \frac{g(u) - g(b)}{u - b}\right] \left[\lim_{x\to a} \frac{f(x) - f(a)}{x - a}\right] = g'(b)f'(a) \ .$$

V.3.10.6. What is wrong with the following "proof" of the Inverse Function Rule? What does this argument actually prove?

Let f be continuous and strictly monotone in a neighborhood of a, and differentiable at a. If $f(a) = b$, then $g = f^{-1}$ is defined, continuous, and strictly monotone in a neighborhood of b by ().
(a) Set $h = g \circ f = f^{-1} \circ f$. Then $h(x) = x$ in a neighborhood of a. Thus h is differentiable at a, and $h'(a) = 1$.
(b) Using the Chain Rule, differentiate h at a to obtain

$$1 = h'(a) = g'(f(a))f'(a) = (f^{-1})'(b)f'(a) \ .$$

(c) If $f'(a) \neq 0$, divide both ends of the equation by $f'(a)$ to obtain

$$(f^{-1})'(b) = \frac{1}{f'(a)} \ .$$

V.3.10.7. Prove the formula $\frac{d}{dx}\left([f(x)]^n\right) = n[f(x)]^{n-1}f'(x)$ for $n \in \mathbb{N}$ by induction using the Product Rule instead of the Chain Rule.

V.3.10.8. (S. Chang[9]) Let $f(x) = \sum_{k=1}^{100} \left[\frac{\sin(2\pi k^2 x)}{4\pi^2 k^5} + \frac{x^2}{2k}\right]$. Note that f is $\mathcal{C}^\infty$ (in fact, analytic) on $\mathbb{R}$. Use a computer algebra system to plot f, f', and f''. Could you have predicted the behavior of f'' from the appearance of the graph of f, or even the graph of f'?

V.3.10.9. (a) Formally define the "derivative" of a polynomial $p(x) = \sum_{k=0}^{n} a_k x^k$ to be

$$p^{\#}(x) = \sum_{k=1}^{n} k a_k x^{k-1} \ .$$

Give a purely algebraic proof that $p \mapsto p^{\#}$ is linear and satisfies the "Product Rule"

$$(pq)^{\#} = p^{\#}q + pq^{\#}$$

for polynomials p and q. (This argument is valid over any field or even any ring.)
(b) Extend the definition of # to the field of rational functions over $\mathbb{R}$ by

$$\left(\frac{p}{q}\right)^{\#} = \frac{qp^{\#} - pq^{\#}}{q^2}$$

for polynomials p and q. Show that the extended # is well defined and also satisfies the product rule, and the quotient rule. See XII.12.2.2..
(c) If f is a polynomial of degree n, expand $f(x+t)$ to obtain

$$f(x+t) = f_0(x) + f_1(x)t + f_2(x)t^2 + \cdots + f_n(x)t^n$$

[9] *American Mathematical Monthly*, v. 119 (October 2012), p. 669.

for polynomials $f_0, f_1, \ldots, f_n$. Show that $f_0 = f$ and $f_1 = f^\#$, and more generally $f_k = \frac{1}{k!} f^{\#\#\cdots\#}$, where the "derivative" is taken k times. This is an algebraic version of Taylor's Theorem (V.10.1.7.).

(d) Let $f(x)$ be a polynomial. Define a polynomial g in two variables by $g(x, y) = f(y) - f(x)$. Show that g is divisible by $y - x$ and if

$$h(x, y) = \frac{g(x, y)}{y - x} = \frac{f(y) - f(x)}{y - x}$$

then $f^\#(x) = h(x, x)$.

The map $\#$ is a derivation (XII.12.1.1.). Compare with III.5.3.7..

V.3.10.10. Let $f(x) = 2x^2 + x^2 \sin \frac{1}{x^2}$ for $x \neq 0$, $f(0) = 0$.

(a) Show that f is differentiable everywhere.

(b) Show that f has a strict local minimum (actually absolute minimum) at 0.

(c) Show that the First Derivative Test (V.3.6.10., V.8.2.6.) fails at 0.

V.3.10.11. There is a unique differentiable function f with $f'(x) = \frac{1}{2}x + x^2 \sin \frac{1}{x^2}$ for $x \neq 0$ and $f(0) = f'(0) = 0$ (V.8.5.9.).

(a) Show that by the first derivative test, f has a local minimum at 0 (which is in fact a strict absolute minimum).

(b) In what sense is the graph of f concave upward around 0?

V.3.10.12. [Gou59] Let f and g be differentiable on an open interval I, and suppose that $f'g = fg'$ on I.

(a) Can we conclude from the quotient rule applied to $\frac{f}{g}$ that $\frac{f}{g}$ is constant on I, i.e. that f is a constant multiple of g on I?

(b) Consider the case where $I = \mathbb{R}$, $g(x) = x^2$ for all x, $f(x) = x^2$ if $x \leq 0$, and $f(x) = 2x^2$ for $x > 0$.

(c) Suppose we assume f and g are $\mathcal{C}^\infty$? [Consider Cauchy's examples V.17.2.3., especially the last two and their reflections.]

V.3.10.13. [Gou59, p. 32] Prove the following theorem:

Theorem. Let f be defined on a neighborhood $\mathcal{U}$ of $a \in \mathbb{R}$ and n-times differentiable at a, with $f(a) = b \in \mathbb{R}$, and let g be defined on a neighborhood $\mathcal{V}$ of b and n-times differentiable at b, with $f(\mathcal{U}) \subseteq \mathcal{V}$. Set $h = g \circ f$ on $\mathcal{U}$. Then h is n-times differentiable at a, and

$$h^{(n)}(a) = \sum_{k=1}^{n} A_{n,k} g^{(k)}(b)$$

where

$$A_{n,k} = \frac{1}{k!} \sum_{j=0}^{k-1} (-1)^j \binom{k}{j} [f(a)]^j \left[\frac{d^n}{dx^n} \left([f(x)]^{k-j} \right) \right] (a) \ .$$

(a) Show that $h^{(n)}(a)$ is of the form $\sum_{k=1}^{n} A_{n,k} g^{(k)}(b)$, where the $A_{n,k}$ are independent of the form of g.

(b) To find the $A_{n,k}$, set $g(u) = u^m$ for $1 \leq m \leq n$ and solve a system of equations. (Try $n = 2$ first.)

(c) Show that this formula is equivalent to Faà di Bruno's formula (V.3.7.8.).

V.3.10.14. Let f be VAN DER WAERDEN'S example (V.3.8.1.).

(a) Let a be a real number whose decimal part is $.d_1d_2d_3\cdots$, where d_n is in the range $0-4$ for n odd and in the range $5-9$ for n even. Show that $D_-f(a) = D_+f(a) = 0$ and $D^-f(a) = D^+f(a) = 1$.

(b) Show that if $m, n \in \mathbb{Z}$, $m < n$, there are uncountably many points a in $[0,1]$ with $D_-f(a) = D_+f(a) = m$ and $D^-f(a) = D^+f(a) = n$. [If $n > 0$, take the first n decimal places in the range $0-4$, the next $n-m$ in the range $5-9$, etc.]

(c) Extend the argument of b to cover the cases where m and/or n is $\pm\infty$. (Both can be $+\infty$ or both $-\infty$, or we can have $m = -\infty$ and $n = +\infty$.)

(d) For how many points a is $D_-f(a) \neq D_+f(a)$ and/or $D^-f(a) \neq D^+f(a)$? What are the possibilities?

V.3.10.15. Let I be an interval in $\mathbb{R}$, and $f : I \to \mathbb{R}$ a one-to-one continuous function.

(a) Show that f cannot have a local maximum or local minimum in the interior of I. [If $a, b, c \in I$, $a < b < c$, and $f(a) < f(b)$, $f(c) < f(b)$, use the Intermediate Value Theorem to show there are x_1, x_2 with $a < x_1 < b < x_2 < c$ and $f(x_1) = f(x_2)$.]

(b) For every $a, b \in I$, $a < b$, f must attain its maximum on $[a,b]$ at a or b, and its minimum at the other.

(c) If there are $a, b, c, d \in I$, $a < b < c < d$, such that f attains its maximum on $[a,b]$ at b and its maximum on $[c,d]$ at c, then f has a local maximum at either b or c, and similarly for minima.

(d) Conclude that f must be either strictly increasing on I or strictly decreasing on I.

V.3.11. Calculus of Finite Differences

> "The analysis of the infinite ... is nothing but a special case of the method of differences, ... wherein the differences are infinitely small ..."
>
> *L. Euler*[10]

> "It appears that Fermat, the true inventor of differential calculus [sic], considered calculus as derived from the calculus of finite differences by neglecting infinitesimals of higher orders as compared with those of a lower order ..."
>
> *Laplace*[11]

In this section, we discuss a discrete version of derivatives. This procedure is important historically as a motivation of much of differential and integral calculus (cf. e.g. [Eul00]), and is still very useful in numerical analysis.

V.3.11.1. We start with a functional relation $y = f(x)$, where f is a function defined, say, on all of $\mathbb{R}$ (simply to make the discussion simpler by not having to worry about domain questions). Fix an $x_0 \in \mathbb{R}$, and an "increment" $h > 0$ (h is often written Δx). Form the sequence

$$x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, \ldots, x_n = x_0 + nh, \ldots$$

which is an arithmetic sequence whose successive differences are each equal to $h = \Delta x$. Set

$$\Delta x_n = x_{n+1} - x_n$$

for each n; of course, $\Delta x_n = h$ for all n since the x_n were chosen evenly spaced.

V.3.11.2. Now form the sequence

$$y_0, y_1, y_2, \ldots$$

where $y_n = f(x_n)$ for each n. This sequence is rarely an arithmetic sequence (it will be arithmetic if f is linear or affine); its properties are closely connected to the properties of the function f. Set

$$\Delta y_n = y_{n+1} - y_n = f(x + (n+1)h) - f(x + nh)$$

for each n. The Δy_n are called the *sequence of first differences* of y (or f) corresponding to the increment $h = \Delta x$. This sequence is rarely constant (it will be constant if f is linear or affine); its properties are closely connected to the properties of the function f. If f is a quadratic polynomial, the first differences will form an arithmetic progression.

V.3.11.3. Go on to form the *second differences*

$$\Delta^2 y_n = \Delta y_{n+1} - \Delta y_n$$

for each n. The second differences are all zero if and only if the first differences are constant, e.g. if f is affine. If f is a quadratic polynomial, the second differences will be constant.

[10] [Eul00, p. 64].
[11] *Lés Intégrales Définies*; cf. [Mor93, 1902].

V.3.11.4. Successively form *higher differences*

$$\Delta^{k+1} y_n = \Delta^k y_{n+1} - \Delta^k y_n$$

for each k and n. The sequence of $(k+1)$'st differences of (y_n) is the sequence of first differences of $(\Delta^k y_n)$. If f is a polynomial of degree $\leq k$, then the sequence of k'th differences will be constant, so the sequence of $(k+1)$'st differences will be 0.

V.3.11.5. The successive differences can be conveniently displayed in a table

$$\begin{array}{ccccccccc}
y_0 & & y_1 & & y_2 & & y_3 & & \cdots \\
& \Delta y_0 & & \Delta y_1 & & \Delta y_2 & & \Delta y_3 & \cdots \\
& & \Delta^2 y_0 & & \Delta^2 y_1 & & \Delta^2 y_2 & & \cdots \\
& & & \Delta^3 y_0 & & \Delta^3 y_1 & & \Delta^3 y_2 & \cdots \\
& & & & \ddots & & \ddots & & \ddots
\end{array}$$

where the first row is computed from the function f and each other entry is the difference of the two entries just above it.

V.3.11.6. We have

$$\Delta y_0 = y_1 - y_0$$
$$\Delta^2 y_0 = (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 - y_0$$
$$\Delta^3 y_0 = y_3 - 3y_2 + 3y_1 - y_0$$

and, for any n,

$$\Delta^n y_0 = \sum_{k=0}^{n} (-1)^k \binom{n}{k} y_{n-k}$$

and, more generally, for any n and m,

$$\Delta^n y_m = \sum_{k=0}^{n} (-1)^k \binom{n}{k} y_{n+m-k}$$

(these formulas can easily be proved by induction). Solving for y_n, we obtain

$$y_n = \sum_{k=0}^{n} \binom{n}{k} \Delta^k y_0$$

and, more generally, for any n and m,

$$y_{n+m} = \sum_{k=0}^{n} \binom{n}{k} \Delta^k y_m$$

(again a straightforward proof by induction).

V.3.11.7. Thus, for any f, x_0, and $n \in \mathbb{N}$,

$$f(x_0 + n\Delta x) = \sum_{k=0}^{n} \binom{n}{k} \Delta^k f$$

where we write $\Delta^k f$ for $\Delta^k f_0$. Thus, if f is defined on the interval $[x_0, x]$, and we take $n \in \mathbb{N}$ and $\Delta x = \frac{x-x_0}{n}$, so $x = x_0 + n\Delta x$ and $n = \frac{x-x_0}{\Delta x}$, we have

$$f(x) = \sum_{k=0}^{n} \frac{n(n-1)\cdots(n-k+1)}{k!} \Delta^k f \approx \sum_{k=0}^{m} \frac{(x-x_0)^k}{k!} \frac{\Delta^k f}{(\Delta x)^k}$$

if m is large and n is much larger, if $\frac{\Delta^k f}{k!(\Delta x)^k} \to 0$ fairly rapidly as $k \to \infty$ (see () for a precise version).

V.3.11.8. The sequence of k'th differences is, roughly speaking, a discrete version of the k'th derivative $f^{(k)}$ (although the differences are defined for any f whether or not it is differentiable or even continuous).

V.3.11.9. EXAMPLES. (i) suppose $f(x) = mx$. Then the table becomes

$$\begin{array}{ccccccccc}
mx_0 & & mx_1 & & mx_2 & & mx_3 & & \cdots \\
& m\Delta x & & m\Delta x & & m\Delta x & & m\Delta x & \cdots \\
& & 0 & & 0 & & 0 & & \cdots \\
& & & 0 & & 0 & & 0 & \cdots \\
& & & & \ddots & & \ddots & & \ddots
\end{array}$$

since $mx_{n+1} - mx_n = m\Delta x$ for all n.
(ii) Let $f(x) = x^2$. Then the table becomes

$$\begin{array}{ccccccccccc}
x_0^2 & & x_1^2 & & x_2^2 & & x_3^2 & & & & \cdots \\
& (x_0+x_1)\Delta x & & (x_1+x_2)\Delta x & & (x_2+x_3)\Delta x & & (x_3+x_4)\Delta x & & & \cdots \\
& & 2(\Delta x)^2 & & 2(\Delta x)^2 & & 2(\Delta x)^2 & & & & \cdots \\
& & & 0 & & 0 & & 0 & & & \cdots \\
& & & & \ddots & & \ddots & & & & \ddots
\end{array}$$

since $x_{n+1}^2 - x_n^2 = (x_n + x_{n+1})(x_{n+1} - x_n) = (x_n + x_{n+1})\Delta x$ for all n. Since $x_{n+1} = x_n + \Delta x$, we have

$\Delta y_n = 2x_n\Delta x + (\Delta x)^2$, so the table becomes

$$\begin{array}{ccccccc}
x_0^2 & & x_1^2 & & x_2^2 & & x_3^2 & & \cdots \\
& 2x_0\Delta x + (\Delta x)^2 & & 2x_1\Delta x + (\Delta x)^2 & & 2x_2\Delta x + (\Delta x)^2 & & 2x_3\Delta x + (\Delta x)^2 & \cdots \\
& & 2(\Delta x)^2 & & 2(\Delta x)^2 & & 2(\Delta x)^2 & & \cdots \\
& & & 0 & & 0 & & 0 & \cdots \\
& & & & \ddots & & \ddots & & \ddots
\end{array}$$

(iii) Let $f(x) = x^k$, $k \in \mathbb{N}$. The binomial theorem shows that

$$\Delta y_n = (x_n + \Delta x)^k - x_n^k = kx_n^{k-1}\Delta x + \binom{k}{2}x_n^{k-2}(\Delta x)^2 + \cdots + kx_n(\Delta x)^{k-1} + (\Delta x)^k$$

$$= (x_n^{k-1} + x_n^{k-2}x_{n+1} + x_n^{k-3}x_{n+1}^2 + \cdots + x_nx_{n+1}^{k-2} + x_{n+1}^{k-1})\Delta x \ .$$

The succeeding rows can be similarly calculated. The general formula is complicated, but the $(k+1)$'st row is constant, $\Delta^k y_n = k!(\Delta x)^k$ for all n, and the succeeding rows are 0.

(iv) Let $f(x) = \frac{1}{x}$. (This is not defined for all $x \in \mathbb{R}$, but we only need that it is defined for $x \geq x_0$ to make the above construction of the differences work, so we just need to take $x_0 > 0$.) Then

$$\Delta y_n = \frac{1}{x_n + \Delta x} - \frac{1}{x_n} = -\frac{1}{x_n(x_n + \Delta x)}\Delta x = -\frac{1}{x_nx_{n+1}}\Delta x$$

$$\Delta^2 y_n = -\frac{1}{x_{n+1}x_{n+2}}\Delta x + \frac{1}{x_nx_{n+1}}\Delta x = \frac{x_{n+2} - x_n}{x_nx_{n+1}x_{n+2}}\Delta x = \frac{2\Delta x}{x_nx_{n+1}x_{n+2}}\Delta x = \frac{2}{x_nx_{n+1}x_{n+2}}(\Delta x)^2$$

and higher differences can be calculated similarly. In this case the rows never become identically zero.

(v) Let $f(x) = \sqrt{x} = x^{1/2}$. Then

$$\Delta y_n = \sqrt{x_n + \Delta x} - \sqrt{x} = (\sqrt{x_n + \Delta x} - \sqrt{x_n})\frac{\sqrt{x_n + \Delta x} + \sqrt{x_n}}{\sqrt{x_n + \Delta x} + \sqrt{x_n}} = \frac{1}{\sqrt{x_n + \Delta x} + \sqrt{x_n}}\Delta x = \frac{1}{\sqrt{x_n} + \sqrt{x_{n+1}}}\Delta x \ .$$

Again, no row of higher differences becomes zero.

(vi) Let $f(x) = a^x$ for fixed $a > 0$ (the comment from (iv) applies to this function). Then the sequence $(y_0, y_1, \dots)$ is a geometric progression with successive ratios $b = a^h$. We have

$$\Delta y_n = a^{x_0 + (n+1)h} - a^{x_0 + nh} = a^{x_0 + nh}(a^h - 1) = cy_n$$

where $c = a^h - 1$. Thus each succeeding row is also a geometric progression.

V.3.11.10. The x_0 is fixed in these calculations. But at the end we want to allow the x_0 to be any number (in the domain), and thus often write x in place of x_0 and y in place of y_0, and hence $\Delta^m y$ in place of $\Delta^m y_0$. Thus the above formulas become:

(i) If $f(x) = mx$, then $\Delta y = m\Delta x$ and $\Delta^k y = 0$ for $k > 1$.

(ii) If $f(x) = x^2$, then $\Delta y = 2x\Delta x + (\Delta x)^2$, $\Delta^2 y = 2(\Delta x)^2$, and $\Delta^m y = 0$ for $m > 2$.

(iii) If $f(x) = x^k$, then

$$\Delta y = kx^{k-1}\Delta x + \binom{k}{2}x^{k-2}(\Delta x)^2 + \cdots + kx(\Delta x)^{k-1} + (\Delta x)^k$$

and $\Delta^k y = k!(\Delta x)^k$.

(iv) If $f(x) = \frac{1}{x}$, then $\Delta y = -\frac{1}{x(x+\Delta x)}\Delta x$.

(v) If $f(x) = \sqrt{x}$, then $\Delta y = \frac{1}{\sqrt{x}+\sqrt{x+\Delta x}}\Delta x$.

(vi) If $f(x) = a^x$, then $\Delta y = ca^x = cy$, where $c = a^{\Delta x} - 1$.

V.3.11.11. The formulas from these examples are analogous to differentiation formulas, and suggest that the derivative of a power function should be a power function whose exponent is reduced by one, and that the derivative of an exponential function should be another exponential function. Here is a nonrigorous but suggestive closer link to differentiation. If we regard $h = \Delta x$ as "infinitesimally small," so that the x_n are essentially equal to $x = x_0$ (at least for n small), in the case $f(x) = x^k$ we have

$$\Delta y_n = (x_n^{k-1} + x_n^{k-2}x_{n+1} + x_n^{k-3}x_{n+1}^2 + \cdots + x_n x_{n+1}^{k-2} + x_{n+1})\Delta x \approx kx^{k-1}\Delta x$$

for small n, and if $f(x) = a^x$ we have

$$\Delta y_n = a^{x+(n+1)h} - a^{x+nh} = a^{x+nh}(a^h - 1) \approx cy\,\Delta x = ca^x\,\Delta x$$

where $c = \lim_{h\to 0}\frac{a^h-1}{h}$.

V.3.11.12. More rigorously, we have

$$f'(x) = \lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}$$

(to be completely rigorous we should also allow $h < 0$) so in the case where $f(x) = x^k$ we get $f'(x) = kx^{k-1}$, and in the case $f(x) = a^x$ we get

$$f'(x) = a^x\left[\lim_{h\to 0}\frac{a^h - 1}{h}\right] .$$

Of course, these are in fact the correct formulas. Note also that in these cases we have

$$f^{(m)}(x) = \lim_{\Delta x\to 0}\frac{\Delta^m y}{(\Delta x)^m}$$

for all m. This is one of the origins of the notation $\frac{d^m y}{dx^m}$ for $f^{(m)}(x)$.

V.3.11.13. One can also consider the inverse problem, which is the discrete antiderivative problem: given the sequence of first differences (Δy_n), find the sequence (y_n) it comes from. The y_n are not uniquely determined: an arbitrary constant can be added to all the y_n. This amounts to making an arbitrary choice of y_0. Once y_0 has been chosen, there is a unique solution:

$$y_n = y_0 + \sum_{m=0}^{n-1} \Delta y_m \ .$$

Factorial Polynomials

Here is another pretty analogy. Our exposition is adapted from [Mil66]. We will normalize to take $h = \Delta x = 1$.

V.3.11.14. If $n \in \mathbb{N}$, define an n'th degree monic polynomial

$$x^{(n)} = x(x-1)(x-2)\cdots(x-n+1)$$

and $x^{(0)} = 1$. The polynomial $x^{(n)}$ is also called $\pi_n(x)$ in the literature (V.11.7.1.). We have

$$\Delta x^{(n)} = (x+1)^{(n)} - x^{(n)} = (x+1)x(x-1)\cdots(x+n-2) - x(x-1)\cdots(x-n+2)(x-n+1)$$

$$= x(x-1)\cdots(x-n+2)[(x+1)-(x-n+1)] = nx(x-1)\cdots(x-n+2) = nx^{(n-1)}$$

motivating the definition as an interesting one. We also have

$$x^{(n)}(x-n)^{(m)} = x^{(n+m)}$$

for $n, m \in \mathbb{N}$ (and even in $\mathbb{N} \cup \{0\}$).

V.3.11.15. The polynomials $\{x^{(0)}, x^{(1)}, \ldots, x^{(n)}\}$ are clearly linearly independent in P_n () since none is a linear combination of the previous ones, which are of smaller degree. Thus they form a basis for the $(n+1)$-dimensional vector space P_n. The change-of-basis matrix from the usual basis $\{1, x, x^2, \ldots, x^n\}$ is triangular, with 1's on the diagonal. Thus any polynomial of degree n

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

can be written uniquely as a *factorial polynomial*

$$p(x) = b_0 x^{(0)} + b_1 x^{(1)} + b_2 x^{(2)} + \cdots + b_n x^{(n)}$$

and $b_n = a_n$ (and in particular $b_0 x^{(0)} + \cdots + b_n x^{(n)}$ is identically 0 if and only if all b_k are 0). For such a polynomial we have

$$\Delta p = b_1 + 2b_2 x^{(1)} + \cdots + n b_n x^{(n-1)} \ .$$

V.3.11.16. If $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ is a polynomial of degree $\leq n$, then Taylor's formula () yields that

$$p(x) = p(0) + p'(0)x + \frac{p''(0)}{2!}x^2 + \cdots + p^{(n)}(0)x^n$$

for all x (this formula can be easily proved by noting that $p^{(k)}(0) = k!a_k$ for all k). Similarly, if

$$p(x) = b_0x^{(0)} + b_1x^{(1)} + \cdots + b_nx^{(n)}$$

(note the unfortunate coincidence of notation: $x^{(k)}$ does not denote a derivative here!) a simple calculation shows that

$$\Delta^k p(0) = k!b_k$$

and hence

$$p(x) = p(0) + \Delta p(0)x^{(1)} + \frac{\Delta^2 p(0)}{2!}x^{(2)} + \cdots + \frac{\Delta^n p(0)}{n!}x^{(n)}$$

(this is a purely algebraic formula not involving any derivatives). This formula is called *Newton's formula* (NEWTON is credited with many formulas), and is the difference version of Taylor's formula.

The factorial polynomials have many applications in interpolation theory; cf. V.11.7.1.. Here is an example related to the discussion in I.6.1.26..

V.3.11.17. EXAMPLE. Here is a sample calculation using this formula (from [Mil66]). Suppose we want to find the unique polynomial $p(x)$ of degree ≤ 7 for which $p(0) = 0$, $p(1) = 7$, $p(2) = 26$, $p(3) = 63$, $p(4) = 124$, $p(5) = 215$, $p(6) = 342$, and $p(7) = 511$ (it may not be hard to find this polynomial by inspection or trial and error in this case; cf. I.6.1.38.). Newton's formula gives a systematic way to do it without solving a system of eight linear equations. Make up a table:

x	0	1	2	3	4	5	6	7
$p(x)$	0	7	26	63	124	215	342	511
$\Delta p(x)$	7	19	37	61	91	127	169	
$\Delta^2 p(x)$	12	18	24	30	36	42		
$\Delta^3 p(x)$	6	6	6	6	6			
$\Delta^4 p(x)$	0	0	0	0				

(the table need not be continued farther once a zero row, or even a constant row, is obtained). Thus

$$p(x) = 0 + 7x^{(1)} + \frac{12}{2!}x^{(2)} + \frac{6}{3!}x^{(3)} + 0x^{(4)}$$

$$= 7x + 6x(x-1) + x(x-1)(x-2) = x^3 + 3x^2 + 3x = (x+1)^3 - 1 \ .$$

V.3.11.18. The definition of $x^{(n)}$ can be extended to negative n. The right way to do it is to preserve the formula $x^{(n)}(x-n)^{(m)} = x^{(n+m)}$, i.e. we should have $x^{(-n)}(x+n)^{(n)} = x^0 = 1$ so we define

$$x^{(-n)} = \frac{1}{(x+n)^{(n)}} = \frac{1}{(x+n)(x+n-1)\cdots(x-1)} \ .$$

A simple calculation then verifies that $x^{(n)}(x-n)^{(m)} = x^{(n+m)}$ for all $n, m \in \mathbb{Z}$, and perhaps more importantly

$$\Delta x^{(n)} = n x^{(n-1)}$$

for all $n \in \mathbb{Z}$.

V.3.12. Exercises

V.3.12.1. Stirling Numbers. The coefficients in the change-of-basis matrix between the ordinary power functions x^k and the factorial polynomials $x^{(k)}$ are called *Stirling numbers*. Specifically, if $0 \le r \le n$ we define S_r^n by

$$x^{(n)} = \sum_{r=0}^{n} S_r^n x^r \ .$$

The S_r^n are the *Stirling numbers of the first kind*. We also define $\mathcal{S}_r^n$ by

$$x^n = \sum_{r=0}^{n} \mathcal{S}_r^n x^{(r)} \ .$$

The $\mathcal{S}_r^n$ are the *Stirling numbers of the second kind*. Define $S_r^n = \mathcal{S}_r^n = 0$ for $r > n$.

(a) Show that $S_0^n = \mathcal{S}_0^n = 0$ and $S_n^n = \mathcal{S}_n^n = 1$ for all $n \in \mathbb{N}$.

(b) Show that the following recurrence relations hold for all $n, r \in \mathbb{N}$:

$$S_r^{n+1} = S_{r-1} n - n S_r^n$$

$$\mathcal{S}_r^{n+1} = \mathcal{S}_{r-1}^n + r\mathcal{S}_r^n \ .$$

(c) Show that the Stirling numbers are completely determined by the conditions in (a) and (b) for all n and all r, $1 \le r \le n$.

(d) Compute the Stirling numbers for small n (say, $n \le 6$). They are always integers; Stirling numbers of the second kind are nonnegative, but some Stirling numbers of the first kind are negative.

V.4. Extensions of the Derivative

There are a number of ways to extend the derivative concept which prove to be useful. We discuss several of these generalizations in this section.

V.4.1. One-Sided Derivatives

The first generalization we consider is one-sided derivatives. The definition is quite straightforward:

V.4.1.1. DEFINITION. (i) Let f be a real-valued function on the interval $[x_0, b)$. If

$$\lim_{h\to 0+} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{x\to x_0+} \frac{f(x)-f(x_0)}{x-x_0}$$

exists (as a real number), then f is *right differentiable* at x_0, and this number is the *right derivative* of f at x_0, denoted $D_R f(x_0)$.

(ii) Let f be a real-valued function on the interval $(a, x_0]$. If

$$\lim_{h\to 0-} \frac{f(x_0+h)-f(x_0)}{h} = \lim_{x\to x_0-} \frac{f(x)-f(x_0)}{x-x_0}$$

exists (as a real number), then f is *left differentiable* at x_0, and this number is the *left derivative* of f at x_0, denoted $D_L f(x_0)$.

Various notations are used for left and right derivatives, such as $D_l f$ and f'_- for $D_L f$ and $D_r f$ and f'_+ for $D_R f$.

The next proposition is obvious from ().

V.4.1.2. PROPOSITION. Let f be a real-valued function on an interval (a, b), and $x_0 \in (a, b)$. Then f is differentiable at x_0 if and only if it is both left and right differentiable at x_0 and $D_L f(x_0) = D_R f(x_0)$. We then have

$$f'(x_0) = D_L f(x_0) = D_R f(x_0) \ .$$

We can have left and right derivatives which are not equal, or one or both may fail to exist:

V.4.1.3. EXAMPLES. (i) Let $f(x) = |x|$. Then $D_L f(0) = -1$ and $D_R f(x) = 1$. Thus f is both left and right differentiable at 0, but not differentiable at 0.

(ii) Let $f(x) = \sqrt{x}$ on $[0, \infty)$. Then $D_R f(0)$ does not exist (we sometimes say $D_R f(0) = +\infty$), and $D_L f(0)$ is undefined since f is undefined to the left of 0. So f is neither left or right differentiable at 0.

(iii) Let $f(x) = \sqrt{x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. Then $D_R f(0)$ does not exist and $D_L f(0) = 0$. So f is left differentiable but not right differentiable at 0.

(iv) Let $f(x) = x \sin \frac{1}{x}$ for $x \neq 0$ and $f(0) = 0$. Then neither $D_R f(0)$ nor $D_L f(0)$ exists. So f is neither left or right differentiable at 0.

All these examples are continuous. We have the following slight generalization of (), whose proof is nearly identical and is left to the reader. Recall the definition of left and right continuity ().

V.4.1.4. PROPOSITION. (i) Let f be a real-valued function on an interval $[x_0, b)$. If f is right differentiable at x_0, then f is right continuous at x_0.
(ii) Let f be a real-valued function on an interval $(a, x_0]$. If f is left differentiable at x_0, then f is left continuous at x_0.

Endpoint Differentiability and Differentiability on an Interval

If x_0 is an endpoint of the domain of a function, it does not make sense to say that f is differentiable at x_0. But one-sided differentiability does make sense, and is called *endpoint differentiability* in this context. So we may make the following definition in analogy with ():

V.4.1.5. DEFINITION. Let f be a real-valued function on an interval $[a, b]$. Then f is *differentiable on* $[a, b]$ if it is differentiable at all points of (a, b), right differentiable at a, and left differentiable at b. Differentiability on an interval of the form $[a, b)$ or $(a, b]$ is defined analogously.

V.4.2. Dini Derivatives

A further generalization is very useful. Recall the definition of lim sup and lim inf from ().

V.4.2.1. DEFINITION. (i) Let f be a real-valued function on the interval $[x_0, b)$. Set

$$D^+f(x_0) = \limsup_{h\to 0+} \frac{f(x_0+h)-f(x_0)}{h} = \limsup_{x\to x_0+} \frac{f(x)-f(x_0)}{x-x_0}$$

$$D_+f(x_0) = \liminf_{h\to 0+} \frac{f(x_0+h)-f(x_0)}{h} = \liminf_{x\to x_0+} \frac{f(x)-f(x_0)}{x-x_0}$$

(as extended real numbers). $D^+f(x_0)$ and $D_+f(x_0)$ are the *upper and lower right Dini derivatives* of f at x_0.
(i) Let f be a real-valued function on the interval $(a, x_0]$. Set

$$D^-f(x_0) = \limsup_{h\to 0-} \frac{f(x_0+h)-f(x_0)}{h} = \limsup_{x\to x_0-} \frac{f(x)-f(x_0)}{x-x_0}$$

$$D_-f(x_0) = \liminf_{h\to 0-} \frac{f(x_0+h)-f(x_0)}{h} = \liminf_{x\to x_0-} \frac{f(x)-f(x_0)}{x-x_0}$$

(as extended real numbers). $D^-f(x_0)$ and $D_-f(x_0)$ are the *upper and lower left Dini derivatives* of f at x_0.

V.4.2.2. Note that the Dini derivatives are always defined, provided only that the function is defined on an interval to the appropriate side of x_0. We clearly always have $D_+f(x_0) \le D^+f(x_0)$ and $D_-f(x_0) \le D^-f(x_0)$.

We then have an obvious characterization of left and right differentiability (cf. ()):

V.4.2.3. PROPOSITION. (i) Let f be a real-valued function on an interval $[x_0, b)$. Then f is right differentiable at x_0 if and only if $D^+f(x_0)$ and $D_+f(x_0)$ are equal and finite. We then have

$$D_Rf(x_0) = D^+f(x_0) = D_+f(x_0) \ .$$

(ii) Let f be a real-valued function on an interval $(a, x_0]$. Then f is left differentiable at x_0 if and only if $D^-f(x_0)$ and $D_-f(x_0)$ are equal and finite. We then have

$$D_Lf(x_0) = D^-f(x_0) = D_-f(x_0) \ .$$

(iii) Let f be a real-valued function on an interval (a, b), and $x_0 \in (a, b)$. Then f is differentiable at x_0 if and only if all four Dini derivatives at x_0 are equal and finite. We then have

$$f'(x_0) = D^+f(x_0) = D_+f(x_0) = D^-f(x_0) = D_-f(x_0) \ .$$

V.4.2.4. EXAMPLES. (i) Let $f(x) = |x|$. Then $D^-f(0) = D_-f(0) = -1$ and $D^+f(0) = D_+f(x) = 1$.
(ii) Let $f(x) = \sqrt{x}$ on $[0, \infty)$. Then $D^+f(0) = D_+f(0) = +\infty$, and $D^-f(0)$ and $D_-f(0)$ are undefined since f is undefined to the left of 0.
(iii) Let $f(x) = \sqrt{|x|}$. Then $D^+f(0) = D_+f(0) = +\infty$ and $D^-f(0) = D_-f(0) = -\infty$.
(iv) Let $f(x) = x \sin \frac{1}{x}$ for $x \neq 0$ and $f(0) = 0$. Then $D^+f(0) = D^-f(0) = 1$ and $D_+f(0) = D_-f(0) = -1$.
(v) Let $f(x) = x \sin \frac{1}{x}$ for $x > 0$ and $f(x) = 0$ for $x \leq 0$. Then $D^+f(0) = 1$, $D_+f(0) = -1$, and $D^-f(0) = D_-f(0) = 0$.
(vi) Let $f(x) = \sqrt{x} \sin \frac{1}{x}$ for $x \neq 0$ and $f(0) = 0$. Then $D^+f(0) = D^-f(0) = +\infty$ and $D_+f(0) = D_-f(0) = -\infty$.
(vii) Let $f(x) = x^{1/3}$. Then $D^+f(0) = D_+f(0) = D^-f(0) = D_-f(0) = +\infty$.

We have a version of () also for Dini derivatives:

V.4.2.5. PROPOSITION. (i) Let f be a real-valued function on an interval $[x_0, b)$. If f is not right continuous at x_0, then at least one right Dini derivative at x_0 is infinite.
(ii) Let f be a real-valued function on an interval $(a, x_0]$. If f is not left continuous at x_0, then at least one left Dini derivative at x_0 is infinite.

PROOF: The proofs of both (i) and (ii) consist of four nearly identical cases. We do only one of the cases for (i), and leave the rest of the cases to the reader.

In (i), if f is not right continuous at x_0, then either $f(x_0) \neq \liminf_{x \to x_0+} f(x)$ or $f(x_0) \neq \limsup_{x \to x_0+} f(x)$ (or both). Suppose $f(x_0) < \liminf_{x \to x_0+} f(x)$. Then

$$\liminf_{x \to x_0+} [f(x) - f(x_0)] > 0$$

and, since $\lim_{x\to x_0+} \dfrac{1}{x-x_0} = +\infty$, by () we have

$$D_+f(x_0) = \liminf_{x\to x_0+} \frac{f(x)-f(x_0)}{x-x_0} = +\infty \ .$$



V.4.2.6. Note that the converse is not true by Examples V.4.2.4.(iii),(v),(vi). In fact, there are (many!) continuous functions f on $\mathbb{R}$ for which $D^+f(x) = D^-f(x) = +\infty$ and $D_+f(x) = D_-f(x) = -\infty$ for *every* $x \in \mathbb{R}$ ().

Local Extrema

Here is a general version of V.3.6.3.. The proof is essentially the same.

V.4.2.7. PROPOSITION. Let f be a function on an interval (a,b), and $x_0 \in (a,b)$.
(i) If f has a local maximum at x_0, then $D_-f(x_0) \geq 0$ and $D^+f(x_0) \leq 0$.
(ii) If f has a local minimum at x_0, then $D^-f(x_0) \leq 0$ and $D_+f(x_0) \geq 0$.

This result also applies when x_0 is an endpoint of an interval in the domain of f, except that the Dini derivatives on the outside of the interval are not defined (or not relevant):

V.4.2.8. PROPOSITION. Let f be a real-valued function on $[a,b]$. Then
(i) If f has an endpoint maximum at a, then $D^+f(a) \leq 0$.
(ii) If f has an endpoint minimum at a, then $D_+f(a) \geq 0$.
(iii) If f has an endpoint maximum at b, then $D_-f(a) \geq 0$.
(iv) If f has an endpoint minimum at b, then $D^-f(a) \leq 0$.

V.4.2.9. Converses of these statements are, of course, false; the counterexamples in the differentiable case, such as $f(x) = x^3$ at 0, are also counterexamples here. But there is a version of the First Derivative Test in this generality ().

V.4.2.10. At a single point, there may be no relation between the left and right Dini derivatives. However, at "most" points both upper Dini derivatives are greater than both lower derivatives:

V.4.2.11. THEOREM. Let f be a real-valued function on an open interval I. Set

$$A = \{x \in I : D^- f(x) < D_+ f(x)\}$$

$$B = \{x \in I : D^+ f(x) < D_- f(x)\} \ .$$

Then A and B are countable.

PROOF: We prove that A is countable; the proof for B is essentially identical. Let $x \in A$. Choose $r_x \in \mathbb{Q}$ with $D^- f(x) < r_x < D_+ f(x)$, and then fix $s_x, t_x \in \mathbb{Q} \cap I$ with $s_x < x < t_x$ and

$$\frac{f(y) - f(x)}{y - x} < r_x \text{ for all } y, s_x < y < x$$

$$\frac{f(y) - f(x)}{y - x} > r_x \text{ for all } y, x < y < t_x \ .$$

Then, for any $y \neq x$ with $s_x < y < t_x$, we have

$$f(y) - f(x) > r_x(y - x) \ .$$

Define $\phi : A \to \mathbb{Q}^3$ by $\phi(x) = (r_x, s_x, t_x)$. We show that ϕ is one-to-one; then, since $\mathbb{Q}^3$ is countable, it will follow that A is countable. Suppose $\phi(x) = \phi(z)$ for $x, z \in A$, $x \neq z$. Since $s_x = s_z < z < t_z = t_x$, we have

$$f(z) - f(x) > r_x(z - x) \ .$$

Similarly, since $s_z < x < t_z$, we have

$$f(x) - f(z) > r_z(x - z) \ .$$

But $r_x = r_z$, a contradiction. 👍

We get the following immediate consequence, which is not so obvious geometrically:

V.4.2.12. COROLLARY. Let f be a function on an interval I. Then the set of $x \in I$ for which $D_L f(x)$ and $D_R f(x)$ are both defined (even allowing infinite values) but unequal is countable. In particular, if f is left and right differentiable everywhere in I, then f is differentiable at all but countably many points of I.

A far deeper result is the remarkable theorem proved successively for larger classes of functions by A. DENJOY, GRACE CHISHOLM YOUNG, and S. SAKS:

V.4.2.13. THEOREM. [DENJOY-YOUNG-SAKS] Let f be an arbitrary real-valued function on an interval $[a, b]$. Then for almost all $x \in [a, b]$, one of the following four conditions holds:

(1) f is differentiable at x.

(2) $D^- f(x) = D_+ f(x) \in \mathbb{R}$, $D_- f(x) = -\infty$, and $D^+ f(x) = +\infty$.

(3) $D^+ f(x) = D_- f(x) \in \mathbb{R}$, $D_+ f(x) = -\infty$, and $D^- f(x) = +\infty$.

(4) $D_-f(x) = D_+f(x) = -\infty$ and $D^-f(x) = D^+f(x) = +\infty$.

In particular, for almost all $x \in [a,b]$, either f is differentiable at x or else $\max(D^-f(x), D^+f(x)) = +\infty$ and $\min(D_-f(x), D_+f(x)) = -\infty$.

If f is measurable, then for almost all x, either condition (1) or condition (4) holds (i.e. conditions (2) and (3) hold only on sets of measure zero).

The proof uses some sophisticated machinery, and will be given in (). This result is closely related to a result about approximate tangents to rectifiable sets in geometric measure theory [Mor09, 3.12].

V.4.3. Strong Derivatives

V.4.4. Symmetric First and Second Derivatives

V.4.5. Approximate and Weak Derivatives

V.4.6. Exercises

V.4.6.1. Define a function $f : [0,1) \to [0,1)$ as follows. If $x \in [0,1)$, write

$$x = .d_1d_2d_3\cdots$$

in its decimal expansion not ending in an infinite string of 9's. Set

$$f(x) = .0d_10d_20d_3\cdots\ .$$

(a) Show that f is a strictly increasing function.

(b) Show that f is right continuous everywhere.

(c) Show that f is left continuous except at the decadent rationals ().

(d) Show that $D_Rf(x)$ exists and equals 0 everywhere.

V.5. Exponential and Trigonometric Functions

Anyone who has studied mathematics through calculus knows that exponential and trigonometric functions are of great importance, both theoretical and practical. However, the theoretical subtleties involved in properly defining these functions and establishing their properties are often underappreciated. Since the functions are so important and useful, they are usually introduced in the mathematics curriculum before the necessary tools (from calculus) for a proper treatment are available; and by the time the right tools are discussed, the functions are so familiar that they are often taken for granted. (These comments also apply to the historical development of exponential and trigonometric functions.)

A careful and rigorous treatment of exponential and/or trigonometric functions can be done in several ways. The two most common methods are: (1) defining one of the inverse functions (log or arcsin) as the antiderivative of a familiar function, or (2) using infinite series. In the case of exponential functions, there is a third approach which proceeds directly from algebraic rules of exponents. This third approach is the one we will take for exponential and logarithm functions in this section; see Exercise () for the first approach via integrals, and () for the infinite series approach. Our treatment of trigonometric functions will follow approach (1) (see () for the infinite series approach here).

V.5.1. Rational Exponents

V.5.1.1. Recall that if $x \in \mathbb{R}$ and $n \in \mathbb{N}$, then $x^n = x \cdot x \cdots x$ (n times). If $x \neq 0$, extend this to $n \in \mathbb{Z}$ by $x^0 = 1$ and $x^{-n} = (x^{-1})^n = (x^n)^{-1}$. The usual rules of exponents hold: $x^{m+n} = x^m x^n$, $x^{m-n} = \frac{x^m}{x^n}$, $(x^m)^n = (x^n)^m = x^{mn}$, $(xy)^n = x^n y^n$ for all $x, y \neq 0$ and all $m, n \in \mathbb{Z}$.

If $n \in \mathbb{N}$ and $x > 1$, then $x^n > x > 1$. So if $n \in \mathbb{N}$ and $0 < x < y$, then $x^n < x^n \left(\frac{y}{x}\right)^n = \left(x \cdot \frac{y}{x}\right)^n = y^n$, so $f(x) = x^n$ is strictly increasing on $(0, \infty)$. Similarly, if $n \in \mathbb{Z}$, $n < 0$, then $f(x) = x^n = (x^{-1})^{-n}$ is strictly decreasing on $(0, \infty)$. For any $n \in \mathbb{Z}$, $f(x) = x^n$ is continuous and even differentiable on $(0, \infty)$, and its range is unbounded above and below in $(0, \infty)$ if $n \neq 0$.

Thus, by the IVT, if $n \in \mathbb{Z}$, $n \neq 0$, then the range of $f(x) = x^n$ is all of $(0, \infty)$, and by () f has a continuous inverse function f^{-1} which also maps $(0, \infty)$ onto $(0, \infty)$, and f^{-1} is strictly increasing if $n > 0$, strictly decreasing if $n < 0$. If $x > 0$, write $x^{1/n}$ for $f^{-1}(x)$. If $n > 0$, the definition of $x^{1/n}$ can also be extended to $x = 0$ by setting $0^{1/n} = 0$, and if n is odd, to $x < 0$ by $x^{1/n} = -(-x)^{1/n}$; the extended function is continuous on its domain.

V.5.1.2. Proposition. If $p, q \in \mathbb{Z}$, $q \neq 0$, for $x > 0$ define $x^{p/q} = (x^p)^{1/q}$. We then have:

(i) For $p, q \in \mathbb{Z}$, $q \neq 0$, and $x > 0$, $[(x^p)^{1/q}]^q = x^p = [(x^{1/q})^q]^p = [(x^{1/q})^p]^q$, so $(x^p)^{1/q} = (x^{1/q})^p$ since $f(x) = x^q$ is one-to-one.

(ii) Similarly, if $p, q, r, s \in \mathbb{Z}$, $q, s \neq 0$, and $x > 0$, then $[(x^{p/q})^{r/s}]^{qs} = x^{pr} = [(x^{r/s})^{p/q}]^{qs}$, so $(x^{p/q})^{r/s} = (x^{r/s})^{p/q} = x^{pr/qs}$. In particular, $x^{ps/qs} = x^{p/q}$, and hence x^t is well defined for any $t \in \mathbb{Q}$.

(iii) If $r, s \in \mathbb{Q}$, $x, y \in (0, \infty)$, we have $(x^r)^s = x^{rs}$, $x^{r+s} = x^r x^s$, $x^{r-s} = \frac{x^r}{x^s}$, $(xy)^r = x^r y^r$.

(iv) If $r \in \mathbb{Q}$, $r \neq 0$, the function $f(x) = x^r$ is a composition of two strictly monotone continuous functions from $(0, \infty)$ onto $(0, \infty)$, hence strictly monotone and continuous, strictly increasing if $r > 0$ and strictly decreasing if $r < 0$.

(v) $1^r = 1$ for all $r \in \mathbb{Q}$, so if $x > 1$, then $x^r > 1$ for $r > 0$ and $0 < x^r < 1$ for $r < 0$.

V.5.1.3. PROPOSITION. If $r \in \mathbb{Q}$, then $f(x) = x^r$ is differentiable on $(0, \infty)$, and $f'(x) = rx^{r-1}$ on $(0, \infty)$.

PROOF: If $n \in \mathbb{N}$, then $\frac{d}{dx}(x^n) = nx^{n-1}$ (**??**). The formula also obviously holds if $n = 0$. If $n \in \mathbb{N}$, then using the Quotient Rule x^{-n} is differentiable on $(0, \infty)$ and

$$\frac{d}{dx}\left(x^{-n}\right) = \frac{d}{dx}\left(\frac{1}{x^n}\right) = \frac{0 - nx^{n-1}}{x^{2n}} = -nx^{-n-1}$$

so the formula holds for all $r \in \mathbb{Z}$. If $q \in \mathbb{Z}$, $q \neq 0$, then by the Inverse Function Theorem (**??**) $x^{1/q}$ is differentiable on $(0, \infty)$ and

$$\frac{d}{dx}\left(x^{1/q}\right) = \frac{1}{q(x^{1/q})^{q-1}} = \frac{1}{qx^{1-1/q}} = \frac{1}{q}x^{(1/q)-1}$$

If $r = \frac{p}{q}$, then by the Chain Rule x^r is differentiable on $(0, \infty)$ and

$$\frac{d}{dx}\left(x^{p/q}\right) = \frac{d}{dx}\left((x^p)^{1/q}\right) = \frac{1}{q}(x^p)^{(1/q)-1} \cdot px^{p-1} = \frac{p}{q}x^{[(p/q)-p]+[p-1]} = \frac{p}{q}x^{(p/q)-1}.$$



V.5.2. Irrational Exponents and Exponential Functions

Now fix $b > 1$ and look at the function $f(x) = b^x$ with domain $\mathbb{Q}$.

V.5.2.1. PROPOSITION. b^x is strictly increasing.

PROOF: If $x, y \in \mathbb{Q}$ with $x < y$, then $b^y = b^{x+(y-x)} = b^x b^{y-x}$, and $b^{y-x} > 1$ by (v).

V.5.2.2. PROPOSITION. $\{b^n : n \in \mathbb{N}\}$ is unbounded.

PROOF: If $S = \{b^n : n \in \mathbb{N}\}$ is bounded, let $a = \sup(S)$. Since $b > 1$, $\frac{a}{b} < a$, so there is an $n \in \mathbb{N}$ with $b^n > \frac{a}{b}$. Then $b^{n+1} > a$, a contradiction.

V.5.2.3. COROLLARY. $\lim_{n\to\infty} b^{1/n} = 1$.

PROOF: The sequence $(b^{1/n})$ is decreasing and bounded below by 1, hence $\lim_{n\to\infty} b^{1/n}$ exists and equals $c := \inf_{n\in\mathbb{N}} b^{1/n}$. Then $c \geq 1$ and $c^n \leq b$ for all $n \in \mathbb{N}$, so $c = 1$ by V.5.2.2..

V.5.2.4. PROPOSITION. If $a \in \mathbb{Q}$, then the function $f(x) = b^x$ is uniformly continuous on $(-\infty, a) \cap \mathbb{Q}$.

PROOF: If $\epsilon > 0$, then there is an $n \in \mathbb{N}$ with $b^{1/n} - 1 < b^{-a}\epsilon$. Set $\delta(\epsilon) = 1/n$. Then, if $x, y \in \mathbb{Q}$, $x < y < a$, $y - x < \delta(\epsilon)$, we have

$$0 < b^y - b^x = b^x(b^{y-x} - 1) < b^a(b^{y-x} - 1) < \epsilon.$$

V.5.2.5. COROLLARY. The function $f(x) = b^x$ ($x \in \mathbb{Q}$) extends to a continuous strictly increasing function on all of $\mathbb{R}$, also denoted b^x, with range $(0, \infty)$.

PROOF: The extension is guaranteed by **??**, and is continuous and increasing. To see that it is strictly increasing, let $x, y \in \mathbb{R}$, $x < y$, and let $r, s \in \mathbb{Q}$, $x < r < s < y$. Then $b^x \leq b^r < b^s \leq b^y$. The range is an interval (**??**) which must be all of $(0, \infty)$ by V.5.2.2..

V.5.2.6. DEFINITION. The function $f(x) = b^x$ thus has a continuous strictly increasing inverse function f^{-1} with domain $(0, \infty)$ and range $\mathbb{R}$. The number $f^{-1}(x)$ is called the *logarithm of x with base b*, denoted $\log_b(x)$.

V.5.2.7. PROPOSITION. The following laws of exponents and logarithms hold: if $x, y \in \mathbb{R}$ and $b, c > 1$, then $b^{x+y} = b^x b^y$, $b^{x-y} = \frac{b^x}{b^y}$, $b^{xy} = (b^x)^y$, $(bc)^x = b^x c^x$; if $s, t > 0$, $r \in \mathbb{R}$, then $\log_b(st) = \log_b(s) + \log_b(t)$, $\log_b(s/t) = \log_b(s) - \log_b(t)$, $\log_b(s^r) = r \log_b(s)$. If $c > 1$, then $\log_b(s) = \log_c(s) \log_b(c)$. In particular, $\log_b 1 = 0$ and $\log_b(c) = \frac{1}{\log_c(b)}$.

PROOF: The laws of exponents have already been established for $x, y \in \mathbb{Q}$. They follow for general x, y by continuity: for example, if (x_n) and (y_n) are sequences of rational numbers with $x_n \to x$ and $y_n \to y$, then $b^{x_n} \to b^x$, $b^{y_n} \to b^y$, $b^{x_n + y_n} \to b^{x+y}$, and $b^{x_n} b^{y_n} \to b^x b^y$ by continuity of b^x and standard limit theorems. The laws of logarithms follow from the laws of exponents by setting $x = \log_b(s)$ and $y = \log_b(t)$, so $s = b^x$ and $t = b^y$: for example, $st = b^x b^y = b^{x+y}$, so $\log_b(st) = x + y$. For the last formula, if $z = \log_c(s)$, then $s = c^z = (b^{\log_b c})^z = b^{z \log_b c}$, so $\log_b(s) = z \log_b c$.

If $0 < b < 1$ and $x \in \mathbb{R}$, we can define $b^x = (b^{-1})^{-x}$. Then b^x is strictly decreasing, as is the inverse function $\log_b(x)$. The same rules of exponents and logarithms hold. We can also define $1^x = 1$ for all x.

An important property of the exponential function b^x ($b > 0$, $b \neq 1$) is that it is strictly convex (**??**):

V.5.2.8. LEMMA. If $b > 1$, $k \in \mathbb{R}$, and $x > 0$, then $b^{k+2x} - b^k > 2(b^{k+x} - b^k)$.

PROOF:

$$b^{k+2x} - b^k = b^k(b^{2x} - 1) = b^k(b^x + 1)(b^x - 1) = (b^x + 1)(b^{k+x} - b^k).$$

V.5.2.9. PROPOSITION. If $b > 1$ and $x, y, z \in \mathbb{R}$ with $z < y < x$, then

$$\frac{b^y - b^z}{y - z} < \frac{b^x - b^z}{x - z} < \frac{b^x - b^y}{x - y}.$$

(It follows that $f(x) = b^x$ is a strictly convex function.)

PROOF: Lemma V.5.2.8. means that on any interval $[x_1, x_2]$ with midpoint c, the point (c, b^c) lies below the line between (x_1, b^{x_1}) and (x_2, b^{x_2}). Thus the first inequality follows from repeated applications of V.5.2.8. if $y - z$ is a dyadic rational multiple of $x - z$ (see Exercise 1). By continuity the first inequality holds with a $\leq$ for all x, y, z. To show that the strict inequality holds in general, choose w with $y < w < x$ and $y - z$ a dyadic rational multiple of $w - z$. Then

$$\frac{b^y - b^z}{y - z} < \frac{b^w - b^z}{w - z} \leq \frac{b^x - b^z}{x - z}.$$

For the second inequality, suppose that $\frac{b^x - b^y}{x - y} \leq \frac{b^x - b^z}{x - z}$. Then

$$b^x - b^z = (y - z)\frac{b^y - b^z}{y - z} + (x - y)\frac{b^x - b^y}{x - y} < (y - z)\frac{b^x - b^z}{x - z} + (x - y)\frac{b^x - b^z}{x - z} = b^x - b^z$$

which is a contradiction. 👍

V.5.2.10. COROLLARY. Let $b > 1$, and $0 < z < y < x$. Then

$$\frac{\log_b x - \log_b y}{x - y} < \frac{\log_b x - \log_b z}{x - z} < \frac{\log_b y - \log_b z}{y - z}.$$

Note that V.5.2.9. and V.5.2.10. also hold for $0 < b < 1$. For V.5.2.9.,

$$\frac{(b^{-1})^{-y} - (b^{-1})^{-x}}{-y + x} < \frac{(b^{-1})^{-z} - (b^{-1})^{-x}}{-z + x} < \frac{(b^{-1})^{-z} - (b^{-1})^{-y}}{-z + y}.$$

V.5.3. The Number e

V.5.3.1. LEMMA. If $x > 0$ and $k > 1$, then $(1 + x)^k > 1 + kx$.

PROOF: Fix $k > 1$ rational, and let $f(x) = (1 + x)^k - 1 - kx$ for $x \geq 0$. Then f is differentiable for $x \geq 0$, $f(0) = 0$, and $f'(x) = k(1 + x)^{k-1} - k = k[(1 + x)^{k-1} - 1] > 0$ for $x > 0$. The formula thus holds for k rational, and thus for k irrational with $\geq$ by continuity. To see that the strict inequality holds for general $k > 1$, choose rational r with $1 < r < k$. Then

$$(1 + x)^k = [(1 + x)^r]^{k/r} > [1 + rx]^{k/r} \geq 1 + \frac{k}{r}rx.$$

👍

V.5.3.2. COROLLARY. The function $g(x) = (1+x)^{1/x}$ is decreasing for $x > 0$.

PROOF: Let $x > 0$ and $k > 1$. Then $(1+kx)^{1/kx} < ([1+x]^k)^{1/kx} = (1+x)^{1/x}$. 👍

Thus, to show that $\lim_{x\to 0+}(1+x)^{1/x}$ exists, it suffices to show that $(1+x)^{1/x}$ is bounded for $x > 0$.

V.5.3.3. THEOREM. Let $h(x) = (1+x)^{1+1/x}$ for $x > 0$, and $f(y) = (1-y)^{-1/y} = \frac{1}{(1-y)^{1/y}}$ for $0 < y < 1$. Then

(i) f and h are increasing.

(ii) $g(x) < h(z) = f\left(\frac{z}{1+z}\right)$ for all $x, z > 0$.

(iii) $\lim_{x\to 0+} f(x)$, $\lim_{x\to 0+} g(x)$ and $\lim_{x\to 0+} h(x)$ exist and are equal.

V.5.3.4. LEMMA. Let $b > 1$, $x > 0$, and $0 < y < 1$. Then

$$\log_b([1+x]^{1/x}) = \frac{\log_b(1+x)}{x} < -\frac{\log_b(1-y)}{y} = \log_b\left(\left[\frac{1}{1-y}\right]^{1/y}\right).$$

PROOF: Apply V.5.2.10. to $1-y, 1, 1+x$. 👍

We now prove Theorem V.5.3.3..
PROOF: For (i) and (ii), fix $b > 1$ and for $z > 0$ set $y = \frac{z}{1+z}$. Then $0 < y < 1$ and $z = \frac{1}{1-y} - 1$, and z increases from 0 to $+\infty$ as y increases from 0 to 1. By V.5.2.10., $-\frac{\log_b(1-y)}{y} = \log_b\left(\left[\frac{1}{1-y}\right]^{1/y}\right)$ increases in y for $0 < y < 1$ [if $0 < y_1 < y_2 < 1$, apply V.5.2.10. to $1-y_2, 1-y_1, 1$], so $\log_b([1+z]^{\frac{1+z}{z}})$ increases in z for $z > 0$, proving (i), and V.5.3.4. with y converted to z gives (ii).

Thus $g(x)$ is increasing and bounded above as $x \to 0+$, and $f(x)$ and $h(x)$ are decreasing and bounded below, so $\lim_{x\to 0+} g(x) = \sup_{x>0} g(x)$ and

$$\lim_{x\to 0+} f(x) = \lim_{x\to 0+} h(x) = \inf_{x>0} h(x) = \inf_{x>0} f(x)$$

exist; call the limits d and e respectively. Then $2 = (1+1)^1 < d \le e$, and $\lim_{x\to 0+} \frac{h(x)}{g(x)} = \frac{e}{d}$. But $\frac{h(x)}{g(x)} = 1+x$, so $\lim_{x\to 0+} \frac{h(x)}{g(x)} = 1 = \frac{e}{d}$, $d = e$. 👍

V.5.3.5. DEFINITION. The number e is the limit $\lim_{x\to 0+}(1+x)^{1/x} = \lim_{x\to+\infty}\left(1+\frac{1}{x}\right)^x = \lim_{x\to+\infty}\left(1-\frac{1}{x}\right)^{-x}$.

The number e is one of the most important constants in mathematics, comparable in importance to π.

V.5.3.6. COROLLARY. $e > (1+1)^1 = 2 > 1$, so $f(x) = e^x$ is strictly increasing and continuous, and its inverse function $\ln x = \log_e x$ is continuous.

The function $\ln x$ is called the *natural logarithm function*, and mathematicians usually write it $\log x$.

V.5.3.7. We can of course get better estimates of the numerical value of e using V.5.3.3. for small x: for example, if $x = .01$, we get $2.704 < 1.01^{100} < e < 1.01^{101} < 2.732$. The average of 1.01^{100} and 1.01^{101}, $2.71833\cdots$, is an even better estimate (the geometric mean, $2.71830\cdots$, is better yet). The exact numerical value of e is $2.718281828\cdots$ (the decimal does not continue to repeat; e is an irrational number, in fact transcendental – see III.9.3.1.). There are more efficient ways to compute e using infinite series; see V.5.5.3.–V.5.5.5..

V.5.4. Derivatives of Exponential and Logarithm Functions

V.5.4.1. THEOREM. The function $f(x) = e^x$ is differentiable, and $f'(x) = e^x$.

V.5.4.2. LEMMA. If $f(x) = e^x$, then f is differentiable at 0 and $f'(0) = \lim_{h\to 0} \frac{e^h-1}{h} = 1$.

PROOF: Let $x = e^h - 1$. Then $h = \ln(1+x)$ and $x \to 0$ as $h \to 0$ (i.e. $\lim_{h\to 0} e^h - 1 = 0$ and $\lim_{x\to 0} \ln(1+x) = 0$ by continuity of e^x and $\ln x$), and $x > 0$ if and only if $h > 0$. Thus

$$\lim_{h\to 0+} \frac{e^h-1}{h} = \lim_{x\to 0+} \frac{x}{\ln(1+x)} = \lim_{x\to 0+} \frac{1}{\frac{1}{x}\ln(1+x)} = \lim_{x\to 0+} \frac{1}{\ln\left([1+x]^{1/x}\right)} = \frac{1}{\ln e} = 1$$

since $\lim_{x\to 0+}[1+x]^{1/x} = e$ and $\ln x$ is continuous at e. We also have

$$\lim_{h\to 0-} \frac{e^h-1}{h} = \lim_{h\to 0+} \frac{e^{-h}-1}{-h} = \lim_{h\to 0+} \frac{e^{-h}(1-e^h)}{-h} = \left[\lim_{h\to 0+} e^{-h}\right]\left[\lim_{h\to 0+} \frac{e^h-1}{h}\right] = 1$$



We now prove Theorem V.5.4.1..
PROOF: Fix x. Then

$$f'(x) = \lim_{h\to 0} \frac{e^{x+h}-e^x}{h} = \lim_{h\to 0} \frac{e^x e^h - e^x}{h} = \lim_{h\to 0} \frac{e^x(e^h-1)}{h} = e^x \lim_{h\to 0} \frac{e^h-1}{h} = e^x$$



V.5.4.3. COROLLARY. Let $b > 0$, and $f(x) = b^x$. Then f is differentiable everywhere and $f'(x) = (\ln b)b^x$ for all x.

PROOF: Since $b^x = e^{x\ln b}$, the result follows from V.5.4.1. and the Chain Rule. 

V.5.4.4. COROLLARY. If $b > 0$, $b \neq 1$, then $f(x) = \log_b(x)$ is differentiable on $(0, \infty)$, and $\frac{d}{dx}(\log_b(x)) = \frac{1}{x \ln b}$. In particular, $\frac{d}{dx}(\ln x) = \frac{1}{x}$.

PROOF: This follows immediately from V.5.4.3. and the Inverse Function Theorem:

$$\frac{d}{dx}(\log_b x) = \frac{1}{(\ln b) b^{\log_b x}} = \frac{1}{x \ln b}.$$

V.5.4.5. COROLLARY. For all $r \in \mathbb{R}$, the function $f(x) = x^r$ is differentiable (hence continuous) on $(0, \infty)$, and the formula $\frac{d}{dx}(x^r) = rx^{r-1}$ holds on $(0, \infty)$.

PROOF: Since $x^r = e^{r \ln x}$ for $x > 0$, f is differentiable by V.5.4.1., V.5.4.4., and the Chain Rule, and we have

$$\frac{d}{dx}(x^r) = \frac{d}{dx}\left(e^{r \ln x}\right) = e^{r \ln x} \frac{r}{x} = x^r \frac{r}{x} = rx^{r-1}.$$

V.5.5. Exercises

V.5.5.1. (a) Write out the details of the proof of V.5.1.2.(ii).
(b) Prove the rules of exponents in V.5.1.2.(iii), using arguments similar to those in (i). (Note that the first rule has already been proved in (ii).)

V.5.5.2. (a) Prove the assertion in the first sentence in the proof of V.5.2.9..
(b) Prove the assertion in the second sentence of the proof of V.5.2.9. by induction, using the statement $P(n)$: "For every $u, v, w \in \mathbb{R}$ with $w < v < u$ and $\frac{v-w}{u-w} = \frac{k}{2^n}$ for some $k \in \mathbb{N}$, $k < 2^n$, $\frac{b^v - b^w}{v-w} < \frac{b^u - b^w}{u-w}$."
To prove $P(n+1)$ for a set u, v, w, apply $P(n)$ to either the first or second half of the interval $[w, u]$ depending where v is.

V.5.5.3. Here is an alternate approach to the definition of e (cf. [?, III.4].) Let $x_n = \left(1 + \frac{1}{n}\right)^n$.
(a) Using the Binomial Theorem, show that

$$x_n = 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \frac{1}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots +$$

$$+ \frac{1}{k!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right) + \cdots + \frac{1}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-1}{n}\right)$$

(b) Using (a), show that $x_n < x_{n+1}$ for any $n \in \mathbb{N}$.
(c) Let $y_n = \sum_{k=0}^{n} \frac{1}{k!}$. Show that $x_n < y_n$ for all n.
(d) Since $k! > 2^{k-1}$ for all $k > 2$, show that $y_n < 2 + \sum_{k=2}^{n} \frac{1}{2^{k-1}} < 3$ for $n > 2$. Thus the sequence (x_n) is a bounded increasing sequence, hence convergent. Call the limit e. Show that $2 < e < 3$.

V.5.5.4. Use the notation of V.5.5.3..
(a) If $1 < k < n$, set

$$z_{k,n} = 1 + 1 + \frac{1}{2!}\left(1 - \frac{1}{n}\right) + \frac{1}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots +$$

$$+ \frac{1}{k!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right)$$

Show that $z_{k,n} < x_n$, so $z_{k,n} < e$.
(b) Fix k. Show $\lim_{n\to\infty} z_{k,n} = y_k$. Conclude that $y_k \le e$ for all k. (In fact, since $y_k < y_{k+1} \le e$, we have $y_k < e$ for all k.)
(c) Since $x_k < y_k < e$ for all k, and $\lim_{k\to\infty} x_k = e$, we have $\lim_{k\to\infty} y_k = e$.

V.5.5.5. Use the notation of V.5.5.3.–V.5.5.4.. The sequence (y_n) converges to e fairly rapidly (unlike (x_n), which converges rather slowly):
(a) For each n and each $m > 1$, show that

$$y_{n+m} - y_n = \frac{1}{(n+1)!}\left(1 + \sum_{k=2}^{m} \frac{1}{(n+2)\cdots(n+k)}\right)$$

$$< \frac{1}{(n+1)!}\left(1 + \sum_{k=2}^{\infty} \frac{1}{(n+2)\cdots(n+k)}\right)$$

and thus, since $\lim_{m\to\infty} y_{n+m} = e$, $e - y_n \le \frac{1}{(n+1)!}\left(1 + \sum_{k=2}^{\infty} \frac{1}{(n+2)\cdots(n+k)}\right)$.
(b) By comparing terms and using $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for $|r| < 1$ (),

$$e - y_n \le \frac{1}{(n+1)!}\left(1 + \sum_{k=2}^{\infty} \frac{1}{(n+2)\cdots(n+k)}\right)$$

$$< \frac{1}{(n+1)!}\sum_{k=0}^{\infty} \frac{1}{(n+2)^k} = \frac{1}{(n+1)!}\frac{n+2}{n+1} = \frac{1}{n!}\frac{n+2}{(n+1)^2}$$

Thus, for example, $e - y_{10} < \frac{12}{10!\cdot 11^2} \approx 2.7 \times 10^{-8}$, so y_{10} agrees with e to 7 decimal places. $e - y_{100} < \frac{102}{100!\cdot 101^2} \approx 10^{-156}$, so y_{100} (which can be quickly computed on a programmable hand calculator) agrees with e to at least 155 decimal places. (In contrast, $x_{100} = 2.704\cdots$ only agrees with e to one decimal place.)
(c) Since $\frac{n(n+2)}{(n+1)^2} = \frac{n^2+2n}{n^2+2n+1} < 1$, we have $y_n < e < y_n + \frac{1}{n!\cdot n}$. We can write $e = y_n + \frac{\theta_n}{n!\cdot n}$ for some θ_n, $0 < \theta_n < 1$.
(d) Using (b) and $y_{n+1} < e$, show that

$$\frac{n}{n+1} < \theta_n < \frac{n^2+2n}{n^2+2n+1}$$

and thus $\theta_n \to 1$ as $n \to \infty$.

V.5.5.6. Continue to use the notation of V.5.5.3.–V.5.5.5.. Show that e is irrational as follows. Suppose $e = \frac{m}{n}$. Then

$$\frac{m}{n} = 1 + \frac{1}{2!} + \cdots + \frac{1}{n!} + \frac{\theta_n}{n! \cdot n}$$

Multiply both sides by $n!$ to conclude that $\frac{\theta_n}{n}$ is an integer. This is a contradiction since $0 < \theta_n < 1$.

V.5.5.7. Compute the derivative of $f(x) = \log x$ directly as follows. Fix $a > 0$.

(a) Write

$$\frac{f(a+h) - f(a)}{h} = \frac{1}{h} \log\left(\frac{a+h}{a}\right) = \frac{1}{h} \log\left(1 + \frac{h}{a}\right) \ .$$

(b) Make the substitution $x = \frac{a}{h}$, so that $\frac{1}{h} = \frac{x}{a}$, and note that

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} = \lim_{x \to \pm\infty} \frac{x}{a} \log\left(1 + \frac{1}{x}\right) = \lim_{x \to \pm\infty} \frac{\log\left(\left[1 + \frac{1}{x}\right]^x\right)}{a}$$

if the limits both exist and are equal.

(c) Use V.5.3.5. and the continuity of the logarithm function to conclude that $f'(a)$ exists and

$$f'(a) = \frac{\log e}{a} = \frac{1}{a} \ .$$

(The limits at $+\infty$ and $-\infty$ must also be calculated separately.)

V.5.5.8. Using V.5.5.7. and the Chain Rule, use the following argument from [Gri21] to derive the Product and Quotient Rules. Suppose f and g are differentiable.

(a) Set $h = fg$. From $\log|h| = \log|f| + \log|g|$, conclude that $\frac{h'}{h} = \frac{f'}{f} + \frac{g'}{g}$.

(b) Set $h = f/g$. From $\log|h| = \log|f| - \log|g|$, conclude that $\frac{h'}{h} = \frac{f'}{f} - \frac{g'}{g}$.

(c) Solve each equation for h' to obtain the Product and Quotient Rules.

Is this argument an adequate proof of the rules?

V.5.5.9. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function with $f(x+y) = f(x) + f(y)$ for all $x, y \in \mathbb{R}$.

(a) Let $a = f(1)$. Show that $f(n) = an$ for all $n \in \mathbb{Z}$.

(b) Show that $f(r) = ar$ for all $r \in \mathbb{Q}$.

(c) Use continuity to show that $f(x) = ax$ for all $x \in \mathbb{R}$. Thus f is linear.

(d) The continuity hypothesis cannot be removed. (However, the continuity hypothesis can be relaxed to Lebesgue measurability; see **??**.) $\mathbb{R}$ is a vector space over $\mathbb{Q}$ of dimension $2^{\aleph_0}$. Let B be a basis for $\mathbb{R}$ over $\mathbb{Q}$. Then every function from B to $\mathbb{R}$ extends to a $\mathbb{Q}$-linear (hence additive) function from $\mathbb{R}$ to $\mathbb{R}$. There are $2^{2^{\aleph_0}}$ functions from B to $\mathbb{R}$. But there are only $2^{\aleph_0}$ $\mathbb{R}$-linear functions from $\mathbb{R}$ to $\mathbb{R}$. Using the Axiom of Choice, construct a discontinuous example.

V.5.5.10. Let g be a continuous function from $\mathbb{R}$ to $\mathbb{R}$ which is not identically 0 and with the property that $g(x+y) = g(x)g(y)$ for all $x, y \in \mathbb{R}$.
(a) Show that g takes only positive values. [$g(x) = g(x/2)^2$.]
(b) Show using Exercise V.5.5.9. that $g(x) = b^x$ for all x, where $b = g(1) > 0$. So g is an exponential function. [Consider $f(x) = \log g(x)$.]
(c) The continuity hypothesis cannot be removed here either: construct a discontinuous example.
(d) Prove similarly that a continuous function $h : (0, \infty) \to \mathbb{R}$ which is not identically 0 and which has the property that $h(xy) = h(x) + h(y)$ for all $x, y > 0$ must be a logarithm function.

V.5.5.11. Redevelop the entire theory of exponential and logarithm functions by first defining $\log x = \int_1^x \frac{1}{t}\, dt$ for $x > 0$ and then letting $f(x) = e^x$ be the inverse function of $\log x$.

V.5.5.12. Here is an alternate approach to proving differentiability of exponential functions. Let $b > 1$. Then $g_b(x) = b^x$ is continuous on $\mathbb{R}$ (**??**).
(a) If $k \in \mathbb{R}$, $\int_{-k-1}^{-k} b^t\, dt < b^{-k}$. Thus, if $n \in \mathbb{N}$, $\int_{-n}^{0} b^t\, dt < \sum_{k=0}^{n-1} b^{-k} = \frac{1-b^{-n}}{1-b^{-1}}$.
(b) Conclude that the improper integral $\int_{-\infty}^{0} b^t\, dt$ converges, and hence $\int_{-\infty}^{x} b^t\, dt$ converges for any x. For $x \in \mathbb{R}$, set $f_b(x) = \int_{-\infty}^{x} b^t\, dt$. Show that f_b is differentiable and $f_b'(x) = g_b(x) = b^x$ for all x.
(c) By changing variables in the improper integral, show that $f_b(x) = b^x \cdot f_b(0)$ for any x. Thus g_b is infinitely differentiable, and $g_b' = k_b g_b$, where $\frac{1}{k_b} = \int_{-\infty}^{0} b^t\, dt$.
(d) If $a, b > 1$, apply the product rule to $g_{ab} = g_a g_b$ to conclude that $k_{ab} = k_a + k_b$. Prove that $\ell(b) = k_b$ is continuous. Thus ℓ is a logarithm function.

V.6. Trigonometric Functions

There is a dark secret behind the usual (precalculus) development of the theory of trigonometric functions: the entire theory is based on logical quicksand. All of trigonometry is based on the assumption that a circle has a finite "circumference" ("arc length"), which is never really defined; the number π is "defined" to be the ratio of the "circumference" to the diameter, and empirically calculated to be a little more than three. This assumption, while eminently believable and experimentally verifiable, is logically untenable, especially in the absence of a real definition of arc length. The notion of arc length can only be rigorously defined using limits (). See I.6.1.5.ff. for a fuller discussion.

In this section, we will give a rigorous and efficient treatment of trigonometric functions, the properties and identities they satisfy, and their calculus. We will actually define one of the inverse trigonometric functions first, an approach which may seem oddly backwards to readers familiar with the traditional approach, but which has the advantage of being logically impeccable.

V.6.1. Arc Length of a Circle and the Arcsine Function

We will consider the unit circle $\{(x, y) : x^2 + y^2 = 1\}$ in $\mathbb{R}^2$.

V.6.1.1. PROPOSITION. The arc length of the circle is finite.

To prove this, consider the arc from $(0, 1)$ to $(1, 0)$, which is a quarter circle. By symmetry, it is enough to prove that this arc has finite length. We give two arguments.

V.6.1.2. First consider the arc from $(0, 1)$ to $(b, \sqrt{1 - b^2})$ for $0 < b < 1$. This arc is the graph of the continuously differentiable function $y = \sqrt{1 - x^2}$ on $[0, b]$, so by **??** the arc length is finite and given by

$$L(0, b) = \int_0^b \sqrt{1 + \frac{x^2}{1 - x^2}}\, dx = \int_0^b \frac{1}{\sqrt{1 - x^2}}\, dx$$

Thus, for example, the arc from $(0, 1)$ to $(\frac{4}{5}, \frac{3}{5})$ has finite length. Since this arc is congruent to the arc from $(\frac{3}{5}, \frac{4}{5})$ to $(1, 0)$ (via interchanging x and y), this arc also has finite length, and the quarter circle is the union, hence also has finite length which we will call $L(0, 1)$.

V.6.1.3. We give an alternate argument which gives a (rather crude) estimate of the length. As $b \to 1$, $L(0, b) \to L(0, 1)$ (**??**), and $L(0, b)$ is increasing as b increases, so we only need show that $\{L(0, b) : 0 \le b < 1\}$ is bounded. But if $0 < x < 1$, $x^2 < x$, so $\frac{1}{\sqrt{1-x^2}} < \frac{1}{\sqrt{1-x}}$, and

$$L(0, b) = \int_0^b \frac{1}{\sqrt{1 - x^2}}\, dx < \int_0^b \frac{1}{\sqrt{1 - x}}\, dx = -2\sqrt{1 - x}\Big|_0^b = 2 - 2\sqrt{1 - b} < 2$$

for all $b < 1$; thus $L(0, 1)$ is finite and less than 2.

V.6.1.4. To get a lower bound for $L(0, 1)$, since it is the length of an arc from $(0, 1)$ to $(1, 0)$, it must be greater than the straight-line distance between the points, which is $\sqrt{2}$. A better lower bound, for example, is the sum of the distance from $(0, 1)$ to $(\frac{4}{5}, \frac{3}{5})$ and the distance from $(\frac{4}{5}, \frac{3}{5})$ to $(1, 0)$, which is $\sqrt{\frac{4}{5}} + \sqrt{\frac{2}{5}} = 1.52688\cdots$.

V.6.1.5. DEFINITION. The number π is the arc length of half the unit circle.

By symmetry, $\pi = 2L(0,1)$. By the above estimates, we have $2(\sqrt{\frac{4}{5}}+\sqrt{\frac{2}{5}}) = 3.05\cdots < \pi < 4$. Of course, it is well known that the exact value of π is $3.1415926535\cdots$; we will discuss efficient ways to calculate π later (**??**).

V.6.1.6. DEFINITION. If $-1 \le x \le 1$, the *arcsine* of x is the number

$$\arcsin x = \int_0^x \frac{1}{\sqrt{1-t^2}}\,dt$$

(If $x = \pm 1$, this is an improper integral; we can define $\arcsin 1 = L(0,1) = \frac{\pi}{2}$ and $\arcsin(-1) = -L(0,1) = -\frac{\pi}{2}$.)

Thus $\arcsin x$ is the antiderivative of $\frac{1}{\sqrt{1-x^2}}$, i.e. by the Fundamental Theorem of Calculus, $\frac{d}{dx}(\arcsin x) = \frac{1}{\sqrt{1-x^2}}$ for $-1 < x < 1$. Thus $f(x) = \arcsin x$ is continuous on $[-1,1]$ and differentiable on $(-1,1)$, and $f'(x) = \frac{1}{\sqrt{1-x^2}} > 0$ on $(-1,1)$, so $\arcsin x$ is strictly increasing on $[-1,1]$. We have $\arcsin 0 = 0$ and $\arcsin(-x) = -\arcsin x$ for $-1 \le x \le 1$.

V.6.2. The Sine and Cosine Functions

Since $f(x) = \arcsin x$ is continuous and strictly increasing on $[-1,1]$, with range $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, it has a continuous inverse function $g(x)$ on $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$.

V.6.2.1. DEFINITION. If $x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, the *sine* of x is $\sin x = g(x)$, and the *cosine* of x is $\cos x = \sqrt{1-g(x)^2} = \sqrt{1-(\sin x)^2}$.

We usually write $\sin^2 x$ instead of $(\sin x)^2$, i.e. $\cos x = \sqrt{1-\sin^2 x}$ for $-\frac{\pi}{2} \le x \le \frac{\pi}{2}$. Thus $\sin^2 x + \cos^2 x = 1$ for all $x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ (of course, $\cos^2 x = (\cos x)^2$.)

V.6.2.2. Since $\arcsin(-x) = -\arcsin(x)$, $\sin x$ has the same property:

$$\sin(-x) = -\sin x$$

$$\cos(-x) = \cos x$$

We have $\sin 0 = 0$, $\cos 0 = 1$, $\sin(\frac{\pi}{2}) = 1$, $\cos(\frac{\pi}{2}) = 0$, $\sin(-\frac{\pi}{2}) = -1$, $\cos(-\frac{\pi}{2}) = 0$.

V.6.2.3. From the Inverse Function Theorem, $g(x) = \sin x$ is differentiable on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and

$$g'(x) = \frac{1}{f'(g(x))} = \frac{1}{\frac{1}{\sqrt{1-\sin^2 x}}} = \sqrt{1-\sin^2 x} = \cos x$$

and from the Chain Rule $h(x) = \cos x$ is differentiable on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and

$$h'(x) = \frac{d}{dx}\left(\sqrt{1-\sin^2 x}\right) = \frac{1}{2}\left(\sqrt{1-\sin^2 x}\right)^{-1/2}(-2\sin x\cos x)$$

$$= -\frac{\sin x\cos x}{\sqrt{1-\sin^2 x}} = -\frac{\sin x\cos x}{\cos x} = -\sin x$$

and thus we obtain:

V.6.2.4. THEOREM. The functions $\sin x$ and $\cos x$ are infinitely differentiable on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$, and

$$\frac{d}{dx}(\sin x) = \cos x\,, \quad \frac{d}{dx}(\cos x) = -\sin x$$

for $-\frac{\pi}{2} < x < \frac{\pi}{2}$.

V.6.2.5. By the properties of inverse functions, if $0 \le x \le \frac{\pi}{2}$, $\sin x = b$ means

$$x = \arcsin b = L(0, b) = \int_0^b \frac{1}{\sqrt{1-t^2}}\, dt$$

so the point P on the circle (in the first quadrant) for which the arc from $(0,1)$ to P has length x has first coordinate $\sin x$. The second coordinate is then $\sqrt{1-\sin^2 x} = \cos x$, i.e. $P = (\sin x, \cos x)$. It is more customary to measure arcs of the circle counterclockwise from $(1,0)$, and then the coordinates of the point of arc length x are $(\cos x, \sin x)$ (just apply the transformation interchanging x and y.)

V.6.2.6. We extend the definition of $\sin x$ and $\cos x$ to all real numbers by taking $(\cos x, \sin x)$ to be the coordinates of the point of arc length x traveling counterclockwise around the circle from $(1,0)$ (clockwise a distance $|x|$ if $x < 0$). Since the circumference of the circle is 2π, we obtain that sin and cos are periodic with period 2π: for all x,

$$\sin(x + 2\pi) = \sin x\,, \quad \cos(x + 2\pi) = \cos x$$

It can in fact be easily checked that, for all x,

$$\sin(-x) = -\sin x\,, \quad \cos(-x) = \cos x$$

$$\sin\left(x + \frac{\pi}{2}\right) = \cos x\,, \quad \cos\left(x + \frac{\pi}{2}\right) = -\sin x$$

since when the point (a, b) is rotated a quarter circle counterclockwise, its coordinates become $(-b, a)$. Then replacing x by $x - \frac{\pi}{2}$,

$$\sin\left(x - \frac{\pi}{2}\right) = -\cos x\,, \quad \cos\left(x - \frac{\pi}{2}\right) = \sin x$$

Repeating the process twice, we obtain

$$\sin(x \pm \pi) = -\sin x\,, \quad \cos(x \pm \pi) = -\cos x$$

for all x. Of course, $|\sin x| \le 1$ and $|\cos x| \le 1$ for all x.

V.6.2.7. THEOREM. The functions $\sin x$ and $\cos x$ are infinitely differentiable on all of $\mathbb{R}$, and

$$\frac{d}{dx}(\sin x) = \cos x\,, \quad \frac{d}{dx}(\cos x) = -\sin x$$

for all $x \in \mathbb{R}$.

PROOF: By V.6.2.4., the result holds on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Then, since $\sin x = \cos\left(x - \frac{\pi}{2}\right)$, $\sin x$ is differentiable on $(0, \pi)$, and by the chain rule

$$\frac{d}{dx}(\sin x) = \frac{d}{dx}\left(\cos\left(x - \frac{\pi}{2}\right)\right) = -\sin\left(x - \frac{\pi}{2}\right) = \cos x$$

on $(0, \pi)$, and similarly $\cos x$ is differentiable on $(0, \pi)$ with derivative $-\sin x$. From $\sin x = -\cos\left(x + \frac{\pi}{2}\right)$ we similarly get that $\sin x$ is differentiable on $(-\pi, 0)$ with derivative $\cos x$, and similarly for $\cos x$. Thus the result holds on the interval $(-\pi, \pi)$. Since $\sin(x \pm \pi) = -\sin x$ and $\cos(x \pm \pi) = -\cos x$ for all x, the result holds on $((n-1)\pi, (n+1)\pi)$ for any $n \in \mathbb{Z}$, and hence for the whole real line. 👍

V.6.3. Other Trigonometric Functions

V.6.3.1. DEFINITION. If $x \in \mathbb{R}$, define the *tangent*, *cotangent*, *secant*, and *cosecant* of x, respectively denoted $\tan x$, $\cot x$, $\sec x$, $\csc x$, by

$$\tan x = \frac{\sin x}{\cos x}, \quad \cot x = \frac{\cos x}{\sin x} = \frac{1}{\tan x}$$

$$\sec x = \frac{1}{\cos x}, \quad \csc x = \frac{1}{\sin x}$$

V.6.3.2. These functions are defined wherever the denominators are nonzero. Thus the tangent and secant are defined except at $\pm\frac{\pi}{2} \pm n\pi$, and the cotangent and cosecant are defined except at (integer) multiples of π. We have

$$\lim_{x \to \frac{\pi}{2}-} \tan x = +\infty, \quad \lim_{x \to -\frac{\pi}{2}+} \tan x = -\infty$$

$$\lim_{x \to \frac{\pi}{2}-} \sec x = +\infty, \quad \lim_{x \to -\frac{\pi}{2}+} \sec x = +\infty$$

$$\lim_{x \to \pi-} \cot x = -\infty, \quad \lim_{x \to 0+} \cot x = +\infty$$

$$\lim_{x \to \pi-} \csc x = +\infty, \quad \lim_{x \to 0+} \csc x = +\infty$$

Since $|\sin x|$ and $|\cos x|$ are always ≤ 1, we have $|\sec x| \geq 1$ and $|\csc x| \geq 1$ whenever defined.

The functions $\sec x$ and $\csc x$ are periodic with period 2π. But since $\sin(x+\pi) = -\sin x$ and $\cos(x+\pi) = -\cos x$, the functions $\tan x$ and $\cot x$ are actually periodic with period π.

If $0 \leq x < \frac{\pi}{2}$, the number $\tan x$ is the slope of the line through the origin and the point $(\cos x, \sin x)$ in the first quadrant whose arc length along the unit circle from $(1, 0)$ is x.

V.6.3.3. These functions, where defined, satisfy the identities

$$1 + \tan^2 x = \frac{\cos^2 x}{\cos^2 x} + \frac{\sin^2 x}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$$

and similarly $1 + \cot^2 x = \csc^2 x$ (where, as usual, $\tan^2 x = (\tan x)^2$, etc.)

V.6.3.4. These functions are infinitely differentiable on their domains, and their derivatives can be calculated using the quotient rule and power rule:

$$\frac{d}{dx}(\tan x) = \frac{(\cos x)(\cos x) - (\sin x)(-\sin x)}{\cos^2 x} = \frac{1}{\cos^2 x} = \sec^2 x$$

$$\frac{d}{dx}(\cot x) = \frac{(\sin x)(-\sin x) - (\cos x)(\cos x)}{\sin^2 x} = -\frac{1}{\sin^2 x} = -\csc^2 x$$

$$\frac{d}{dx}(\sec x) = \frac{d}{dx}\left((\cos x)^{-1}\right) = (-1)(\cos x)^{-2}(-\sin x) = \frac{\sin x}{\cos^2 x} = \sec x \tan x$$

$$\frac{d}{dx}(\csc x) = \frac{d}{dx}\left((\sin x)^{-1}\right) = (-1)(\sin x)^{-2}(\cos x) = -\frac{\cos x}{\sin^2 x} = -\csc x \cot x$$

V.6.4. The Inverse Trigonometric Functions

We have already defined an inverse function, $\arcsin x$, for $\sin x$. Actually, if we start with the sine function and try to define an inverse, since $\sin x$ is periodic and not one-to-one, we must restrict it to an interval on which it is one-to-one; it is conventional to take the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Similarly, for $\cos x$ we conventionally take the interval $[0, \pi]$, which is mapped one-one onto $[-1, 1]$.

V.6.4.1. DEFINITION. The inverse function to $\cos x$ with domain $[-1, 1]$ and range $[0, \pi]$ is called the *arccosine* function, denoted $\arccos x$.

V.6.4.2. Since the function $g(x) = \cos x$ is continuous on $[0, \pi]$ and differentiable on $(0, \pi)$ with nonzero derivative, the function $f(x) = \arccos x$ is continuous on $[-1, 1]$ and differentiable on $(-1, 1)$ with derivative given by the Inverse Function Theorem:

$$\frac{d}{dx}(\arccos x) = f'(x) = \frac{1}{g'(f(x))} = \frac{1}{-\sin(\arccos x)}$$

$$= -\frac{1}{\sqrt{1 - \cos^2(\arccos x)}} = -\frac{1}{\sqrt{1 - x^2}}$$

Actually, from $\sin x = \cos\left(\frac{\pi}{2} - x\right)$ it follows that $\arccos x = \frac{\pi}{2} - \arcsin x$.

V.6.4.3. Now consider $\tan x$. The interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ is one period of the tangent, and $\tan x$ maps this interval one-one onto $(-\infty, \infty)$, so we conventionally take the inverse function for the tangent function to have this range:

V.6.4.4. DEFINITION. The inverse function to $\tan x$ with domain $(-\infty, \infty)$ and range $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ is called the *arctangent* function, denoted $\arctan x$.

Since the function $g(x) = \tan x$ is differentiable on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ with nonzero derivative, $f(x) = \arctan x$ is differentiable on $(-\infty, \infty)$ with derivative

$$\frac{d}{dx}(\arctan x) = f'(x) = \frac{1}{g'(f(x))} = \frac{1}{\sec^2(\arctan x)} = \frac{1}{1 + \tan^2(\arctan x)} = \frac{1}{1 + x^2}$$

V.6.4.5. We can similarly define an *arccotangent* function with domain $(-\infty, \infty)$, range $(0, \pi)$, and derivative $-\frac{1}{1+x^2}$. Actually, $\operatorname{arccot} x = \frac{\pi}{2} - \arctan x$ since $\cot x = \tan\left(\frac{\pi}{2} - x\right)$.

We could also define arcsecant and arccosecant functions, but these are rarely used.

It is very common to write $\sin^{-1} x$ for $\arcsin x$, etc. This notation is consistent with the usual notation for inverse functions, but incompatible with the convention of writing $\sin^2 x$ for $(\sin x)^2$, etc., and thus can be confusing. (Note that $\sin^{-1} x \neq (\sin x)^{-1}$!)

> "$\sin^2 \phi$ is odious to me, even though Laplace made use of it; should it be feared that $\sin \phi^2$ might become ambiguous, which would perhaps never occur, or at most very rarely when speaking of $\sin(\phi^2)$, well then, let us write $(\sin \phi)^2$, but not $\sin^2 \phi$, which by analogy should signify $\sin(\sin \phi)$."
>
> *C. F. Gauss*[12]

V.6.5. The Area of a Circle

V.6.5.1. We can use trigonometry to calculate the area of a circle. Actually, we should first define what we mean by "area" of a subset of the plane. This is not as easy as it sounds: the general problem is essentially the basic subject matter of measure theory. It would take us too far afield to try to treat this question here (cf. ()), so we will just assume that area is defined for sufficiently well-behaved subsets of the plane, and satisfies

(i) If f is a nonnegative continuous function on a closed bounded interval $[a, b]$, then the region $R = \{(x, y) : a \leq x \leq b, 0 \leq y \leq f(x)\}$ has area $A(R) = \int_a^b f(x)\, dx$.

(ii) If $R \subseteq S$, then $A(R) \leq A(S)$.

(iii) If $R_1, \ldots, R_n$ are nonoverlapping regions ("nonoverlapping" means the intersection of any two is contained in the boundary of both), then

$$A(R_1 \cup \cdots \cup R_n) = A(R_1) + \cdots + A(R_n).$$

(iv) Congruent regions have the same area.

(There is actually a subtle technical problem with (iii), since the boundary of a region can have positive measure, but this will not cause difficulty if we consider only regions bounded by finitely many smooth curves.)

A circle (technically a disk) or any wedge of a circle has a well-defined area under these assumptions. If D is the unit disk, then $A(D) = 2\int_{-1}^{1} \sqrt{1 - x^2}\, dx < 4$. We want to calculate the exact value of $A(D)$.

V.6.5.2. THEOREM. $A(D) = \pi$.

The simplest way to calculate this area is using vector calculus, but we will proceed in an alternate way. We have that $A(D) = \pi p$ for some $p > 0$; we want to show that $p = 1$.

[12] [Mor93, 1886].

V.6.5.3. Define a function $\phi : [0, 2\pi] \to \mathbb{R}$ by setting $\phi(x)$ equal to the area of a wedge of a circle with arc length x. Any two such wedges are congruent, so $\phi(x)$ is well defined, and strictly increasing; we have $\phi(0) = 0$ and $\phi(2\pi) = A(D) = \pi p$. For each $n \in \mathbb{N}$, we have $\phi\left(\frac{2\pi}{n}\right) = \frac{\pi p}{n}$ by (iii), i.e. $\phi(x) = \frac{p}{2}x$ if $x = \frac{2\pi}{n}$ for some n. (Actually this formula holds for all $x \in [0, 2\pi]$, although we will not need this fact for the Proof: it holds for $x = \frac{2\pi m}{n}$ for $m \leq n$ by (iii), and hence for all x since such numbers are dense in $[0, 2\pi]$ and ϕ is strictly increasing.)

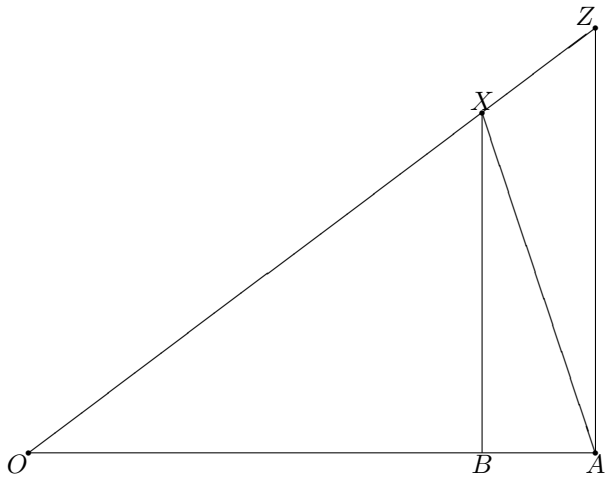


Figure V.8:

If $0 < x < \frac{\pi}{2}$, draw the picture in Figure 1, with $\overline{OA} = 1$, $\overline{OX} = 1$, and $\angle AOX = x$, so $X = (\cos x, \sin x)$. Then $\overline{BX} = \sin x$ and $\overline{AZ} = \tan x$. Let T_1 be the triangle $\triangle AOX$ and T_2 the triangle $\triangle AOZ$. Then $A(T_1) \leq \phi(x) \leq A(T_2)$, so if $x_n = \frac{2\pi}{n}$ for $n > 4$ we have

$$\frac{1}{2}\sin x_n \leq \frac{p}{2}x_n \leq \frac{1}{2}\tan x_n$$

$$\frac{\sin x_n}{x_n} \leq p \leq \frac{\tan x_n}{x_n}$$

Since $\lim_{n\to\infty} \frac{\sin x_n}{x_n} = \lim_{n\to\infty} \frac{\tan x_n}{x_n} = 1$ (), we have $p = 1$. This completes the proof of the theorem.

V.6.5.4. It follows from the argument of the proof that $\phi(x) = \frac{1}{2}x$ for all x, giving an alternate geometric interpretation of the angle x as area rather than arc length. See Exercise () for another derivation of this formula.

V.6.6. The Hyperbolic Functions

The trigonometric functions are sometimes called the *circular functions*, for obvious reasons. We can define analogous functions related to a hyperbola which are formally similar to the circular functions; it turns out that the connections are much more than formal, although these deep connections are beyond the scope of this chapter (cf. XII.1.1.8.).

There are actually two apparently totally different ways the hyperbolic functions can be defined, one using the hyperbola in the way the circle was used for the trigonometric functions, and the other using exponential functions. It is remarkable that the same functions are obtained by the two methods. We will describe both methods in this section.

V.6.6.1. We begin with the unit hyperbola $y^2 - x^2 = 1$. (It is more customary to use the hyperbola $x^2 - y^2 = 1$ with base point $(1, 0)$, but our approach facilitates computations.) Since the upper half of the hyperbola has equation $y = \sqrt{1 + x^2}$, in analogy with the formula defining $\arcsin x$ we define the *hyperbolic arcsine* $\operatorname{arcsinh} x$ by

$$\operatorname{arcsinh} x = \int_0^x \frac{1}{\sqrt{1 + t^2}}\, dt$$

Unfortunately, there is no interpretation of this as arc length in the hyperbolic case. (See Exercise () for a connection with area similar to the one for the trigonometric functions.)

Then $\operatorname{arcsinh} x$ is defined for all x and is differentiable on $(-\infty, \infty)$ with derivative $\frac{1}{\sqrt{1+x^2}}$. We obviously have $\operatorname{arcsinh}(-x) = -\operatorname{arcsinh} x$ for all x.

V.6.6.2. PROPOSITION. The range of $\operatorname{arcsinh} x$ is $(-\infty, \infty)$.

PROOF: Since $\lim_{x\to+\infty} \frac{x}{\sqrt{1+x^2}} = 1$, there is an n such that $\frac{1}{\sqrt{1+x^2}} > \frac{1}{2x}$ for $x > n$. Thus, if $b > n$, we have

$$\int_n^b \frac{1}{\sqrt{1 + x^2}}\, dx > \int_n^b \frac{1}{2x}\, dx = (\ln b - \ln n)$$

which goes to infinity as $b \to +\infty$. So $\operatorname{arcsinh} x$ becomes arbitrarily large as $x \to +\infty$. Since $\operatorname{arcsinh}(-x) = -\operatorname{arcsinh} x$ and the range is an interval, it must be the whole real line. 👍

V.6.6.3. So $f(x) = \operatorname{arcsinh} x$ is strictly increasing and has an inverse function which we call the *hyperbolic sine*, denoted $\sinh x$. By V.6.6.2., the domain of $\sinh x$ is all of $\mathbb{R}$. We then define the *hyperbolic cosine* $\cosh x$ to be $\sqrt{1 + \sinh^2 x}$, so $\cosh^2 x - \sinh^2 x = 1$ for all x. We have $\sinh(-x) = -\sinh x$ and $\cosh(-x) = \cosh x$ for all x, $\sinh 0 = 0$, and $\cosh 0 = 1$.

V.6.6.4. The function $g(x) = \sinh x$ is differentiable on $\mathbb{R}$, with derivative given by

$$\frac{d}{dx}(\sinh x) = g'(x) = \frac{1}{f'(g(x))} = \frac{1}{\frac{1}{\sqrt{1+sinh^2 x}}} = \sqrt{1 + \sinh^2 x} = \cosh x$$

and by the Chain Rule,

$$\begin{aligned}\frac{d}{dx}(\cosh x) = \frac{d}{dx}\left(\sqrt{1 + \sinh^2 x}\right) &= \frac{1}{2}(1 + \sinh^2 x)^{-1/2}(2 \sinh x \cosh x) \\ &= \frac{\sinh x \cosh x}{\sqrt{1 + \sinh^2 x}} = \frac{\sinh x \cosh x}{\cosh x} = \sinh x\end{aligned}$$

and thus $\sinh x$ and $\cosh x$ are infinitely differentiable.

The functions $\sinh x$ and $\cosh x$ thus behave formally in a similar way to $\sin x$ and $\cos x$, with occasional sign differences. The analogy only goes so far, however: $\sinh x$ and $\cosh x$ are not periodic and unbounded. In fact, $\cosh x \geq 1$ for all x.

V.6.6.5. We can define analogs of the other trigonometric functions, called the *hyperbolic tangent, cotangent, secant, and cosecant*:

$$\tanh x = \frac{\sinh x}{\cosh x}\,, \quad \coth x = \frac{\cosh x}{\sinh x} = \frac{1}{\tanh x}$$
$$\operatorname{sech} x = \frac{1}{\cosh x}\,, \quad \operatorname{csch} x = \frac{1}{\sinh x}$$

It is useful to have a one-syllable pronunciation of the names of the hyperbolic functions, since the names "hyperbolic sine," etc., are cumbersome. There is an obvious pronunciation for "cosh", and also for the seldom-used "coth" and "sech". The names "sinh" and "tanh" are normally pronounced "cinch" and "tanch" respectively. Your guess is as good as mine about how to pronounce "csch" (or "csc" for that matter!)

V.6.6.6. The domain of $\tanh x$ and $\operatorname{sech} x$ is all of $\mathbb{R}$; the domain of $\coth x$ and $\operatorname{csch} x$ is $\mathbb{R} \setminus \{0\}$. We have $\tanh 0 = 0$, $\tanh(-x) = -\tanh x$ and $|\tanh x| < 1$ for all x, and $\lim_{x\to+\infty} \tanh x = 1$, $\lim_{x\to-\infty} \tanh x = -1$. Similarly, $\operatorname{sech} 0 = 1$, $\operatorname{sech}(-x) = \operatorname{sech} x$ and $|\operatorname{sech} x| \le 1$ for all x, and $\lim_{x\to+\infty} \operatorname{sech} x = \lim_{x\to-\infty} \operatorname{sech} x = 0$. They satisfy identities similar to the ones for trigonometric functions; for example:

$$1 - \tanh^2 x = 1 - \frac{\sinh x}{\cosh x} = \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \frac{1}{\cosh^2 x} = \operatorname{sech}^2 x$$

V.6.6.7. These hyperbolic functions are all infinitely differentiable on their domains, with derivatives given by standard rules:

$$\frac{d}{dx}(\tanh x) = \frac{(\cosh x)(\cosh x) - (\sinh x)(\sinh x)}{\cosh^2 x} = \frac{1}{\cosh^2 x} = \operatorname{sech}^2 x$$
$$\frac{d}{dx}(\coth x) = \frac{(\sinh x)(\sinh x) - (\cosh x)(\cosh x)}{\sinh^2 x} = -\frac{1}{\sinh^2 x} = -\operatorname{csch}^2 x$$
$$\frac{d}{dx}(\operatorname{sech} x) = \frac{d}{dx}\left((\cosh x)^{-1}\right) = (-1)(\cosh x)^{-2}(\sinh x) = -\frac{\sinh x}{\cosh^2 x} = -\operatorname{sech} x \tanh x$$
$$\frac{d}{dx}(\operatorname{csch} x) = \frac{d}{dx}\left((\sinh x)^{-1}\right) = (-1)(\sinh x)^{-2}(\cosh x) = -\frac{\cosh x}{\sinh^2 x} = -\operatorname{csch} x \coth x$$

V.6.6.8. Other inverse hyperbolic functions can be defined. For example, the one most often used, the inverse hyperbolic tangent $f(x) = \operatorname{arctanh} x$ is the inverse function of $g(x) = \tanh x$, with domain $(-1, 1)$, range $(-\infty, \infty)$, and derivative

$$\frac{d}{dx}(\operatorname{arctanh} x) = f'(x) = \frac{1}{g'(f(x))} = \frac{1}{\operatorname{sech}^2(\operatorname{arctanh} x)}$$
$$= \frac{1}{1 - \tanh^2(\operatorname{arctanh} x)} = \frac{1}{1 - x^2}$$

Since $\frac{1}{1-x^2} = \frac{1}{2}\frac{1}{1+x} + \frac{1}{2}\frac{1}{1-x}$ by partial fractions, the derivative of $\operatorname{arctanh} x$ is the same as the derivative of $h(x) = \frac{1}{2}\ln(1+x) - \frac{1}{2}\ln(1-x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$. Since $h(0) = \operatorname{arctanh} 0 = 0$, we have $h(x) = \operatorname{arctanh} x$ for all x, i.e.

$$\operatorname{arctanh} x = \frac{1}{2}\ln(1+x) - \frac{1}{2}\ln(1-x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right) \quad (-1 < x < 1)$$

There is therefore a close connection between inverse hyperbolic functions and logarithm functions, and hence the hyperbolic functions are actually closely related to exponential functions. Since the connection is not so easy to motivate, we will simply define some new functions and show that they are the same as the hyperbolic functions.

V.6.6.9. DEFINITION. Define $s(x) = \frac{e^x - e^{-x}}{2}$ and $c(x) = \frac{e^x + e^{-x}}{2}$.

V.6.6.10. THEOREM. We have $s(x) = \sinh x$ and $c(x) = \cosh x$ for all x.

To prove this theorem, note that $s(x)$ and $c(x)$ are differentiable on all of $\mathbb{R}$, and that $s'(x) = c(x)$, $c'(x) = s(x)$. Since $c(x) > 0$ for all x, it follows that $s(x)$ is strictly increasing on $\mathbb{R}$. We also have that $s(0) = 0$, $s(-x) = -s(x)$ for all x, and $\lim_{x\to+\infty} s(x) = +\infty$, so it follows that the range of $s(x)$ is all of $\mathbb{R}$. We have, for all x,

$$1 + (s(x))^2 = 1 + \frac{(e^x - e^{-x})^2}{4} = \frac{4 + e^{2x} - 2 + e^{-2x}}{4}$$
$$= \frac{e^{2x} + 2 + e^{-2x}}{4} = \frac{(e^x + e^{-x})^2}{4} = (c(x))^2$$

and thus $c(x) = \sqrt{1 + (s(x))^2}$ for all x.

Thus $s(x)$ has an inverse function $h(x)$, which is defined and differentiable on all of $\mathbb{R}$, with derivative given by the Inverse Function Theorem:

$$h'(x) = \frac{1}{s'(h(x))} = \frac{1}{c(h(x))} = \frac{1}{\sqrt{1 + (s(h(x)))^2}} = \frac{1}{\sqrt{1 + x^2}}$$

so the derivative of $h(x)$ is the same as the derivative of $\operatorname{arcsinh} x$. Since $h(0) = 0 = \operatorname{arcsinh} 0$, we have $h(x) = \operatorname{arcsinh} x$ for all x, and thus the inverse functions are also equal: $s(x) = \sinh x$. The derivatives of these functions are then also equal: $c(x) = \cosh x$. This completes the proof of the theorem.

V.6.6.11. From this, we get formulas for the other hyperbolic functions:

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \coth x = \frac{\cosh x}{\sinh x} = \frac{e^x + e^{-x}}{e^x - e^{-x}}$$
$$\operatorname{sech} x = \frac{1}{\cosh x} = \frac{2}{e^x + e^{-x}}, \quad \operatorname{csch} x = \frac{1}{\sinh x} = \frac{2}{e^x - e^{-x}}$$

See Exercise () for a direct algebraic derivation of the formula for $\tanh x$.

V.6.6.12. Actually, the trigonometric functions also have similar formulas in terms of exponential functions, but the formulas involve complex numbers:

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}$$

(we will not attempt to explain these formulas here.) The rich and beautiful subject of complex analysis (calculus for functions on the complex numbers) reveals many deep and fascinating connections between trigonometric, hyperbolic, and exponential functions; see XII.1.1.8..

V.6.7. Exercises

V.6.7.1. If $0 \leq x \leq \frac{\pi}{2}$, show that the function $\phi(x)$ of the proof of V.6.5.2. is given by

$$\phi(x) = \frac{1}{2}\cos x \sin x + \int_{\cos x}^{1} \sqrt{1-t^2}\, dt$$

(see Figure ().) Using (), show that $\phi'(x)$ is the constant function $\frac{1}{2}$, so $\phi(x) = \frac{1}{2}x$.

V.6.7.2. If $x \geq 0$, let $\psi(x)$ be the area of the wedge bounded by the hyperbola $x^2 - y^2 = 1$, the line through the origin and $(\cosh x, \sinh x)$, and the x-axis (Figure ().) Show that $\psi(x)$ is given by

$$\psi(x) = \frac{1}{2}\cosh x \sinh x - \int_{1}^{\cosh x} \sqrt{t^2-1}\, dt$$

As in Exercise (), show that $\psi'(x)$ is the constant function $\frac{1}{2}$, so $\psi(x) = \frac{1}{2}x$.

V.6.7.3. Derive the exponential formula for $\tanh x$ from the formula $\operatorname{arctanh} x = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$ by solving the equation

$$2x = \ln\left(\frac{1+y}{1-y}\right)$$

for y.

V.6.7.4. [Bri98] Prove the sum formulas for sine and cosine

$$\sin(\alpha+\beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$$

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

for $\alpha, \beta \in \mathbb{R}$, as follows.

(a) Fix $\beta \in \mathbb{R}$, and replace α by a variable x, and define

$$f(x) = [\sin(x+\beta) - \sin x\cos\beta - \cos x\sin\beta]^2 + [\cos(x+\beta) - \cos x\cos\beta + \sin x\sin\beta]^2 \ .$$

Then f is differentiable on $\mathbb{R}$.

(b) Compute f' using the Chain Rule, and show that f' is identically 0. So f is constant.

(c) Since $f(-\beta) = 0$, f is identically 0. Conclude that the sum formulas hold.

(d) Prove the following identities from the above sum formulas, for every $\alpha, \beta \in \mathbb{R}$ which make sense in the formulas:

$$\sin(\alpha-\beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$$

$$\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta$$

$$\sin\alpha\cos\beta = \frac{1}{2}[\sin(\alpha+\beta) + \sin(\alpha-\beta)]$$

$$\cos\alpha\cos\beta = \frac{1}{2}[\cos(\alpha+\beta) + \cos(\alpha-\beta)]$$

$$\sin\alpha\sin\beta = \frac{1}{2}\left[\cos(\alpha-\beta) - \cos(\alpha+\beta)\right]$$

$$\sin 2\alpha = 2\sin\alpha\cos\alpha$$

$$\cos 2\alpha = \cos^2\alpha - \sin^2\alpha = 2\cos^2\alpha - 1 = 1 - 2\sin^2\alpha$$

$$\cos^2\frac{\alpha}{2} = \frac{1+\cos\alpha}{2}$$

$$\sin^2\frac{\alpha}{2} = \frac{1-\cos\alpha}{2}$$

$$\tan(\alpha+\beta) = \frac{\tan\alpha + \tan\beta}{1 - \tan\alpha\tan\beta}$$

V.6.7.5. Here is a function which appears in some texts, particularly older ones, and connects trigonometric and hyperbolic functions without the need to use complex numbers.

DEFINITION. The *Gudermannian* is the function

$$gd(x) = \tan^{-1}(\sinh x)\ .$$

(a) Show that the Gudermannian is a strictly increasing differentiable function from $\mathbb{R}$ onto $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

(b) Show that $gd'(x) = \operatorname{sech} x$, so, since $gd(0) = 0$, for $x \in \mathbb{R}$,

$$gd(x) = \int_0^x \operatorname{sech} t\, dt\ .$$

(c) The inverse function gd^{-1} is a strictly increasing differentiable function from $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ onto $\mathbb{R}$. Show that it is given by the formulas

$$gd^{-1}(x) = \sinh^{-1}(\tan x) = \log(\sec x + \tan x)$$

for $-\frac{\pi}{2} < x < \frac{\pi}{2}$.

(d) Show that $(gd^{-1})'(x) = \sec x$, so, since $gd^{-1}(0) = 0$, for $x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$,

$$gd^{-1}(x) = \int_0^x \sec t\, dt\ .$$

V.7. Big-O and Little-o Notation

In this section, we introduce some notation which is commonly used to describe the asymptotic behavior of functions or sequences, the big-O and little-o notation. Although some care must be exercised in using this notation, it is very helpful in simplifying formulas and arguments containing constants whose exact value is unimportant or functions which become negligible as a limit is approached, in addition to giving a clean and precise description of the essential quantitative behavior of limits of functions. Not only does use of this system often simplify arguments in form, but it also sometimes helps conceptually in that the essential ideas of the argument are clarified and highlighted; in addition, arguments written this way are usually closer in spirit to the way the founding fathers of calculus phrased their work (although our arguments have a much higher level of rigor than those of the early days of the subject.) In fact, a version of the little-o notation was used by Newton (it is unclear whether he invented it), who wrote:

> "When I am investigating a truth or the solution to a Probleme I use all sorts of approximations and neglect to write down the letter o, but when I am demonstrating a Proposition I always write down the letter o & proceed exactly by the rules of Geometry."
>
> *Isaac Newton*[13]

Definition. Let a be an extended real number, and f and g functions defined in a deleted neighborhood of a (this does not preclude the possibility that they are also defined at a if $a \in \mathbb{R}$).

(i) If $a \in \mathbb{R}$, then $g(x) = O(f(x))$ as $x \to a$ if there exist constants $M \geq 0$ and $\delta > 0$ such that $|g(x)| \leq M|f(x)|$ whenever $0 < |x - a| < \delta$.

(ii) If $a \in \mathbb{R}$, then $g(x) = o(f(x))$ as $x \to a$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that $|g(x)| \leq \epsilon|f(x)|$ whenever $0 < |x - a| < \delta$.

If $a = \pm\infty$, the δ and $0 < |x - a| < \delta$ must be replaced in the usual way by B and $x > B$ for $a = +\infty$, and by A and $x < A$ if $a = -\infty$.

An analogous definition can be made for one-sided approach if f and g are defined in a deleted one-sided neighborhood of a, and for sequences. Thus we may write "$g(x) = O(f(x))$ as $x \to a+$" or "$y_n = o(x_n)$ as $n \to \infty$."

In the definition of O, we can clearly assume that $M > 0$ since there is no harm in increasing M.

The a is usually fixed throughout an argument and is often understood.

The O and o stand for "order." The expression $g(x) = O(f(x))$ as $x \to a$ roughly means "$|g(x)|$ is of the same order as (or not essentially larger than) $|f(x)|$ for x near a"; $g(x) = o(f(x))$ roughly means "$|g(x)|$ is of smaller order than (or negligibly small compared to) $|f(x)|$ for x near a."

Examples.

(i) $g(x) = O(f(x))$ as $x \to a$ roughly means that "$\frac{g(x)}{f(x)}$ is bounded in a neighborhood of a." In particular, $g(x) = O(1)$ as $x \to a$ exactly means that g is bounded in a neighborhood of a. If $\lim_{x\to a} \frac{g(x)}{f(x)}$ exists, then $g(x) = O(f(x))$ as $x \to a$. But note that $g(x) = O(f(x))$ as $x \to a$ does not imply that $\lim_{x\to a} \frac{g(x)}{f(x)}$ exists; for example, $\sin \frac{1}{x} = O(1)$ as $x \to 0$.

[13] [?, p. 204].

(ii) $g(x) = o(f(x))$ as $x \to a$ roughly means that $\lim_{x\to a} \frac{g(x)}{f(x)} = 0$; if $\lim_{x\to a} \frac{g(x)}{f(x)} = 0$, then $g(x) = o(f(x))$ as $x \to a$, but not quite conversely: $x \sin \frac{1}{x} = o(\sin \frac{1}{x})$ as $x \to 0$. But $g(x) = o(1)$ as $x \to a$ exactly means that $\lim_{x\to a} g(x) = 0$.

(iii) $x^m = O(x^n)$ as $x \to 0$ if and only if $m \geq n$; $x^m = o(x^n)$ as $x \to 0$ if and only if $m > n$. But $x^m = O(x^n)$ as $x \to +\infty$ if and only if $m \leq n$, and $x^m = o(x^n)$ as $x \to +\infty$ if and only if $m < n$.

PROPOSITION. Let a be an extended real number, and f, g, and h real-valued functions defined in a neighborhood of a.

(i) If $g(x) = o(f(x))$ as $x \to a$, then $g(x) = O(f(x))$ as $x \to a$.

(ii) If $g(x) = O(f(x))$ and $h(x) = O(g(x))$ as $x \to a$, then $h(x) = O(f(x))$ as $x \to a$.

(iii) If $g(x) = o(f(x))$ and $h(x) = O(g(x))$ as $x \to a$, then $h(x) = o(f(x))$ as $x \to a$.

(iv) If $g(x) = O(f(x))$ and $h(x) = o(g(x))$ as $x \to a$, then $h(x) = o(f(x))$ as $x \to a$.

In particular, if $g(x) = O(f(x))$ and $f(x) = O(1)$, then $g(x) = O(1)$; if $g(x) = o(f(x))$ and $f(x) = O(1)$, or $g(x) = O(f(x))$ and $f(x) = o(1)$, then $g(x) = o(1)$.

PROOF: (i) is obvious from the definitions.
(ii): By hypothesis, there is an $M > 0$ and $\delta_1 > 0$ such that $|g(x)| \leq M|f(x)|$ for $0 < |x - a| < \delta_1$, and an $N > 0$ and $\delta_2 > 0$ such that $|h(x)| \leq N|g(x)|$ for $0 < |x - a| < \delta_2$. Set $\delta = \min(\delta_1, \delta_2)$. Then, if $0 < |x - a| < \delta$, we have

$$|h(x)| \leq N|g(x)| \leq MN|f(x)| \ .$$

(iii): Let $\epsilon > 0$. By hypothesis, there is an $M > 0$ and $\delta_1 > 0$ such that $|h(x)| \leq M|g(x)|$ for $0 < |x-a| < \delta_1$. There is then a $\delta_2 > 0$ such that $|g(x)| \leq \frac{\epsilon}{M}|f(x)|$ for $0 < |x - a| < \delta_2$. Set $\delta = \min(\delta_1, \delta_2)$. Then, for $0 < |x - a| < \delta$,

$$|h(x)| \leq M|g(x)| \leq M \cdot \frac{\epsilon}{M}|f(x)| = \epsilon|f(x)| \ .$$

(iv) is almost identical to (iii), and is left to the reader. 👍

We often just write "$O(f(x))$" or "$o(f(x))$" (along with "as $x \to a$" if it is not understood from the context) to denote some function $g(x)$ which satisfies $g(x) = O(f(x))$ or $g(x) = o(f(x))$ (as $x \to a$).

Note that the expressions $O(f(x))$ and $o(f(x))$ do not necessarily represent specific functions of x (or, more precisely, different uses of $O(f(x))$ or $o(f(x))$ may represent different specific functions of x). Consequently, these expressions do not necessarily follow the usual algebraic rules. For example, from

$$g(x) + o(f(x)) = h(x) + o(f(x))$$

it cannot be concluded that $g(x) = h(x)$; it only follows that $g(x) = h(x)+o(f(x))$ (i.e. $g(x)-h(x) = o(f(x))$.) Here are some rules that are valid:

PROPOSITION.

(i) If $g_1(x) = O(f(x))$ and $g_2(x) = O(f(x))$ as $x \to a$, then $g_1(x) + g_2(x) = O(f(x))$ as $x \to a$.

(ii) If $g_1(x) = o(f(x))$ and $g_2(x) = o(f(x))$ as $x \to a$, then $g_1(x) + g_2(x) = o(f(x))$ as $x \to a$.

(iii) If $g_1(x) = O(f_1(x))$ and $g_2(x) = O(f_2(x))$ as $x \to a$, then $g_1(x)g_2(x) = O(f_1(x)f_2(x))$ as $x \to a$. In particular, if $g(x) = O(f(x))$ and $h(x) = O(1)$ as $x \to a$, then $g(x)h(x) = O(f(x))$ as $x \to a$.

(iv) If $g_1(x) = o(f_1(x))$ and $g_2(x) = O(f_2(x))$ as $x \to a$, then $g_1(x)g_2(x) = o(f_1(x)f_2(x))$ as $x \to a$. In particular, if $g(x) = o(f(x))$ and $h(x) = O(1)$ as $x \to a$, then $g(x)h(x) = o(f(x))$ as $x \to a$.

These results are often abbreviated: $O(f(x))+O(f(x)) = O(f(x))$, $o(f(x))+o(f(x)) = o(f(x))$, $O(f(x))\cdot O(g(x)) = O(f(x)g(x))$, $o(f(x))\cdot O(g(x)) = o(f(x)g(x))$. One of the simplest and most common applications of the last two results is that $K \cdot O(f(x)) = O(f(x))$ and $K \cdot o(f(x)) = o(f(x))$ for any constant K.

PROOF: We prove only (iv); the other proofs are very similar (but slightly easier). Let $\epsilon > 0$. By hypothesis, there is an $M > 0$ and $\delta_2 > 0$ such that $|g_2(x)| \le M|f_2(x)|$ for $0 < |x - a| < \delta_2$. Then there exists a $\delta_1 > 0$ such that $|g_1(x)| \le \frac{\epsilon}{M}|f_1(x)|$ for $0 < |x - a| < \delta_1$. Set $\delta = \min(\delta_1, \delta_2)$. Then, for $0 < |x - a| < \delta$,

$$|g_1(x)g_2(x)| \le \frac{\epsilon}{M}|f_1(x)| \cdot M|f_2(x)| = \epsilon|f_1(x)f_2(x)| \ .$$

The next result is fairly special, but is useful in dealing with quotients. It is awkward to even state this result without using the o-notation (try it!)

PROPOSITION. Let a be an extended real number, and g, h, and k real-valued functions defined on a deleted neighborhood of a, with $h(x) \neq 0$ for all x in a deleted neighborhood of a. Suppose $g(x) = O(1)$, $\frac{1}{h(x)} = O(1)$, and $k(x) = o(1)$ as $x \to a$ (i.e. g is bounded and h is bounded away from 0 in a deleted neighborhood of a, and $\lim_{x\to a} k(x) = 0$). Then

$$\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} - \frac{g(x)}{[h(x)]^2}k(x) + O([k(x)]^2) \text{ as } x \to a \ .$$

PROOF: The proof uses the algebraic formula

$$\frac{1}{1+t} = 1 - t + \frac{t^2}{1+t}$$

valid for $t \neq -1$, which is easily proved by multiplying both sides by $1 + t$. (This formula is a special case of the formula of ().)

Under the hypotheses, there is an $M > 0$ and $\delta_1 > 0$ such that $|h(x)| \ge M$ for $0 < |x - a| < \delta_1$, and there is a $\delta_2 > 0$ such that $|k(x)| < M/2$ for $0 < |x - a| < \delta_2$. There is also an $N \ge 0$ and $\delta_3 > 0$ such that $|g(x)| \le N$ for $0 < |x - a| < \delta_3$. Set $\delta = \min(\delta_1, \delta_2, \delta_3)$. Then if $0 < |x - a| < \delta$, $|h(x) + k(x)| > M/2$ and in particular is nonzero, so $\frac{g(x)}{h(x)+k(x)}$ is defined in a deleted neighborhood of 0. Also, if $0 < |x - a| < \delta$, $\left|\frac{k(x)}{h(x)}\right| < 1/2$, so $\frac{k(x)}{h(x)} \neq -1$. Thus, for $0 < |x - a| < \delta$, we have

$$\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} \cdot \frac{1}{1 + \frac{k(x)}{h(x)}}$$

$$= \frac{g(x)}{h(x)} \left(1 - \frac{k(x)}{h(x)} + \frac{\left[\frac{k(x)}{h(x)}\right]^2}{1 + \frac{k(x)}{h(x)}} \right)$$

$$= \frac{g(x)}{h(x)} - \frac{g(x)}{[h(x)]^2} k(x) + \frac{g(x)}{[h(x)]^2 (h(x) + k(x))} [k(x)]^2 \ .$$

The last term is $O([k(x)]^2)$ as $x \to a$ since

$$\left| \frac{g(x)}{[h(x)]^2 (h(x) + k(x))} \right| \leq \frac{N}{M^2 \left(\frac{M}{2}\right)} = \frac{2N}{M^3}$$

for $0 < |x - a| < \delta$. 👍

COROLLARY. Let a be an extended real number, and f, g, h, and k real-valued functions defined on a deleted neighborhood of a, with $h(x) \neq 0$ for all x in a deleted neighborhood of a. Suppose $g(x) = O(1)$, $\frac{1}{h(x)} = O(1)$, and $f(x) = o(1)$ as $x \to a$ (i.e. g is bounded and h is bounded away from 0 in a deleted neighborhood of a, and $\lim_{x \to a} f(x) = 0$).

(i) If $k(x) = O(f(x))$ as $x \to a$, then

$$\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} + O(f(x)) \text{ as } x \to a.$$

(ii) If $k(x) = o(f(x))$ as $x \to a$, then

$$\frac{g(x)}{h(x) + k(x)} = \frac{g(x)}{h(x)} + o(f(x)) \text{ as } x \to a.$$

This result is sometimes abbreviated by the following curious algebraic formula: "Under the hypotheses on f, g, and h,

$$\frac{g(x)}{h(x) + o(f(x))} = \frac{g(x)}{h(x)} + o(f(x))"$$

(and similarly for $O(f(x))$.) The explanation of why this expression does not violate the rules of algebra is that the $o(f(x))$'s on the two sides of the equation stand for (usually) different functions of x.

PROOF: First observe that $\frac{g(x)}{[h(x)]^2}$ is bounded in a deleted neighborhood of a. Using the notation of the proof of V.7., it is bounded by $\frac{N}{M^2}$ for $0 < |x - a| < \delta$. Thus the second term is $O(k(x))$, which is $O(f(x))$ if $k(x) = O(f(x))$, and $o(f(x))$ if $k(x) = o(f(x))$. Also, the third term is $O([k(x)]^2) = o(k(x)) = o(f(x))$ in either part since $k(x) = o(1)$. 👍

The o is probably used more than the O, since o denotes a function which becomes negligible and thus is very useful in limit or differentiability arguments (although the O, if used carefully, gives a more precise description of asymptotic behavior). The next result, an almost immediate consequence of the definition of differentiability, is probably the most important example of use of the little-o notation:

PROPOSITION. Let f be a real-valued function defined in a neighborhood of $a \in \mathbb{R}$. Then f is differentiable at a if and only if there is an $m \in \mathbb{R}$ such that

$$f(x) = f(a) + m(x-a) + o(x-a) \text{ as } x \to a$$

(or, more precisely, $f(x) - f(a) - m(x-a) = o(x-a)$ as $x \to a$). Such an m, if it exists, is unique and equals $f'(a)$.

As an example of use of the little-o notation in simplifying arguments, here is a proof of the Product Rule (). Note that this proof shows how the product rule follows naturally from the algebraic rule for multiplying multinomials. In addition, it avoids having to use the trick/technique of adding and subtracting something from an expression which may not be obvious, and also avoids explicitly using the result that a differentiable function is continuous.

PROOF: Let f and g be real-valued functions defined in a neighborhood of $a \in \mathbb{R}$, and differentiable at a. Then $f(x) = f(a) + f'(a)(x-a) + o(x-a)$ and $g(x) = g(a) + g'(a)(x-a) + o(x-a)$ in a neighborhood of a. So, in a neighborhood of a,

$$\begin{aligned} f(x)g(x) &= [f(a) + f'(a)(x-a) + o(x-a)][g(a) + g'(a)(x-a) + o(x-a)] \\ &= f(a)g(a) + f'(a)g(a)(x-a) + g(a)o(x-a) + f(a)g'(a)(x-a) + f'(a)g'(a)(x-a)^2 \\ &\quad + g'(a)(x-a)o(x-a) + f(a)o(x-a) + f'(a)(x-a)o(x-a) + o(x-a)\cdot o(x-a) \\ &= f(a)g(a) + [f'(a)g(a) + f(a)g'(a)](x-a) \\ &\quad + [g(a)o(x-a) + f'(a)g'(a)(x-a)^2 + g'(a)(x-a)o(x-a) \\ &\quad + f(a)o(x-a) + f'(a)(x-a)o(x-a) + o(x-a)\cdot o(x-a)] \end{aligned}$$

The expressions in the last brackets are all $o(x-a)$ since $f(a)$, $g(a)$, $f'(a)g'(a)$, $f'(a)(x-a)$, $g'(a)(x-a)$, and $o(x-a)$ are all $O(1)$ as $x \to a$, and $(x-a)^2 = o(x-a)$ as $x \to a$. Thus

$$f(x)g(x) = f(a)g(a) + [f'(a)g(a) + f(a)g'(a)](x-a) + o(x-a)$$

in a neighborhood of a, and so by V.7., fg is differentiable at a and

$$(fg)'(a) = f'(a)g(a) + f(a)g'(a) \ .$$

We now give a proof of the Quotient Rule using the o-notation. This proof is not as clean or satisfying as the above proof of the Product Rule, but gives a better idea of where the Quotient Rule comes from algebraically than the usual proof, and also avoids the "adding and subtracting" trick/technique. It does, however, use in one place the fact that a differentiable function is continuous.

PROOF: Let f and g be defined in a neighborhood of $a \in \mathbb{R}$ and differentiable at a, with $g(a) \neq 0$. Then, since f and g are continuous at a, there is a $\delta_1 > 0$ such that $|f(x)| < |f(a)| + 1$ for $|x-a| < \delta_1$, and there is $\delta_2 > 0$ such that $|g(x)| > \frac{|g(a)|}{2}$ for $|x-a| < \delta_2$. Thus g is nonzero in a neighborhood of a, so $\frac{f}{g}$ is defined in a neighborhood of a, and f and $\frac{1}{g}$ are bounded in a neighborhood of a. By V.7., we can write $f(x) = f(a) + f'(a)(x-a) + o(x-a)$ and $g(x) = g(a) + g'(a)(x-a) + o(x-a)$ in a neighborhood of a, so

$$\frac{f(x)}{g(x)} = \frac{f(a) + f'(a)(x-a) + o(x-a)}{g(a) + g'(a)(x-a) + o(x-a)} \ .$$

First applying V.7.(ii), and then applying V.7., we obtain

$$\begin{aligned}\frac{f(x)}{g(x)} &= \frac{f(a)+f'(a)(x-a)+o(x-a)}{g(a)+g'(a)(x-a)+o(x-a)}\\ &= \frac{f(a)+f'(a)(x-a)+o(x-a)}{g(a)+g'(a)(x-a)}+o(x-a)\\ &= \frac{f(a)+f'(a)(x-a)+o(x-a)}{g(a)}\\ &\quad -\frac{f(a)+f'(a)(x-a)+o(x-a)}{[g(a)]^2}g'(a)(x-a)+O([x-a]^2)+o(x-a)\\ &= \frac{f(a)+f'(a)(x-a)+o(x-a)}{g(a)}\\ &\quad -\frac{f(a)+f'(a)(x-a)+o(x-a)}{[g(a)]^2}g'(a)(x-a)+o(x-a)\ .\end{aligned}$$

Expanding and rearranging terms, we obtain:

$$\begin{aligned}\frac{f(x)}{g(x)} &= \frac{f(a)}{g(a)}+\frac{f'(a)}{g(a)}(x-a)+\frac{o(x-a)}{g(a)}-\frac{f(a)g'(a)}{[g(a)]^2}(x-a)\\ &\quad -\frac{f'(a)g'(a)}{[g(a)]^2}(x-a)^2-\frac{o(x-a)}{[g(a)]^2}(x-a)+o(x-a)\\ &= \frac{f(a)}{g(a)}+\left[\frac{f'(a)}{g(a)}-\frac{f(a)g'(a)}{[g(a)]^2}\right](x-a)\\ &\quad +\left[\frac{o(x-a)}{g(a)}-\frac{f'(a)g'(a)}{[g(a)]^2}(x-a)^2-\frac{o(x-a)}{[g(a)]^2}(x-a)+o(x-a)\right]\ .\end{aligned}$$

All the terms in the last brackets are $o(x-a)$, so in a neighborhood of a we have

$$\frac{f(x)}{g(x)} = \frac{f(a)}{g(a)}+\left[\frac{f'(a)}{g(a)}-\frac{f(a)g'(a)}{[g(a)]^2}\right](x-a)+o(x-a)\ .$$

Thus by V.7., $\frac{f}{g}$ is differentiable at a and

$$\left(\frac{f}{g}\right)'(a) = \frac{f'(a)}{g(a)}-\frac{f(a)g'(a)}{[g(a)]^2} = \frac{g(a)f'(a)-f(a)g'(a)}{[g(a)]^2}\ .$$



V.7.1. Exercises

1. Give careful definitions of (1) "$g(x) = O(f(x))$ as $x \to +\infty$" (2) "$g(x) = O(f(x))$ as $x \to a+$" (3) "$y_n = O(x_n)$ as $n \to \infty$", and similarly for o, and give versions of all the results of this section for these notions.

V.8. The Mean Value Theorem

> "The Mean Value Theorem is the midwife of calculus – not very important or glamorous by itself, but often helping to deliver other theorems that are of major significance."
>
> *E. Purcell and D. Varberg*[14]

The Mean Value Theorem (MVT) is one of the major theoretical results of calculus. It is the primary tool allowing information about a differentiable function to be obtained from properties of its derivative.

V.8.1. Rolle's Theorem and the Mean Value Theorem

V.8.1.1. THEOREM. [MEAN VALUE THEOREM] Let $a, b \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). Then there is at least one $c \in (a, b)$ for which

$$f(b) - f(a) = f'(c)(b - a) \ .$$

Alternatively, there is at least one $c \in (a, b)$ with

$$f'(c) = \frac{f(b) - f(a)}{b - a} \ .$$

A geometric interpretation is that there is at least one point on the graph between a and b at which the tangent line is parallel to the secant line between $(a, f(a))$ and $(b, f(b))$ (Figure V.9).

V.8.1.2. The number $f'(c)$ can be interpreted as the instantaneous rate of change of f at c, while the number $\frac{f(b)-f(a)}{b-a}$ is the average rate of change of f between a and b. Thus, for example, if a driver's average speed is 60 MPH over a period of time, there must be at least one instant of time during the period at which his instantaneous speed is exactly 60 MPH.

V.8.1.3. Three important points to note about the conclusion of the MVT:

(i) The c depends on the numbers a and b, as well as on the function f.

(ii) For a given f, a, and b, there may be more than one c, even infinitely many, satisfying the conclusion. (In Figure V.9 there are five, denoted $c_1 - c_5$.) Indeed, in the extreme case where f is linear, every $c \in (a, b)$ works.

(iii) In practice it may be difficult if not impossible to explicitly find a c that works for a given a and b, even for relatively well-behaved f. The MVT is a prime example of a *pure existence theorem*, in which something is shown to exist without any clue as to how to find a specific example. As we will see, however, just the knowledge that such a c exists will be of great importance in obtaining conclusions about the function f.

[14] ?

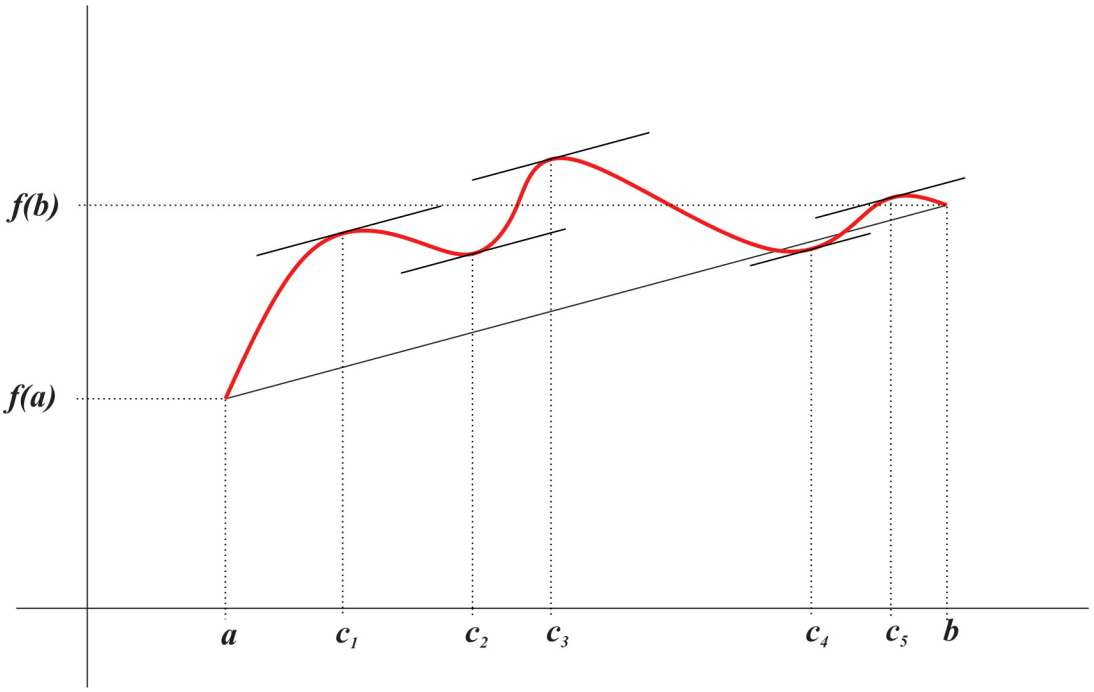


Figure V.9: The Mean Value Theorem

Rolle's Theorem

We will give the standard modern proof of the MVT, due to O. Bonnet in 1868. Cauchy previously gave a more complicated and flawed proof (cf. [Bre07]). The essential step in Bonnet's proof is to prove a special case, which is important enough in its own right to have a standard name:

V.8.1.4. Theorem. [Rolle's Theorem] Let $a, b \in \mathbb{R}$, $a < b$, and let f be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then there is at least one $c \in (a, b)$ with $f'(c) = 0$.

It is ironic that a theorem of calculus should be named after Rolle, since according to [Bal60, p. 318], Rolle, a contemporary of Newton and Leibniz, believed that calculus was "nothing but a collection of ingenious fallacies." (It has been suggested that he later changed his mind.) Many references claim that Rolle first stated and proved this theorem; however, Rolle's statement and proof only referred to polynomials and did not use any explicit calculus. It is unclear who first stated the result in its modern form. See the chapter by J. Barrow-Green in [RS09] for the full story. See also Exercise V.8.6.11..

Proof: Since f is continuous on a closed bounded interval, it attains a maximum M and a minimum m on $[a, b]$ (Max-Min Theorem). Then, since $m \leq f(a) \leq M$, at least one of the following is true:

(i) $M > f(a)$.

(ii) $m < f(a)$.

(iii) f is constant on $[a,b]$.

In case (i), there is a $c \in (a,b)$ with $f(c) = M$. Then $f'(c) = 0$ by V.3.6.3..
In case (ii), there is a $c \in (a,b)$ with $f(c) = m$. Again by V.3.6.3. we have $f'(c) = 0$.
And in case (iii), $f'(c) = 0$ for all $c \in (a,b)$. 👍

Before going on to the MVT, we record some simple consequences of Rolle's theorem which are basically just restatements or special cases of the theorem.

V.8.1.5. Corollary. Let $a, b \in \mathbb{R}$, $a < b$, and let f be continuous on $[a,b]$ and differentiable on (a,b). If $f'(x) \neq 0$ for all $x \in (a,b)$, then $f(a) \neq f(b)$.

The next corollary is useful in locating the zeros of a function such as a polynomial:

V.8.1.6. Corollary. Let f be differentiable on an interval I. Then between any two zeros of f in I there is at least one zero of f'.

The Mean Value Theorem

We now turn to the MVT. This theorem was first stated by Lagrange in (). We will actually prove the following more general theorem, due to Cauchy (with a flawed proof) in 1821; the usual MVT is the special case $g(x) = x$.

V.8.1.7. Theorem. [Cauchy Mean Value Theorem] Let $a, b \in \mathbb{R}$, $a < b$. Let f and g be functions which are continuous on $[a,b]$ and differentiable on (a,b). Then there is at least one $c \in (a,b)$ for which

$$g'(c)[f(b) - f(a)] = f'(c)[g(b) - g(a)] \ .$$

Proof: For $x \in [a,b]$, set

$$h(x) = f(x)[g(b) - g(a)] - g(x)[f(b) - f(a)] \ .$$

Then h is continuous on $[a,b]$ and differentiable on (a,b), and $h(a) = h(b)$. By V.8.1.4. there is a $c \in (a,b)$ such that

$$0 = h'(c) = f'(c)[g(b) - g(a)] - g'(c)[f(b) - f(a)] \ .$$

👍

The following special case is commonly stated as Cauchy's Mean Value Theorem:

V.8.1.8. Corollary. Let $a, b \in \mathbb{R}$, $a < b$. Let f and g be functions which are continuous on $[a, b]$ and differentiable on (a, b), such that $g'(x) \neq 0$ for all $x \in (a, b)$. Then $g(a) \neq g(b)$, and there is at least one $c \in (a, b)$ for which

$$\frac{f(b) - f(a)}{g(b) - g(a)} = \frac{f'(c)}{g'(c)} .$$

Proof: The first conclusion follows from V.8.1.5., and then the second conclusion is V.8.1.7.. 👍

V.8.1.9. This corollary has a geometric interpretation. If $\phi(t) = (g(t), f(t))$, then ϕ is the parametrization of a smooth curve in $\mathbb{R}^2$. The slope of the line through the endpoints of the curve is

$$\frac{f(b) - f(a)}{g(b) - g(a)}$$

and the slope of the tangent line at the point $\phi(t)$ is $\frac{f'(t)}{g'(t)}$. Cauchy's MVT says that there is at least one point on the curve where the tangent line is parallel to the line through the two endpoints. (The requirement that $g' \neq 0$ on (a, b) insures that this curve has no vertical tangents, and it turns out to actually be the graph of a differentiable function (cf. V.8.2.14.). The geometric result is true more generally as long as the endpoints are distinct and f' and g' do not simultaneously vanish (Exercise V.8.6.1.).)

V.8.2. Applications of the Mean Value Theorem

There are many important theoretical consequences of the MVT. We begin with a crucial one:

V.8.2.1. Corollary. Let $a, b \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) \geq 0$ for all $x \in (a, b)$, then f is nondecreasing on $[a, b]$, and in particular $f(a) \leq f(b)$. If f' is not identically 0 on (a, b), then $f(a) < f(b)$.

Proof: If $a \leq x_1 < x_2 \leq b$, applying the MVT to the interval $[x_1, x_2]$ we have

$$f(x_2) - f(x_1) = f'(c)(x_2 - x_1) \geq 0$$

for some $c \in (x_1, x_2)$. Since f is nondecreasing on $[a, b]$, if $f(a) = f(b)$ we must have f constant on $[a, b]$, and hence f' is identically 0. 👍

Similarly, we have:

V.8.2.2. Corollary. Let $a, b \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) \leq 0$ for all $x \in (a, b)$, then f is nonincreasing on $[a, b]$, and in particular $f(a) \geq f(b)$. If f' is not identically 0 on (a, b), then $f(a) > f(b)$.

Putting the last two together, we get a result of basic significance:

V.8.2.3. Corollary. Let $a, b \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) = 0$ for all $x \in (a, b)$, then f is constant on $[a, b]$.

V.8.2.4. Corollary. Let $a, b \in \mathbb{R}$, $a < b$. Let f and g be functions which are continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) = g'(x)$ for all $x \in (a, b)$, then f and g differ by a constant.

Proof: Apply V.8.2.3. to $f - g$. 👍

V.8.2.5. Corollary. Let $a, b \in \mathbb{R}$, $a < b$. Let f and g be functions which are continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = g(a)$ and $f'(x) \geq g'(x)$ for all $x \in (a, b)$, then $f(x) \geq g(x)$ for all $x \in [a, b]$, with strict inequality unless $f' = g'$ on (a, x).

Proof: Apply V.8.2.1. to $f - g$. 👍

The First Derivative Test for Local Extrema

One of the standard tests for local extrema is a direct consequence of the MVT:

V.8.2.6. Theorem. [First Derivative Test] Let f be continuous on an interval I, and a an interior point of I. Suppose

(i) There is an $r \in I$, $r < a$, with f differentiable on (r, a) and $f'(x) \geq 0$ for all $x \in (r, a)$.

(ii) There is an $s \in I$, $s > a$, with f differentiable on (a, s) and $f'(x) \leq 0$ for all $x \in (a, s)$.

Then f has a local maximum at a. It is a strict local maximum unless f' is identically zero in an interval to the left or right of a.
Similarly, suppose

(iii) There is an $r \in I$, $r < a$, with f differentiable on (r, a) and $f'(x) \leq 0$ for all $x \in (r, a)$.

(iv) There is an $s \in I$, $s > a$, with f differentiable on (a, s) and $f'(x) \geq 0$ for all $x \in (a, s)$.

Then f has a local minimum at a. It is a strict local minimum unless f' is identically zero in an interval to the left or right of a.

Note that differentiability of f at a is not assumed (if $f'(a)$ exists, it must be 0 as a result of this theorem and V.3.6.3., or from V.8.2.15..)
Proof: In the first case, apply V.8.2.1. to $[r, a]$ and V.8.2.2. to $[a, s]$, and vice versa in the second case. 👍

A one-sided version of this test also applies at endpoints of I; the statement is left to the reader.

Differential Approximation with Error Estimates

We can obtain a variation of differential approximation with error estimates:

V.8.2.7. Corollary. Let $a, b, M \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). If $f'(x) < M$ for all $x \in (a, b)$, then

$$f(x_2) - f(x_1) < M(x_2 - x_1)$$

for all $x_1, x_2 \in [a, b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

Proof: Let $x_1, x_2 \in [a, b]$, $x_1 < x_2$. Then, applying the MVT to $[x_1, x_2]$, for some $c \in (x_1, x_2)$ we have

$$f(x_2) - f(x_1) = f'(c)(x_2 - x_1) < M(x_2 - x_1) \ .$$

Similarly, we have

V.8.2.8. Corollary. Let $a, b, m \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). If $m < f'(x)$ for all $x \in (a, b)$, then

$$m(x_2 - x_1) < f(x_2) - f(x_1)$$

for all $x_1, x_2 \in [a, b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

Combining, we get the two-sided version:

V.8.2.9. Corollary. Let $a, b, m, M \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). If $m < f'(x) < M$ for all $x \in (a, b)$, then

$$m(x_2 - x_1) < f(x_2) - f(x_1) < M(x_2 - x_1)$$

for all $x_1, x_2 \in [a, b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

Taylor's theorem gives a higher-order version of these inequalities ().

V.8.2.10. V.8.2.7. is often called the *Mean Value Inequality*. It is an immediate consequence of the special case V.8.2.1. and linearity of the derivative. There is a direct proof of V.8.2.1. (in the special case where f is differentiable on $[a, b]$) by the Method of Successive Bisections () (Exercise V.8.6.11.). This proof is arguably simpler and more elementary than the proof of the MVT since it only uses the Nested Intervals Theorem and not the Extreme Value Theorem, and several authors, e.g. [**?**, p. 152], [vRS82, p. 93], [Boa95, p. 8], [Kör04, p. 59], argue that the Mean Value Inequality is sufficient for most applications in analysis and that the MVT itself is overrated. I personally do not regard the proof of the MVT as very difficult given the Extreme Value Theorem, which one would want to include anyway in any treatment of real analysis.

Lipschitz Continuity

A similar consequence which will be important later is:

V.8.2.11. Corollary. Let $a, b, M \in \mathbb{R}$, $a < b$, $M \geq 0$. Let f be a function which is continuous on $[a, b]$ and differentiable on (a, b). If $|f'(x)| \leq M$ for all $x \in (a, b)$, then

$$|f(x_2) - f(x_1)| \leq M|x_2 - x_1|$$

for all $x_1, x_2 \in [a, b]$ (i.e. if f' is bounded on (a, b), then f is a *Lipschitz function* () on $[a, b]$).

Proof: Let $x_1, x_2 \in [a, b]$, $x_1 < x_2$. Then, applying the MVT to $[x_1, x_2]$, for some $c \in (x_1, x_2)$ we have

$$|f(x_2) - f(x_1)| = |f'(c)||x_2 - x_1| \leq M|x_2 - x_1| \ .$$

There is also a generalized version whose statement is left to the reader (it is very similar to V.8.4.9.).

The Intermediate Value Property for Derivatives

Although derivatives can be quite discontinuous (cf. V.12.1.), they rather remarkably satisfy the conclusion of the Intermediate Value Theorem.

V.8.2.12. Definition. A function g defined on an interval I has the *Intermediate Value Property*, or *Darboux Property*, if, whenever $a, b \in I$, $a < b$, and r is a number strictly between $g(a)$ and $g(b)$, there is a $c \in (a, b)$ with $g(c) = r$.

A function which has the Intermediate Value Property is sometimes called *Darboux continuous*.

Every continuous function has the Intermediate Value Property by V.2.3.1.. So do derivatives:

V.8.2.13. Theorem. If f is a differentiable function on an interval I, then f' has the Intermediate Value Property on I.

Proof: Let $a, b \in I$, $a < b$, and r a number strictly between $f'(a)$ and $f'(b)$. Set $g(x) = f(x) - rx$; then g is differentiable on I and $g'(x) = f'(x) - r$ for $x \in I$.

If $f'(a) > r > f'(b)$, then $g'(a) > 0$ and $g'(b) < 0$. Since g is continuous on $[a, b]$, it has a maximum on this interval at some $c \in [a, b]$ (V.2.4.1.). The maximum cannot occur at a or b since $g'(a) > 0$ and $g'(b) < 0$ (V.3.6.12.); thus $c \in (a, b)$, and $g'(c) = 0$ by V.3.6.3.. Then $f'(c) = r$.

The case $f'(a) < r < f'(b)$ is nearly identical, considering the minimum of g on $[a, b]$.

V.8.2.14. Corollary. If f is differentiable on an interval I, and $f'(x) \neq 0$ for all $x \in I$, then f' does not change sign on I, i.e. either $f'(x) > 0$ for all $x \in I$ or $f'(x) < 0$ for all $x \in I$. In particular, f is either strictly increasing on I or strictly decreasing on I (V.8.2.1., V.8.2.2.).

Here is another related approximate continuity property of derivatives:

V.8.2.15. THEOREM. Let f be differentiable on an open interval (a,b), and $c \in (a,b)$. Then

$$\liminf_{x\to c+} f'(x) \le f'(c) \le \limsup_{x\to c+} f'(x)$$

$$\liminf_{x\to c-} f'(x) \le f'(c) \le \limsup_{x\to c-} f'(x) \ .$$

In particular, if $\lim_{x\to c+} f'(x)$ exists, it equals $f'(c)$, and similarly for $\lim_{x\to c-} f'(x)$.

PROOF: We prove only the first inequality; the others are virtually identical. If

$$f'(c) < \liminf_{x\to c+} f'(x)$$

there is an $\epsilon > 0$ and a d, $c < d \le b$, such that $f'(x) \ge f'(c) + \epsilon$ for all $x \in (c,d)$. Then, by the MVT, for every $x, y \in (c,d)$, $x \ne y$, we have

$$\frac{f(y) - f(x)}{y - x} \ge f'(c) + \epsilon \ .$$

Fix $y \in (c,d)$; we then get

$$\frac{f(y) - f(c)}{y - c} = \lim_{x\to c+} \frac{f(y) - f(x)}{y - x} \ge f'(c) + \epsilon$$

since f is continuous at c. But then we have

$$f'(c) = \lim_{y\to c+} \frac{f(y) - f(c)}{y - c} \ge f'(c) + \epsilon$$

which is a contradiction. 👍

In fact, every Darboux continuous function has this approximate continuity property (Exercise V.8.6.4.).

See V.9.4.11. for an extension of the last statement. See also V.12.2.5..

V.8.3. Geometric Interpretation of the Second Derivative

Using the MVT, we can now give careful justification of the interpretation of the second derivative of a function in terms of concavity.

V.8.3.1. DEFINITION. Let f be a continuous function on an interval I. Then f is *concave upward* on I if, whenever $a, b, c \in I$ with $a < c < b$, we have

$$f(c) < f(a) + \frac{f(b) - f(a)}{b - a}(c - a) \ .$$

The function f is *concave downward* on I if, whenever $a, b, c \in I$ with $a < c < b$, we have

$$f(c) > f(a) + \frac{f(b) - f(a)}{b - a}(c - a) \ .$$

Geometrically, f is concave upward on I if on every closed bounded subinterval of I, the graph of f on the interior lies strictly below the secant line through the endpoints, and similarly the graph lies strictly above the secant line if f is concave downwards. "Concave upward" is synonymous with "strictly convex" (cf. XIX.3.1.1.).

V.8.3.2. THEOREM. Let f be continuous on $[a,b]$ and differentiable on (a,b). Then f is concave upward on $[a,b]$ if and only if f' is strictly increasing on (a,b), and f is concave downward on $[a,b]$ if and only if f' is strictly decreasing on (a,b).

PROOF: We prove the first statement. Suppose f' is strictly increasing on $[a,b]$. Let $L(x)$ be the linear function

$$L(x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$$

whose graph is the secant line through $(a,f(a))$ and $(b,f(b))$. Suppose there is an $x_0 \in (a,b)$ with $f(x_0) \geq L(x_0)$. Then by the MVT there are $c_1 \in (a,x_0)$ and $c_2 \in (x_0,b)$ with

$$f'(c_1) = \frac{f(x_0)-f(a)}{x_0-a} \geq \frac{L(x_0)-L(a)}{x_0-a} = \frac{f(b)-f(a)}{b-a} \geq \frac{f(b)-f(x_0)}{b-x_0} = f'(c_2)$$

which contradicts that f' is strictly increasing on (a,b) since $a < c_1 < c_2 < b$. The converse is a special case of XIX.3.1.16.. The proof of the second statement is identical with inequalities reversed. 👍

V.8.3.3. COROLLARY. Let f be continuous on $[a,b]$ and twice differentiable on (a,b). If $f''(x) \geq 0$ for all $x \in (a,b)$, and is not identically 0 on any subinterval, then f is concave upward on $[a,b]$. If $f''(x) \leq 0$ for all $x \in (a,b)$, and is not identically 0 on any subinterval, then f is concave downward on $[a,b]$. In particular, if $f'' > 0$ [resp. $f'' < 0$] on an interval I, then f is concave upward [resp. concave downward] on I.

PROOF: If $f'' \geq 0$ on (a,b) and not 0 on any subinterval, then f' is strictly increasing on (a,b) (V.8.2.1.), and similarly if $f'' \leq 0$. 👍

V.8.3.4. A function which is concave upward need not be differentiable everywhere, for example $f(x) = |x| + x^2$ (but it must be differentiable except on a countable set, cf. XIX.3.1.16.). A linear function is twice differentiable, but not concave up or concave down on any interval. And even a nonlinear twice differentiable function need not be either concave upward or concave downward on any interval: consider an antiderivative for the example of ().

There is a pointwise analog:

V.8.3.5. THEOREM. Let I be an interval, f be a function from I to $\mathbb{R}$, and a in the interior of I. Suppose f is twice differentiable at a (which implies that f' exists in a neighborhood of a and is continuous at a). Let L be the linear function whose graph is the tangent line to the graph of f at a, i.e. $L(x) = f(a)+f'(a)(x-a)$.
(i) If $f''(a) > 0$, then there is an $\epsilon > 0$ such that $f(x) > L(x)$ for all x, $0 < |x-a| < \epsilon$.
(ii) If $f''(a) < 0$, then there is an $\epsilon > 0$ such that $f(x) < L(x)$ for all x, $0 < |x-a| < \epsilon$.

Geometrically, if $f''(a) > 0$, the graph of f near a lies strictly above the tangent line at a, and strictly below if $f''(a) < 0$.

PROOF: (i): We have that f' is increasing at a (V.3.6.8.), i.e. there is an $\epsilon > 0$ such that $f'(x) < f'(a)$ if $x \in (a-\epsilon, a)$ and $f'(x) > f'(a)$ if $x \in (a, a+\epsilon)$. If $x \in (a-\epsilon, a)$, then by the MVT there is a $c \in (x, a)$ with

$$\frac{f(x) - f(a)}{x - a} = f'(c) < f'(a) = \frac{L(x) - L(a)}{x - a}$$

and it follows that $f(x) - f(a) > L(x) - L(a) = L(x) - f(a)$ since $x - a < 0$. The argument if $x \in (a, a+\epsilon)$ is essentially identical. The proof of (ii) is the same with inequalities reversed. 👍

A special case is particularly important:

V.8.3.6. COROLLARY. [SECOND DERIVATIVE TEST] Let I be an interval, f be a function from I to $\mathbb{R}$, and a in the interior of I. Suppose f is twice differentiable at a (which implies that f' exists in a neighborhood of a and is continuous at a).

(i) If $f'(a) = 0$ and $f''(a) > 0$, then f has a strict local minimum at a.

(ii) If $f'(a) = 0$ and $f''(a) < 0$, then f has a strict local maximum at a.

PROOF: The function $L(x)$ is the constant function $f(a)$. 👍

V.8.3.7. This result only goes one way. If f has a relative extremum at a, it need not be twice differentiable (or even once differentiable!) at a, and even if it is we can have $f''(a) = 0$ (consider $f(x) = x^4$ at 0). See V.10.7.12. for an extension of the second derivative test.

Points of Inflection

A point of inflection of a function is a place where the function changes concavity:

V.8.3.8. DEFINITION. Let f be continuous on an interval I, and c an interior point of I. Then f has a *point of inflection* at c if there are $a, b \in I$ with $a < c < b$ such that f is concave upward on $[a, c]$ and concave downward on $[c, b]$, or vice versa.

V.8.3.9. If f is differentiable on I, and has a point of inflection at $c \in I$, then by V.8.3.2. and V.8.2.15. f' must have a strict local extremum at c. In particular, if f is twice differentiable at c, then $f''(c) = 0$. However, the converses are not true: if $f(x) = x^4$, then $f''(0) = 0$, but f does not have a point of inflection at 0, and in fact is concave up on all of $\mathbb{R}$. And if g is an antiderivative for the function f of V.3.10.11., then g is twice differentiable everywhere with $g''(0) = 0$, and g' has a strict local minimum at 0, but g does not have a point of inflection at 0: in fact, g is neither concave up or concave down in any interval with endpoint 0. ✋

V.8.4. Refinements of the Mean Value Theorem

We can give a refinement of V.8.2.1.. We do not obtain this as a consequence of the MVT, but give a direct proof in the spirit of the MVT.

V.8.4.1. THEOREM. Let f be a continuous function on $[a,b]$, and A a countable subset of $[a,b]$. Suppose f is differentiable on $[a,b] \setminus A$.

(i) If $f'(x) > 0$ for all $x \in [a,b] \setminus A$, then f is strictly increasing on $[a,b]$.

(ii) If $f'(x) \geq 0$ for all $x \in [a,b] \setminus A$, then f is nondecreasing on $[a,b]$.

(iii) If $f'(x) = 0$ for all $x \in [a,b] \setminus A$, then f is constant on $[a,b]$.

Similar statements to (i) and (ii) hold if $f'(x) < 0$ or $f'(x) \leq 0$ for all $x \in [a,b] \setminus A$.

Of course, this theorem includes the special cases $A = \emptyset$ and $A = \{a,b\}$, which is V.8.2.1..

PROOF: (i): It suffices to show that f is nondecreasing on $[a,b]$, since if f is nondecreasing but not strictly increasing it would be constant on some subinterval, hence have $f' = 0$ on some interval. In fact, it suffices to show that $f(a) \leq f(b)$, since the interval $[a,b]$ can be replaced by any closed subinterval.

So suppose $f(a) > f(b)$. Since $f(A)$ is countable, we can find a number $d \notin f(A)$ with $f(a) > d > f(b)$. Let $B = \{x \in [a,b] : f(x) = d\}$. B is nonempty by the Intermediate Value Theorem. Since f is continuous, B is closed, and thus has a largest element c. We have $a < c < b$ since $f(a) > d > f(b)$, and $c \notin A$ since $f(c) \notin f(A)$. So f is differentiable at c and $f'(c) > 0$. Since $f(b) < d$ and f never takes the value d on (c,b), we must have $f(x) < d$ for all $x \in (c,b)$ by the Intermediate Value Theorem. Hence

$$\lim_{x \to c+} \frac{f(x) - f(c)}{x - c} \leq 0$$

contradicting the assumption that $f'(c) > 0$.

(ii): Let $\epsilon > 0$. Apply (i) to $g(x) = f(x) + \epsilon x$ to conclude that $f(a) \leq f(b) + \epsilon(b-a)$. Since ϵ is arbitrary, $f(a) \leq f(b)$.

(iii): Apply (ii) to f and $-f$. 👍

The differentiability of f on $[a,b] \setminus A$ can be relaxed to the assumption that the upper right Dini derivative of f is positive (nonnegative, zero) on $[a,b] \setminus A$ (the proof of (iii) is a bit more subtle in this case, however; see ()). Hence a more general statement can be proved with the same argument:

V.8.4.2. THEOREM. Let f be a continuous function on an interval $[a,b]$, and let A be a countable subset of $[a,b]$. Then

(i) If $D^+f(x) > 0$ for all $x \in [a,b] \setminus A$, then f is strictly increasing on $[a,b]$.

(ii) If $D^+f(x) \geq 0$ for all $x \in [a,b] \setminus A$, then f is nondecreasing on $[a,b]$.

(iii) If $D^+f(x) = 0$ for all $x \in [a,b] \setminus A$, then f is constant on $[a,b]$.

Similar statements to (i) and (ii) hold if $D_+(x) < 0$ or $D_+(x) \leq 0$ for all $x \in [a,b] \setminus A$.

V.8.4.3. Note that the only property of A needed for the proof of (i) is that $f(A)$ does not contain any interval. However, there is a subtlety in the proof of (ii) (and hence (iii)) if only this assumption is made: if $f(A)$ does not contain any interval, if A is uncountable it is not obvious that $g(A)$ contains no interval if $g(x) = f(x) + \epsilon x$, and in fact this may fail (cf. [vRS82, 21.G]). Thus, to obtain a more general version of V.8.4.2. in this setting, we would have to add a clumsy set of additional hypotheses, although there are situations where a generalization is possible (e.g. XIX.5.1.18.). We can state part (i) in this setting:

V.8.4.4. THEOREM. Let f be a continuous function on an interval $[a,b]$, and let A be a subset of $[a,b]$ such that $f(A)$ contains no interval. Then

(i) If $D^+f(x) > 0$ for all $x \in [a,b] \setminus A$, then f is strictly increasing on $[a,b]$.

(ii) If $D_+f(x) < 0$ for all $x \in [a,b] \setminus A$, then f is strictly decreasing on $[a,b]$.

Perhaps the most general version of V.8.4.2.(ii) which can be stated reasonably simply is:

V.8.4.5. THEOREM. Let f be a continuous function on an interval $[a,b]$. If $D^+f(x) \geq 0$ for all x not in a set A of measure 0, and $B = \{x \in [a,b] : D^+f(x) = -\infty\}$ is at most countable, then f is nondecreasing on $[a,b]$.

(Under the hypotheses, the sets B, and even A, would turn out to be empty, but this need not be assumed. Actually, the result is interesting and useful even in the case $B = \emptyset$.)
PROOF: Let g be a nondecreasing function on $[a,b]$ with $D_+g(x) = +\infty$ for all $x \in A$ (). Fix $\epsilon > 0$ and let $h_\epsilon = f + \epsilon g$. Then $D^+h_\epsilon(x) \geq D^+f(x) \geq 0$ for $x \notin A \cup B$ since g is nondecreasing. If $x \in A \setminus B$, then $D^+f(x) > -\infty$, so $D^+h_\epsilon(x) = +\infty$ by (). Thus $D^+h_\epsilon(x) \geq 0$ for all $x \notin B$, so h_ϵ is nondecreasing by V.8.4.2.(ii). Since $f = \lim_{\epsilon \to 0} h_\epsilon$, f is also nondecreasing. 👍

V.8.4.6. Note that the hypothesis that B be countable cannot be replaced by an assumption that B is a null set: the negative of the Cantor function () is a counterexample.

Using V.8.4.1., we can obtain generalizations of V.8.2.7. and V.8.2.8.. As explained in V.8.4.3., to obtain a clean statement we must assume the conditions hold except on a countable subset.

V.8.4.7. COROLLARY. Let $a, b, M \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a,b]$. Let A be a countable subset of $[a,b]$. If $D_+f(x) < M$ for all $x \in [a,b] \setminus A$, then

$$f(x_2) - f(x_1) < M(x_2 - x_1)$$

for all $x_1, x_2 \in [a,b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

PROOF: Apply V.8.4.2. to $h(x) = Mx - f(x)$. 👍

V.8.4.8. COROLLARY. Let $a, b, m \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a,b]$. Let A be a countable subset of $[a,b]$. If $m < D^+f(x)$ for all $x \in [a,b] \setminus A$, then

$$m(x_2 - x_1) < f(x_2) - f(x_1)$$

for all $x_1, x_2 \in [a,b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

PROOF: Apply V.8.4.2. to $g(x) = f(x) - mx$. 👍

V.8.4.9. Corollary. Let $a, b, m, M \in \mathbb{R}$, $a < b$. Let f be a function which is continuous on $[a, b]$. Let A be a countable subset of $[a, b]$. If $m < D^+f(x)$ and $D_+f(x) < M$ for all $x \in [a, b] \setminus A$, then

$$m(x_2 - x_1) < f(x_2) - f(x_1) < M(x_2 - x_1)$$

for all $x_1, x_2 \in [a, b]$, $x_1 < x_2$. The result also holds if $<$ is replaced by $\leq$ throughout.

The following special case of V.8.4.2.(iii) is useful:

V.8.4.10. Corollary. Let f and g be continuous functions on an interval I. If there is a countable subset A of I such that $D_Rf(x)$ and $D_Rg(x)$ exist and are equal for all $x \in I \setminus A$, then f and g differ by a constant on I.

Proof: Apply V.8.4.2.(iii) to $f - g$ (on any closed bounded subinterval of I). 👍

V.8.5. Antiderivatives and Primitives

Antiderivatives are usually associated with integration, but we can use the results of the previous subsection to obtain antiderivatives for certain well-behaved functions without discussing integration, at least explicitly.

V.8.5.1. Definition. Let f be a function on an interval I. An *antiderivative* for f on I is a function F on I which is differentiable on I and with $F' = f$ on I.
A *primitive* for f on I is a continuous function F on I such that, for some countable subset A of I, $F'(x)$ exists and equals $f(x)$ for all $x \in I \setminus A$.

V.8.5.2. Since a differentiable function on an interval is continuous, every antiderivative for f on I is a primitive for f on I (with $A = \emptyset$). The terms "antiderivative" and "primitive" are often used as synonyms, but we make a slight technical distinction between them which will be useful.

V.8.5.3. Antiderivatives and primitives are not unique: a constant can be added to any antiderivative (primitive) to obtain another antiderivative (primitive). By V.8.2.5., this is the only change possible for antiderivatives: any two antiderivatives for a function f differ by a constant. This is also true for primitives by V.8.4.10..

Step Functions

Recall () that a *step function* on an interval $[a, b]$ corresponding to a partition

$$\mathcal{P} = \{a = x_0, x_1, \ldots, x_n = b\}$$

is a function which is constant on (x_{k-1}, x_k) for $1 \leq k \leq n$.

V.8.5.4. PROPOSITION. Let f be a step function on $[a,b]$. Then f has a primitive on $[a,b]$.

PROOF: This is quite simple and straightforward, but notationally a bit complicated. The function F will be "piecewise linear" or "polygonal". Suppose f takes the value a_k on (x_{k-1}, x_k). Define f as follows:

For $x_0 \le x \le x_1$, let $F(x) = a_1(x - x_0)$.

For $x_1 < x \le x_2$, let $F(x) = a_1(x_1 - x_0) + a_2(x - x_1)$.

For $x_2 < x \le x_3$, let $F(x) = a_1(x_1 - x_0) + a_2(x_2 - x_1) + a_3(x - x_2)$.

...

For $x_{k-1} < x \le x_k$, let $F(x) = \sum_{j=1}^{k-1} a_j(x_j - x_{j-1}) + a_k(x - x_{k-1})$.

...

It is obvious that F is continuous on $[a,b]$, and that $F'(x) = f(x)$ for $x \in [a,b] \setminus \{x_0, \dots, x_n\}$. 👍

The next theorem is the most important result of this subsection. It will be used for other purposes than the ones of this subsection, such as justifying term-by-term differentiation and integration of infinite series of functions. In many, but not all, applications, the set A in the statement of the theorem will be the empty set. See XIX.5.4.14. for a sophisticated (almost) strengthened version.

V.8.5.5. THEOREM. Let (F_n) and (f_n) be sequences of functions on an interval I, and A a countable subset of I. Suppose

(a) Each F_n is continuous on I.

(b) For every $x \in I \setminus A$, each F_n is differentiable at x and $F_n'(x) = f_n(x)$.

(c) The sequence (f_n) converges u.c. on I to a function f.

(d) There is an $x_0 \in I$ such that the sequence $(F_n(x_0))$ converges.

Then:

(i) The sequence (F_n) converges u.c. on I to a (continuous) function F.

(ii) For each $x \in I \setminus A$, F is differentiable at x and $F'(x) = f(x)$.

PROOF: (i): Let J be a closed bounded subinterval of I. We will show that (F_n) is a uniform Cauchy sequence on J, hence uniformly convergent to a (continuous) limit function F by (). By expanding J, we may assume $x_0 \in J$.

Let $\epsilon > 0$. Choose N such that for all $n, m \ge N$,

$$\sup_{x \in J} |f_n(x) - f_m(x)| < \frac{\epsilon}{2\ell(J)}$$

$$|F_n(x_0) - F_m(x_0)| < \frac{\epsilon}{2} .$$

Then, if $n, m \geq N$ and $x \in J$, by V.8.4.7. we have, since $(F_n - F_m)' = f_n - f_m$ on $J \setminus A$,

$$|[F_n(x) - F_m(x)] - [F_n(x_0) - F_m(x_0)]| < \frac{\epsilon}{2\ell(J)}|x - x_0| \leq \frac{\epsilon}{2}$$

and so

$$\begin{aligned} |F_n(x) - F_m(x)| &= |[F_n(x) - F_n(x_0)] - [F_m(x) - F_m(x_0)] + [F_n(x_0) - F_m(x_0)]| \\ &\leq |[F_n(x) - F_n(x_0)] - [F_m(x) - F_m(x_0)]| + |F_n(x_0) - F_m(x_0)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon . \end{aligned}$$

(ii): Let J be as above, and $\epsilon > 0$. There is an N such that

$$\sup_{x \in J} |f_n(x) - f(x)| < \frac{\epsilon}{4}$$

for all $n \geq N$. Fix $n \geq N$. For all $m \geq N$, we have

$$\sup_{x \in J} |f_n(x) - f_m(x)| < \frac{\epsilon}{2}$$

and hence by V.8.4.7. we have

$$|[F_m(y) - F_m(x)] - [F_n(y) - F_n(x)]| = |[F_m(y) - F_n(y)] - [F_m(x) - F_n(x)]| < \frac{\epsilon}{2}|y - x|$$

for all $x, y \in J$. Since $F_m \to F$ uniformly on J, letting $m \to \infty$ we have

$$|[F(y) - F(x)] - [F_n(y) - F_n(x)]| \leq \frac{\epsilon}{2}|y - x|$$

for all $x, y \in J$.

Now fix $x \in J \setminus A$. Then there is a $\delta > 0$ such that

$$\left| \frac{F_n(y) - F_n(x)}{y - x} - f_n(x) \right| < \frac{\epsilon}{4}$$

whenever $0 < |y - x| < \delta$. Then, for $0 < |y - x| < \delta$, we have

$$\begin{aligned} \left| \frac{F(y) - F(x)}{y - x} - f(x) \right| &= \left| \frac{[F(y) - F(x)] - [F_n(y) - F_n(x)]}{y - x} + \frac{F_n(y) - F_n(x)}{y - x} - f_n(x) + f_n(x) - f(x) \right| \\ &\leq \left| \frac{[F(y) - F(x)] - [F_n(y) - F_n(x)]}{y - x} \right| + \left| \frac{F_n(y) - F_n(x)}{y - x} - f_n(x) \right| + |f_n(x) - f(x)| < \frac{\epsilon}{2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon \end{aligned}$$

and thus $F'(x) = f(x)$. 👍

V.8.5.6. A simpler proof using the Fundamental Theorem of Calculus can be given in the crucial special case where $A = \emptyset$ and the f_n are continuous (IX.2.9.15.).

V.8.5.7. COROLLARY. Let (f_n) be a sequence of functions on an interval I, converging u.c. on I to a function f.

(i) If each f_n has a primitive on I, then f has a primitive on I.

(ii) If each f_n has an antiderivative on I, then f has an antiderivative on I.

PROOF: Fix $x_0 \in I$. By adjusting by a constant, f_n has a primitive (antiderivative in case (ii)) F_n on I with $F_n(x_0) = 0$. Let A_n be the set of $x \in I$ for which $F_n'(x) = f_n(x)$ does not hold; then A_n is countable, and $A_n = \emptyset$ in case (ii). Set $A = \cup_{n=1}^{\infty} A_n$. Then (f_n), (F_n), A satisfy the hypotheses of V.8.5.5.. 👍

V.8.5.8. PROPOSITION. Let f be a continuous function on an interval I. Then there is a sequence (f_n) of step functions on I converging u.c. to f on I. If A is any dense subset of I, the f_n may be chosen to have discontinuities only in A.

PROOF: This is an almost immediate consequence of (). Let $\{[a_n, b_n]\}$ be an increasing sequence of closed bounded intervals with union I. Then f is uniformly continuous on $[a_n, b_n]$. For each n fix $\delta_n > 0$ such that $x, y \in [a_n, b_n]$, $|x - y| < \delta_n$ implies $|f(x) - f(y)| < \frac{1}{n}$. Then let $\mathcal{P}_n$ be a partition of $[a_n, b_n]$ with interior points in A with mesh $< \delta_n$, and t_k a point in the k'th subinterval of the partition. Let f_n be the step function with value $f(t_k)$ on the k'th subinterval of the partition $\mathcal{P}_n$ (and, say, left continuous at the points of the partition), and extend f_n to I by making it constant to the left and right of $[a_n, b_n]$ if necessary. Then $f_n \to f$ u.c. on I (Exercise V.8.6.6.). 👍

V.8.5.9. THEOREM. Let f be a continuous function on an interval I. Then f has an antiderivative on I.

PROOF: Fix $x_0 \in I$, and let A and B be disjoint countable dense subsets of I (e.g. $A = \mathbb{Q} \cap I$, $B = [\mathbb{Q} + \sqrt{2}] \cap I$). Then there is a sequence (f_n) of step functions on I with discontinuities only in A, converging u.c. to f on I. It follows from V.8.5.5. that f has a primitive F_1 on I satisfying $F_1'(x) = f(x)$ for all $x \in I \setminus A$ and $F_1(x_0) = 0$. Similarly, there is a primitive F_2 of f on I such that $F_2'(x) = f(x)$ for all $x \in I \setminus B$ and $F_2(x_0) = 0$. By V.8.4.10. (applied using the countable set $A \cup B$), F_1 and F_2 differ by a constant. But $F_1(x_0) = F_2(x_0)$, so $F_1 = F_2$. Thus $F_1'(x)$ exists and equals $f(x)$ for all $x \in (I \setminus A) \cup (I \setminus B) = I$. 👍

V.8.5.10. We will give another proof of this theorem in IX.2.9.4.. (Deep down, it is really the same proof.)

Another example where a similar argument applies is monotone functions:

V.8.5.11. PROPOSITION. Let f be a monotone function on an interval I. Then there is a sequence (f_n) of step functions on I converging u.c. to f on I.

V.8.5.12. COROLLARY. Let f be a monotone function on an interval I. Then f has a primitive on I.

V.8.5.13. It can be shown with a little additional work that if f is monotone on I and F is a primitive of f on I, then the (countable) set A of x for which $F'(x) = f(x)$ does not hold is exactly the set of discontinuities of f. See V.8.6.7. for a more general result.

V.8.6. Exercises

V.8.6.1. Let C be a smooth curve from p to q in $\mathbb{R}^2$ (i.e. with a parametrization $(g(t), f(t))_{a \le t \le b}$ with f and g differentiable on $[a, b]$ and f' and g' not simultaneously 0). If $p \neq q$ and r is a point on C of maximum distance from the line L through p and q (why does such a point exist?), show that the tangent line to C at r is parallel to L. Extend to the case where f and g are continuous on $[a, b]$ and differentiable on (a, b).

V.8.6.2. (a) Prove V.8.4.2.(i) and (ii) in the case where the upper right Dini derivative D^+f is positive or nonnegative on $[a, b] \setminus A$. Show why (iii) does not immediately follow from (ii) in this case.

(b) Suppose $D^+f(x) = 0$ for all $x \in [a, b] \setminus A$. Apply (ii) to conclude that f is nondecreasing on $[a, b]$. Thus the lower right Dini derivative D_+f satisfies $D_+f(x) \ge 0$ for all $x \in [a, b]$.

(c) In the situation of (b), conclude that $D_+f(x) = 0$ for all $x \in [a, b] \setminus A$; thus (ii) can be applied to $-f$ to conclude that $-f$ is increasing on $[a, b]$. Thus f is constant on $[a, b]$.

V.8.6.3. Use the Intermediate Value Property for derivatives () to give an alternate proof of (). ?????

V.8.6.4. Let g be a Darboux continuous function on an open interval (a, b), and $c \in (a, b)$. Show that

$$\liminf_{x \to c+} g(x) \le g(c) \le \limsup_{x \to c+} g(x)$$

$$\liminf_{x \to c-} g(x) \le g(c) \le \limsup_{x \to c-} g(x) \ .$$

V.8.6.5. Let f be a continuous function on an open interval I. Set

$$M^+ = \sup\{D^+f(x) : x \in I\}$$

$$m^+ = \inf\{D^+f(x) : x \in I\}$$

(m^+ and/or M^+ may be infinite), and let

$$S = \left\{ \frac{f(y) - f(x)}{y - x} : x, y \in I, x \neq y \right\} \ .$$

(a) Show that $(m^+, M^+) \subseteq S \subseteq [m^+, M^+]$. [To show $(m^+, M^+) \subseteq S$, fix $\lambda \in (m^+, M^+)$ and let $g(x) = f(x) - \lambda x$. Then g has both positive and negative right Dini derivatives on I and hence cannot be one-to-one by (). For $S \subseteq [m^+, M^+]$, let $\lambda \le m^+$ and $\mu \ge M^+$, and apply V.8.4.2.(ii) to $f(x) - \lambda x$ and $\mu x - f(x)$.]

(b) Show that $m^+ \in S$ if and only if f is linear with slope m^+ on a subinterval of I, and similarly for M^+.

(c) Repeat (a) with

$$M_+ = \sup\{D_+f(x) : x \in I\}$$
$$m_+ = \inf\{D_+f(x) : x \in I\}$$

and with M^-, m^- and M_-, m_- defined similarly using D^-f and D_-f. Conclude that

$$m^+ = m_+ = m^- = m_- \text{ and } M^+ = M_+ = M^- = M_- \ .$$

(d) Use (a) to give an alternate proof of the Darboux property for derivatives. [Show that if f is differentiable on an open interval I, then $f'(J)$ is an interval for every open subinterval J of I.]

V.8.6.6. Show that if f is as in V.8.5.8., and (f_n) is the sequence defined in the proof, then $f_n \to f$ u.c. on I.

V.8.6.7. In [**?**, IX, Book IV, Chapter II, §1.3], a function on an interval I is called *regulated* (French *réglée*, also sometimes called *ruled*) on I if it is a u.c. limit of a sequence of step functions on I. Note that every regulated function has a primitive by V.8.5.5. and V.8.5.4.. By V.8.5.8. and V.8.5.11., every continuous function and every monotone function is regulated.

(a) Show that a function on an interval I which has left and right limits at every interior point of I has only countably many discontinuities on I. [Mimic the proof of V.4.2.11..]

(b) Show that a function f on an interval I is regulated if and only if it has left and right limits at every interior point of I and one-sided limits at any endpoints of I which are in I.

(c) Show that if f is regulated on I and F is a primitive for f on I, then the (countable) set A of all $x \in I$ for which $F'(x) = f(x)$ does not hold is precisely the set of discontinuities of f.

(d) Show that a regulated function on an interval I is bounded on every closed bounded subinterval of I. [Step functions are bounded.]

(e) Any (finite) linear combination of regulated functions on an interval I is regulated. Thus every function of bounded variation () on every closed bounded subinterval of I is regulated on I. Not every regulated function (indeed, not every continuous function) has bounded variation, even on closed bounded subintervals of I ().

See IX.2.12.4. for more on regulated functions.

V.8.6.8. (a) Show by direct calculation that every polynomial has an antiderivative on $\mathbb{R}$.

(b) Use V.8.5.5. and XX.10.1.1. to give an alternate proof of V.8.5.9..

V.8.6.9. [**?**, Exercise 7.6.21] (a) Prove the following generalization of Cauchy's Mean Value Theorem:

THEOREM. Let f, g, h be continuous on $[a, b]$ and differentiable on (a, b). Then there is a $c \in (a, b)$ such that

$$0 = Det \begin{bmatrix} f(a) & g(a) & h(a) \\ f(b) & g(b) & h(b) \\ f'(c) & g'(c) & h'(c) \end{bmatrix}$$

$$= f'(c)[g(a)h(b) - g(b)h(a)] - g'(c)[f(a)h(b) - f(b)h(a)] + h'(c)[f(a)g(b) - f(b)g(a)] \ .$$

Cauchy's MVT is the special case where h is the constant function 1. [Consider the function

$$\phi(x) = Det \begin{bmatrix} f(a) & g(a) & h(a) \\ f(b) & g(b) & h(b) \\ f(x) & g(x) & h(x) \end{bmatrix}$$

and apply Rolle's Theorem. Recall that a (square) matrix with two identical rows has determinant 0.]

(b) Extend the result to the case $n > 3$:

THEOREM. Let $n \geq 4$, and let $f_1, \ldots, f_n$ be continuous on $[a, b]$ and differentiable on (a, b). Let

$$\{d_{ij} : 3 \leq i \leq n-1, 1 \leq j \leq n\}$$

be arbitrary constants. Then there is a $c \in (a, b)$ such that

$$0 = Det \begin{bmatrix} f_1(a) & f_2(a) & \cdots & f_n(a) \\ f_1(b) & f_2(b) & \cdots & f_n(b) \\ d_{31} & d_{32} & \cdots & d_{3n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n-1,1} & d_{n-1,2} & \cdots & d_{n-1,n} \\ f_1'(c) & f_2'(c) & \cdots & f_n'(c) \end{bmatrix}.$$

(c) Here is another version:

THEOREM. Let $n \geq 3$, and let $\mathcal{P} = \{a = x_1, x_2, \ldots, x_{n-1} = b\}$ be a partition of $[a, b]$ into $n-2$ subintervals. Let $f_1, \ldots, f_n$ be continuous on $[a, b]$ and differentiable on $[a, b] \setminus \mathcal{P}$. Then, for each $k \leq n-2$, there is a $c \in (x_k, x_{k+1})$ such that

$$0 = Det \begin{bmatrix} f_1(x_1) & f_2(x_1) & \cdots & f_n(x_1) \\ f_1(x_2) & f_2(x_2) & \cdots & f_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_1(x_{n-1}) & f_2(x_{n-1}) & \cdots & f_n(x_{n-1}) \\ f_1'(c) & f_2'(c) & \cdots & f_n'(c) \end{bmatrix}.$$

V.8.6.10. The Mean Value Inequality. This problem gives a direct proof of V.8.2.7. (in the case where f is differentiable on all of $[a, b]$) by the Method of Successive Bisections ().

(a) Show that to prove V.8.2.7. it suffices to prove V.8.2.1..

(b) Show that to prove V.8.2.1. it suffices to prove the following. [For each $\epsilon > 0$ set $f_\epsilon(x) = f(x) + \epsilon x$. To show that $f(a) \leq f(b)$, it suffices to show $f_\epsilon(a) \leq f_\epsilon(b)$ for all $\epsilon > 0$.]

PROPOSITION. Let f be differentiable on $[a, b]$. If $f'(x) > 0$ for all $x \in [a, b]$, then $f(a) \leq f(b)$.

(c) To prove this proposition, set $I_1 = [a, b]$, and c the midpoint of I_1. If $f(a) > f(b)$, then either $f(a) > f(c)$ or $f(c) > f(b)$. Let $I_2 = [a, c]$ if the first holds and $I_2 = [c, b]$ otherwise. Repeat the process to obtain a nested sequence (I_n) of intervals, where $I_n = [a_n, b_n]$ with $f(a_n) > f(b_n)$ for all n.

(d) Let $\cap_{n=1}^{\infty} I_n = \{x\}$. Show that $f'(x) \leq 0$, contradicting that $f'(x) > 0$.

V.8.6.11. The aim of this problem is to use a purely algebraic argument, along with the Intermediate Value Theorem for polynomials, to prove Rolle's Theorem for polynomials: between any two real roots of a polynomial p with real coefficients is a root of p'.

(a) Let $p(x)$ be a nonzero polynomial with real coefficients, and a and b consecutive real roots of p. Write

$$p(x) = (x-a)^r(x-b)^s q(x)$$

where $r, s \in \mathbb{N}$ and q is a polynomial with real coefficients with no roots in $[a,b]$, and which therefore does not change sign on $[a,b]$.

(b) Use the product rule to show that

$$p'(x) = (x-a)^{r-1}(x-b)^{s-1}\left[r(x-b)q(x) + s(x-a)q(x) + (x-a)(x-b)q'(x)\right] .$$

(c) Show that the polynomial in brackets in (b) takes values of opposite sign at a and b.

(d) Apply the Intermediate Value Theorem for polynomials to conclude that p' has a root in (a,b).

No actual differential calculus is needed for this problem, since the derivative of a polynomial can be defined and calculated formally, and the product rule for polynomials proved algebraically from the formal definition of derivative; cf. Exercise V.3.10.9.. (Even the Intermediate Value Theorem for polynomials has a purely algebraic proof in real-closed fields; see, for example, [BPR06].) In fact, Rolle did something like this, and it is unclear whether he even realized he was computing derivatives: he generated the sequence of derivatives of the given polynomial p in a somewhat different (algebraic) way and called them the "cascade of polynomials" in using them to locate the (real) roots of p. See the chapter by J. Barrow-Green in [RS09].

V.8.6.12. Let f be differentiable on an interval I, and suppose $f'(x) > 0$ for all $x \in I$. Use V.3.6.8. and the Max-Min Theorem () to prove that f is strictly increasing on I without using the MVT. [If $x_1, x_2 \in I$, $x_1 < x_2$, apply the Max-Min Theorem to $[x_1, x_2]$.]

V.8.6.13. Let f be continuous on an interval I and n-times differentiable on the interior of I. Suppose f has m distinct zeroes in I. Show by repeated application of Rolle's Theorem that for $1 \leq k \leq \min(n,m)$, $f^{(k)}$ has at least $m-k$ distinct zeroes in the interior of I.

V.9. L'Hôpital's Rule

L'Hôpital's Rule is a very useful technique for evaluating limits of indeterminate forms involving quotients. There are a number of variations of the rule, which all fall under the general name "l'Hôpital's Rule."

The rule is named for the French nobleman Marquis Guillaume François Antoine de l'Hôpital (sometimes spelled l'Hospital; his own spelling was apparently Lhospital) and was first published in 1696 in the calculus text *L'Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes* (the first actual calculus text!) written by l'Hôpital; this name for the rule apparently first appeared in print in [Gou59] in 1902. The rule (or at least one form of it) was actually discovered by Johann (or Jean) Bernoulli, who had taught calculus to l'Hôpital, in 1694; Bernoulli had a business arrangement with l'Hôpital giving l'Hôpital, who was himself a respectable mathematician, the right to publish Bernoulli's results. (See [**?**] for the full story.) L'Hôpital's text, which fairly closely followed lecture notes by Bernoulli, was very influential, and editions remained in print for almost 100 years (a copy is available online at http://www.archive.org/details/infinimentpetits1716lhos00uoft); before this text became available, calculus was known and understood by only a handful of top European mathematicians. L'Hôpital, besides being a gifted expositor (he also wrote a text on conic sections which was the standard for that subject throughout the eighteenth century), seems to have been a considerably more pleasant and genial person than Bernoulli: Bernoulli, although he was a great mathematician and had a reputation as an inspiring teacher (Euler was his prize pupil), was hypercompetitive and was continually embroiled in bitter disputes with other mathematicians, the most vicious of which were with his own brother Jakob (who was himself described as "self-willed, obstinate, aggressive, vindictive, beset by feelings of inferiority and yet firmly convinced of his own abilities.")![15] [Hof71] [Do not confuse the various Bernoullis; there were several outstanding mathematicians in this family, extending over at least three generations, but Johann and Jakob (or Jacques), and Johann's son Daniel, were the best.]

V.9.1. L'Hôpital's Rule, Version 1

The first version we give of l'Hôpital's Rule is an easy consequence of the definition of derivative, and is often not regarded as a form of l'Hôpital's Rule at all, although it is the version stated (imprecisely) by l'Hôpital himself (and, apparently, by Johann Bernoulli).

V.9.1.1. Theorem. [L'Hôpital's Rule, Version 1] Let f and g be functions on an interval I, and a an interior point of I. Suppose

(1) $\lim_{x\to a} f(x) = \lim_{x\to a} g(x) = 0$.

(2) f and g are differentiable at a and $g'(a) \neq 0$.

Then $\lim_{x\to a} \frac{f(x)}{g(x)}$ exists and equals $\frac{f'(a)}{g'(a)}$.

[15] For example, when Johann solved the catenary problem before Jakob, he gloated, "The efforts of my brother were without success; for my part, I was more fortunate, for I found the skill (I say it without boasting, why should I conceal the truth?) to solve it in full." (cf. [Dun90]). He added: "You know the disposition of my brother. He would sooner have taken away from me, if he could have done so honestly, the honor of being the first to solve it, rather than letting me take part by myself, let alone ceding the place, if it had really been his." ([**?**]). Johann was, however, known to try to pass off results of both Jakob and Daniel as his own, and treated Daniel vindictively. Jakob for his part was often petulantly resentful of the mathematical abilities of his younger brother. In [**?**] the Bernoullis were called "the most dysfunctional family in the history of mathematics."

PROOF: First note that since f and g are differentiable at a, they are continuous at a, and thus $f(a) = \lim_{x\to a} f(x) = 0$ and $g(a) = \lim_{x\to a} g(x) = 0$. We have

$$\frac{f'(a)}{g'(a)} = \frac{\lim\limits_{x\to a} \dfrac{f(x)-f(a)}{x-a}}{\lim\limits_{x\to a} \dfrac{g(x)-g(a)}{x-a}}$$

Note that $g(x) = g(x) - g(a) \neq 0$ in a deleted neighborhood of a since

$$g'(a) = \lim_{x\to a} \frac{g(x)-g(a)}{x-a} \neq 0 \ .$$

So, by (),

$$\frac{f'(a)}{g'(a)} = \lim_{x\to a} \left[\frac{\frac{f(x)-f(a)}{x-a}}{\frac{g(x)-g(a)}{x-a}} \right] = \lim_{x\to a} \frac{f(x)-f(a)}{g(x)-g(a)} = \lim_{x\to a} \frac{f(x)}{g(x)} \ .$$



V.9.1.2. This argument has an informal but geometrically appealing interpretation. If x changes from a by Δx, then $f(x)$ changes from $f(a)(= 0)$ by Δf, which by () is approximately $df = f'(a)\,\Delta x$ if Δx is small. Similarly, $g(x)$ changes from $g(a)$ by approximately $dg = g'(a)\,\Delta x$. So, if x is close to a,

$$\frac{f(x)}{g(x)} = \frac{\Delta f}{\Delta g} \approx \frac{df}{dg} = \frac{f'(a)\,\Delta x}{g'(a)\,\Delta x} = \frac{f'(a)}{g'(a)} \ .$$

In fact, it seems likely that BERNOULLI used an argument like this to discover the rule.

Version 1 of the rule is adequate for many of the elementary applications.

V.9.1.3. EXAMPLE. To evaluate

$$\lim_{x\to 0} \frac{e^x - 1}{\sin x}$$

let $f(x) = e^x - 1$, $g(x) = \sin x$, $a = 0$. We have $\lim_{x\to 0} f(x) = \lim_{x\to 0} g(x) = 0$, $f'(0) = g'(0) = 1$, so the hypotheses of Version I are satisfied and

$$\lim_{x\to 0} \frac{e^x - 1}{\sin x} = \frac{f'(0)}{g'(0)} = \frac{1}{1} = 1 \ .$$

(Note that $g(x) \neq 0$ on $(-\pi, \pi) \setminus \{0\}$.)

V.9.2. L'Hôpital's Rule, Version 2

There is another version (the commonly stated version) of l'Hôpital's Rule which is more complicated to prove, but more widely applicable. There are several slight variations of this version, which will collectively be called Version 2 of l'Hôpital's Rule.

V.9.2.4. THEOREM. [L'HÔPITAL'S RULE, VERSION 2A] Let f and g be differentiable functions on an interval (a,b), with g' nonzero on (a,b). Suppose

(1) $\lim_{x\to a+} f(x) = \lim_{x\to a+} g(x) = 0$.

(2) $\lim_{x\to a+} \dfrac{f'(x)}{g'(x)}$ exists in either the usual () or extended () sense.

Then $\lim_{x\to a+} \dfrac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to a+} \dfrac{f'(x)}{g'(x)}$.

Note that it is *not* quite correct to say that if f and g are differentiable on (a,b) with g' nonzero, and $\lim_{x\to a+} f(x) = \lim_{x\to a+} g(x) = 0$, then

$$\lim_{x\to a+} \frac{f(x)}{g(x)} = \lim_{x\to a+} \frac{f'(x)}{g'(x)} \ ;$$

it can happen that $\lim_{x\to a+} \dfrac{f(x)}{g(x)}$ exists but $\lim_{x\to a+} \dfrac{f'(x)}{g'(x)}$ does not exist (Exercise V.9.4.4.), even if f and g are C^∞ and $g(x)$ and $g'(x)$ are nonzero for $x > a$ (Exercise V.9.4.7..)

PROOF: If necessary, extend f and g to be continuous on $[a,b)$ by setting $f(a) = g(a) = 0$. If $x \in (a,b)$, since g is continuous on $[a,x]$ and differentiable on (a,x), by the Mean Value Theorem () we have

$$g(x) - g(a) = g'(d)(x-a)$$

for some $d \in (a,x)$; since $g'(d) \neq 0$, we have $g(x) = g(x) - g(a) \neq 0$. Thus $\frac{f}{g}$ is defined on (a,b).

Set $L = \lim_{x\to a+} \dfrac{f'(x)}{g'(x)} \in \mathbb{R} \cup \{\pm\infty\}$. By the Sequential Criterion (), it suffices to show that if (x_n) is a sequence in (a,b) converging to a, we have $\lim_{n\to\infty} \frac{f(x_n)}{g(x_n)} = L$. By the Cauchy Mean Value Theorem we have, for each n,

$$\frac{f(x_n)}{g(x_n)} = \frac{f(x_n) - f(a)}{g(x_n) - g(a)} = \frac{f'(c_n)}{g'(c_n)}$$

for some $c_n \in (a, x_n)$. Since $a < c_n < x_n$ and $x_n \to a$, we have $c_n \to a$ by the Squeeze Theorem () and hence

$$\frac{f(x_n)}{g(x_n)} = \frac{f'(c_n)}{g'(c_n)} \to L$$

by the Sequential Criterion applied to $\lim_{x\to a+} \dfrac{f'(x)}{g'(x)} = L$.

V.9.2.5. Actually, the hypothesis that g' is nonzero on (a,b) is redundant: it is implicit in the existence of the limit $\lim_{x\to a+} \frac{f'(x)}{g'(x)}$ that $\frac{f'}{g'}$ is defined, i.e. $g' \neq 0$, in an open interval (a,r) for some $r > a$. Thus the hypothesis can be eliminated at the possible price of reducing the b to a smaller number still strictly larger than a. We have left the requirement in the statement of the theorem as a reminder that this hypothesis must be checked when the theorem is applied (cf. Exercise V.9.4.5..)

Note that the hypothesis that $g' \neq 0$ on (a,b) implies that g' does not change sign on (a,b) (V.8.2.14.).

An essentially identical argument shows:

V.9.2.6. THEOREM. [L'HÔPITAL'S RULE, VERSION 2B] Let f and g be differentiable functions on an interval (a, b), with g' nonzero on (a, b). Suppose

(1) $\lim_{x\to b-} f(x) = \lim_{x\to b-} g(x) = 0$.

(2) $\lim_{x\to b-} \dfrac{f'(x)}{g'(x)}$ exists in either the usual or extended sense.

Then $\lim_{x\to b-} \dfrac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to b-} \dfrac{f'(x)}{g'(x)}$.

Combining the two previous versions together, we get:

V.9.2.7. THEOREM. [L'HÔPITAL'S RULE, VERSION 2C] Let f and g be differentiable functions on a deleted neighborhood U of $a \in \mathbb{R}$, with g' nonzero on U. Suppose

(1) $\lim_{x\to a} f(x) = \lim_{x\to a} g(x) = 0$.

(2) $\lim_{x\to a} \dfrac{f'(x)}{g'(x)}$ exists in either the usual or extended sense.

Then $\lim_{x\to a} \dfrac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to a} \dfrac{f'(x)}{g'(x)}$.

There is also a version for limits at ∞:

V.9.2.8. THEOREM. [L'HÔPITAL'S RULE, VERSION 2D] Let f and g be differentiable functions on an interval $(a, +\infty)$, with g' nonzero on $(a, +\infty)$. Suppose

(1) $\lim_{x\to+\infty} f(x) = \lim_{x\to+\infty} g(x) = 0$.

(2) $\lim_{x\to+\infty} \dfrac{f'(x)}{g'(x)}$ exists in either the usual or extended sense.

Then $\lim_{x\to+\infty} \dfrac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to+\infty} \dfrac{f'(x)}{g'(x)}$.
A similar statement holds for limits as $x \to -\infty$ for functions differentiable on an interval $(-\infty, a)$.

PROOF: We may assume $a > 0$. Set $L = \lim_{x\to+\infty} \frac{f'(x)}{g'(x)}$. For $t \in \left(0, \frac{1}{a}\right)$, set $F(t) = f(1/t)$ and $G(t) = g(1/t)$. Then F and G are differentiable on $\left(0, \frac{1}{a}\right)$, $G'(t) = -\frac{1}{t^2} g'(1/t) \neq 0$, and $G(t) \neq 0$. We have

$$\lim_{t\to 0+} \frac{F'(t)}{G'(t)} = \lim_{t\to 0+} \frac{-\frac{1}{t^2} f'(1/t)}{-\frac{1}{t^2} g'(1/t)} = \lim_{t\to 0+} \frac{f'(1/t)}{g'(1/t)} = \lim_{x\to+\infty} \frac{f'(x)}{g'(x)} = L$$

and so, by V.9.2.4.,

$$\lim_{x\to+\infty}\frac{f(x)}{g(x)}=\lim_{t\to0+}\frac{F(t)}{G(t)}=L\ .$$



L'Hôpital's Rule also applies to indeterminate forms of type $\frac{\infty}{\infty}$. The statements are essentially the same, but the proof is a little more complicated. In fact, it is unnecessary to assume the limit of the numerator is $\pm\infty$ (if $\lim\frac{f(x)}{g(x)}\neq0$ and $\lim g(x)=\pm\infty$, it follows automatically that $\lim f(x)=\pm\infty$; this can also be proved directly from $\lim\frac{f'(x)}{g'(x)}\neq0$ without applying l'Hôpital's rule, cf. V.9.4.8.).

V.9.2.9. THEOREM. [L'HÔPITAL'S RULE, VERSION 2A′] Let f and g be differentiable functions on an interval (a,b), with g' nonzero on (a,b). Suppose

(1) $\lim_{x\to a+}g(x)=\pm\infty$.

(2) $\lim_{x\to a+}\frac{f'(x)}{g'(x)}$ exists in either the usual or extended sense.

Then $\lim_{x\to a+}\frac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to a+}\frac{f'(x)}{g'(x)}$.

V.9.2.10. THEOREM. [L'HÔPITAL'S RULE, VERSION 2B′] Let f and g be differentiable functions on an interval (a,b), with g' nonzero on (a,b). Suppose

(1) $\lim_{x\to b-}g(x)=\pm\infty$.

(2) $\lim_{x\to b-}\frac{f'(x)}{g'(x)}$ exists in either the usual or extended sense.

Then $\lim_{x\to b-}\frac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to b-}\frac{f'(x)}{g'(x)}$.

V.9.2.11. THEOREM. [L'HÔPITAL'S RULE, VERSION 2C′] Let f and g be differentiable functions on a deleted neighborhood U of $a\in\mathbb{R}$, with g' nonzero on U. Suppose

(1) $\lim_{x\to a}g(x)=\pm\infty$.

(2) $\lim_{x\to a}\frac{f'(x)}{g'(x)}$ exists in either the usual or extended sense.

Then $\lim_{x\to a}\frac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to a}\frac{f'(x)}{g'(x)}$.

V.9.2.12. THEOREM. [L'HÔPITAL'S RULE, VERSION 2D′] Let f and g be differentiable functions on an interval $(a, +\infty)$, with g' nonzero on $(a, +\infty)$. Suppose

(1) $\lim_{x\to+\infty} g(x) = \pm\infty$.

(2) $\lim_{x\to+\infty} \dfrac{f'(x)}{g'(x)}$ exists in either the usual or extended sense.

Then $\lim_{x\to+\infty} \dfrac{f(x)}{g(x)}$ exists (in the same sense) and equals $\lim_{x\to+\infty} \dfrac{f'(x)}{g'(x)}$.
A similar statement holds for limits as $x \to -\infty$ for functions differentiable on an interval $(-\infty, a)$.

PROOF: We give the proof of Version 2d′ for variety; proofs of the other versions are similar.

We first show that we may assume $g \neq 0$ on (a, ∞). We could give an argument using the Mean Value Theorem (cf. Exercise ()), but there is a simpler argument: since $\lim_{x\to+\infty} g(x) = \pm\infty$, there is automatically an interval $(r, +\infty)$ on which $|g| > 1$, say. Thus by increasing a if necessary we may assume that $g \neq 0$ on (a, ∞).

Let $L = \lim_{x\to+\infty} \dfrac{f'(x)}{g'(x)}$. As before, it suffices to show that if (x_n) is a sequence in $(a, +\infty)$ with $x_n \to +\infty$, then $\lim_{n\to\infty} \dfrac{f(x_n)}{g(x_n)} = L$. We will assume $L \in \mathbb{R}$; the case $L = \pm\infty$ is an easy variation. By Cauchy's Mean Value Theorem, for each m and n for which $x_m \neq x_n$ there is a $d_{m,n}$ between x_m and x_n such that

$$f(x_n) - f(x_m) = \frac{f'(d_{m,n})}{g'(d_{m,n})}[g(x_n) - g(x_m)] \ .$$

We then have

$$\frac{f(x_n)}{g(x_n)} - \frac{f(x_m)}{g(x_n)} = \frac{f(x_n) - f(x_m)}{g(x_n)} = \frac{f'(d_{m,n})}{g'(d_{m,n})}\frac{g(x_n) - g(x_m)}{g(x_n)}$$

$$\frac{f(x_n)}{g(x_n)} - L = \left[\frac{f'(d_{m,n})}{g'(d_{m,n})} - L\right] - \frac{f'(d_{m,n})}{g'(d_{m,n})}\frac{g(x_m)}{g(x_n)} + \frac{f(x_m)}{g(x_n)} \ .$$

Fix $\epsilon > 0$ and let R such that

$$\left|\frac{f'(x)}{g'(x)} - L\right| < \frac{\epsilon}{3}$$

for $x > R$, and fix m_0 such that $x_n > R$ for all $n \geq m_0$. Thus, for $n > m_0$,

$$\left|\frac{f'(d_{m_0,n})}{g'(d_{m_0,n})} - L\right| < \frac{\epsilon}{3} \ .$$

Set $M = |L| + \frac{\epsilon}{3}$. Since $\lim_{x\to+\infty} |g(x)| = +\infty$, choose $N > m_0$ such that $\left|\dfrac{f(x_{m_0})}{g(x_n)}\right| < \dfrac{\epsilon}{3}$ and $\left|\dfrac{g(x_{m_0})}{g(x_n)}\right| < \dfrac{\epsilon}{3M}$ for all $n \geq N$. Then, for $n \geq N$,

$$\left|\frac{f(x_n)}{g(x_n)} - L\right| \leq \left|\frac{f'(d_{m_0,n})}{g'(d_{m_0,n})} - L\right| + \left|\frac{f'(d_{m_0,n})}{g'(d_{m_0,n})}\right|\left|\frac{g(x_{m_0})}{g(x_n)}\right| + \left|\frac{f(x_{m_0})}{g(x_n)}\right| < \epsilon.$$

V.9.2.13. In fact, if g' is never zero and $\lim \dfrac{f'(x)}{g'(x)}$ exists, then $\lim \dfrac{f(x)}{g(x)}$ always exists, but does not necessarily equal $\lim \dfrac{f'(x)}{g'(x)}$ unless $\lim f(x) = \lim g(x) = 0$ or $\lim g(x) = \pm\infty$ (Exercise V.9.4.9.; statements apply to any kind of one-sided limits, but not two-sided).

V.9.2.14. Version 2 of l'Hôpital's Rule (in its various forms) is commonly used even when Version 1 applies. There are also situations where Version 2 can be used although Version 1 does not apply. (The reverse can also happen: there are situations where Version 1 applies but Version 2 does not; cf. Exercise V.9.4.4..)

V.9.3. Applications

V.9.3.1. Examples. (i) As in V.9.1.3., we can apply Version 2c to obtain

$$\lim_{x\to 0} \frac{e^x - 1}{\sin x} = \lim_{x\to 0} \frac{e^x}{\cos x} = 1 \ .$$

(ii) An example of a limit where Version 1 does not apply but Version 2 does is:

$$\lim_{x\to 0+} \frac{\sin x}{\sqrt{x}} \ .$$

Version 2a says that if

$$\lim_{x\to 0+} \frac{\cos x}{\frac{1}{2}x^{-1/2}} = \lim_{x\to 0+} 2\sqrt{x}\cos x$$

exists, the first limit also exists and has the same value. But the second limit is 0 by (); thus

$$\lim_{x\to 0+} \frac{\sin x}{\sqrt{x}} = 0 \ .$$

This limit could have also been evaluated by

$$\lim_{x\to 0+} \frac{\sin x}{\sqrt{x}} = \lim_{x\to 0+} \left[\sqrt{x}\cdot \frac{\sin x}{x}\right] = 0\cdot 1 = 0 \ .$$

(iii) Besides applications of Versions 2d and 2a′–2d′, which have no Version 1 analogues, the most common situation where Version 2 is necessary (i.e. Version 1 does not apply) is illustrated by the following example. Let $f(x) = 1 - \cos x$, $g(x) = x^2$. Then

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = \lim_{x\to 0} \frac{1-\cos x}{x^2}$$

is an indeterminate form of type $\frac{0}{0}$ to which we might try to apply l'Hôpital's Rule. If

$$\lim_{x\to 0} \frac{f'(x)}{g'(x)} = \lim_{x\to 0} \frac{\sin x}{2x}$$

exists, then Version 2c implies that the first limit exists and has the same value. The second limit is also an indeterminate form of type $\frac{0}{0}$, so we may apply Version 2c again to convert it to

$$\lim_{x\to 0} \frac{f''(x)}{g''(x)} = \lim_{x\to 0} \frac{\cos x}{2}$$

i.e. if this limit exists, then the second limit and thus also the first limit exists and has the same value. But the last limit is not an indeterminate form, and equals $\frac{1}{2}$ by direct evaluation. We usually just abbreviate the whole argument as

$$\lim_{x\to 0} \frac{1-\cos x}{x^2} = \lim_{x\to 0} \frac{\sin x}{2x} = \lim_{x\to 0} \frac{\cos x}{2} = \frac{1}{2}$$

where the first two steps are justified by "l'Hôpital's Rule" (specifically, Version 2c); logically, we technically do the second application first and use the result to justify the first application. We could also justify the first step by Version 2c and calculate the second limit directly by Version 1.

(iv) In applying l'Hôpital's Rule, it must be checked that the hypotheses apply. This pitfall is particularly important to watch for when the Rule is used repeatedly as in (iii). For example, one might argue

$$\lim_{x\to 0} \frac{x^2}{x^2+\sin x} = \lim_{x\to 0} \frac{2x}{2x+\cos x} = \lim_{x\to 0} \frac{2}{2-\sin x} = \frac{2}{2} = 1 \ .$$

However, the second limit is not an indeterminate form and the second application of l'Hôpital's Rule is not justified. The second limit can be evaluated directly to be 0, and thus the first limit is also 0.

See Exercise V.9.4.7. for a more subtle example where the hypotheses of l'Hôpital's Rule are not satisfied.

Sometimes some cleverness is required to successfully use l'Hôpital's Rule:

V.9.3.2. Example. Consider

$$\lim_{x\to 0+} \frac{e^{-1/x}}{x} \ .$$

The hypotheses of Version 2a are satisfied with $f(x) = e^{-1/x}$ and $g(x) = x$. Thus the limit can be converted to

$$\lim_{x\to 0+} \frac{e^{-1/x}\frac{1}{x^2}}{1} = \lim_{x\to 0+} \frac{e^{-1/x}}{x^2} \ .$$

But this limit is worse than the one we started with, and repeating l'Hôpital's Rule will make things successively even worse.

Instead, we can rewrite the limit as

$$\lim_{x\to 0+} \frac{1/x}{e^{1/x}}$$

and note that Version 2a′ applies to give

$$\lim_{x\to 0+} \frac{1/x}{e^{1/x}} = \lim_{x\to 0+} \frac{-1/x^2}{-e^{1/x}\frac{1}{x^2}} = \lim_{x\to 0+} e^{-1/x} = 0 \ .$$

We could also make the substitution $u = 1/x$ and use Version 2d′ to give

$$\lim_{x\to 0+} \frac{1/x}{e^{1/x}} = \lim_{u\to+\infty} \frac{u}{e^u} = \lim_{u\to+\infty} \frac{1}{e^u} = 0 \ .$$

Indeterminate forms of other types can often be converted into a form to which l'Hôpital's rule can be applied, by algebraic manipulation or taking logarithms or exponentials:

V.9.3.3. EXAMPLES. (i) $\lim_{x\to+\infty} xe^{-x}$ is an indeterminate form of type $\infty \cdot 0$. It can be converted to $\lim_{x\to+\infty} \frac{x}{e^x}$ and evaluated by Version 2d′. Another similar example is in Exercise ()(d).

(ii) $\lim_{x\to 0} \left[\frac{1}{\sin x} - \frac{1}{x}\right]$ is an indeterminate form of type $\infty - \infty$. It can be converted into a fraction and evaluated by

$$\lim_{x\to 0} \left[\frac{1}{\sin x} - \frac{1}{x}\right] = \lim_{x\to 0} \frac{x - \sin x}{x \sin x}$$

$$= \lim_{x\to 0} \frac{1 - \cos x}{\sin x + x\cos x} = \lim_{x\to 0} \frac{\sin x}{2\cos x - x\sin x} = \frac{0}{2} = 0$$

by two applications of l'Hôpital's Rule. (For the first application, it must be verified that $h(x) = \sin x + x\cos x$ is nonzero in a deleted neighborhood of 0; this follows from the fact that $h'(0) \neq 0$.)

(iii) $\lim_{x\to 0+} x^x$ is an indeterminate form of type 0^0. By taking logarithms we can reduce the problem to computing $\lim_{x\to 0+} x\log x$, which is 0 (Exercise ()). Thus

$$\lim_{x\to 0+} x^x = \lim_{x\to 0+} e^{x\log x} = e^0 = 1$$

by continuity of the exponential function at 0. Similarly,

$$\lim_{x\to+\infty} x^{1/x} = 1$$

(indeterminate form of type ∞^0). See IV.1.12.8. for a direct argument.

(iv) $\lim_{x\to+\infty} \left(1 + \frac{1}{x}\right)^x$ is of the form 1^∞. Taking logarithms, we reduce to evaluating $\lim_{x\to+\infty} x\log\left(1 + \frac{1}{x}\right)$. This is of type $\infty \cdot 0$, so we can convert it to

$$\lim_{x\to+\infty} \frac{\log(1 + 1/x)}{1/x} = \lim_{x\to+\infty} \frac{\frac{1}{1+1/x}(-1/x^2)}{-1/x^2} = \lim_{x\to+\infty} \frac{1}{1 + 1/x} = 1$$

so

$$\lim_{x\to+\infty} \left(1 + \frac{1}{x}\right)^x = e^1 = e$$

again by continuity of the exponential function. (This example is somewhat artificial, since direct evaluation of this limit was used in defining e and the exponential function.)

V.9.4. Exercises

V.9.4.1. Express and justify Version 1 of l'Hôpital's Rule using little-o notation (V.7.) by writing $f(x) = f'(a)(x-a) + o(x-a)$ and $g(x) = g'(a)(x-a) + o(x-a)$.

V.9.4.2. Show that Version 1 and Version 2c of l'Hôpital's Rule are equivalent in the (common) case where f and g are differentiable in a neighborhood of a, f' and g' are continuous at a, $f(a) = g(a) = 0$, and $g'(a) \neq 0$.

V.9.4.3. Let $f(x) = x^2$ for $x \in \mathbb{Q}$ and $f(x) = 0$ for $x \notin \mathbb{Q}$, and let $g(x) = x$ for all x. Then Version 1 of L'Hôpital's Rule applies for $a = 0$. But f is not differentiable, or even continuous, except at 0, so no variant of Version 2 applies.

V.9.4.4. Let $f(x) = x^2 \sin \frac{1}{x^2}$ for $x \neq 0$, and $f(0) = 0$; and let $g(x) = \sin x$. Then f and g are differentiable everywhere, and $f'(0) = 0$ (), $g'(0) = 1$.

(a) Version 1 of L'Hôpital's Rule applies to show that

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = \lim_{x\to 0} \frac{x^2 \sin \frac{1}{x^2}}{\sin x} = \frac{0}{1} = 0 \ .$$

(With a moderate amount of work, this limit can be evaluated directly using the Squeeze Theorem instead of L'Hôpital's Rule.)

(b) $\lim_{x\to 0} f'(x) = \lim_{x\to 0} \left[2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2}\right]$ and $\lim_{x\to 0} \frac{f'(x)}{g'(x)}$ do not exist. In particular, Version 2 of L'Hôpital's rule does not apply.

(c) The existence of $\lim_{x\to a} \frac{f(x)}{g(x)}$ does not imply the existence of $\lim_{x\to a} \frac{f'(x)}{g'(x)}$ even for everywhere differentiable f and g with $g' \neq 0$ in a neighborhood of a. Thus Version 2 is not more general than Version 1 even for functions which are everywhere differentiable.

(d) There are simpler examples for limits at ∞: for example, let $f(x) = x + \sin x$, $g(x) = x$. Then $\lim_{x\to+\infty} f(x) = \lim_{x\to+\infty} g(x) = +\infty$ and $\lim_{x\to+\infty} \frac{f(x)}{g(x)}$ exists (and equals 1), but $\lim_{x\to+\infty} \frac{f'(x)}{g'(x)}$ does not exist.

V.9.4.5. ([**?**]; cf. [Boa86]) What is wrong with the following argument?

(a) Let $f(x) = x + \sin x \cos x$. Show that $f'(x) = 2\cos^2 x$.

(b) Let $g(x) = e^{\sin x} f(x)$. Show that $\lim_{x\to+\infty} f(x) = \lim_{x\to+\infty} g(x) = +\infty$.

(c) We have

$$\lim_{x\to+\infty} \frac{f'(x)}{g'(x)} = \lim_{x\to+\infty} \frac{2\cos^2 x}{e^{\sin x}[f(x)\cos x + 2\cos^2 x]}$$
$$= \lim_{x\to+\infty} \frac{2\cos x}{e^{\sin x}[f(x) + 2\cos x]} = 0$$

since the numerator is bounded and the denominator approaches $+\infty$. Thus, by l'Hôpital's Rule, Version 2d′,

$$\lim_{x\to+\infty} \frac{f(x)}{g(x)} = \lim_{x\to+\infty} e^{-\sin x} = 0 \ .$$

But $e^{-1} \leq e^{-\sin x} \leq e$ for all x, and $\lim_{x\to+\infty} e^{-\sin x}$ does not exist.

V.9.4.6. (a) Using l'Hôpital's Rule, prove that, for every $a > 0$ and $r \in \mathbb{R}$,

$$\lim_{x\to+\infty} \frac{e^{ax}}{x^r} = +\infty \ .$$

[Prove this by induction for $r \in \mathbb{N}$, and use the Squeeze Theorem for other r.]
(b) Using l'Hôpital's Rule, prove that for all $r > 0$,

$$\lim_{x\to+\infty} \frac{\log x}{x^r} = 0 \ .$$

(c) Use (b) to give a proof of (a) not requiring induction.
(d) Using l'Hôpital's Rule, show that for all $r > 0$,

$$\lim_{x\to 0+} x^r \log x = \lim_{x\to 0+} \frac{\log x}{x^{-r}} = 0 \ .$$

For a proof of (a) not using l'Hôpital's Rule, note that, for every x,

$$e^{ax} = \sum_{k=0}^{\infty} \frac{a^k x^k}{k!}$$

and all terms are positive if $a > 0$ and $x > 0$, and hence for any n and m we have

$$e^{ax} > \frac{a^{n+m} x^{n+m}}{(n+m)!}$$

$$\frac{e^{ax}}{x^n} > \frac{a^{n+m}}{(n+m)!} x^m \ .$$

V.9.4.7. [**?**] Let $f(x) = x \sin\left(\frac{1}{x^4}\right) e^{-1/x^2}$ and $g(x) = e^{-1/x^2}$ for $x \neq 0$ and $f(0) = g(0) = 0$.
(a) Show that f and g are C^∞ on $\mathbb{R}$ (), and that $g(x)$ and $g'(x)$ are nonzero for $x \neq 0$.
(b) Show that $\lim_{x\to 0} \frac{f(x)}{g(x)} = 0$.
(c) Show that $\frac{f'(x)}{g'(x)}$ is unbounded in every neighborhood of 0, and that $\lim_{x\to 0} \frac{f'(x)}{g'(x)}$ does not exist.

V.9.4.8. Suppose f and g satisfy the hypotheses Version 2d′ of l'Hôpital's Rule. Recall (V.8.2.14.) that g' does not change sign on $(a, +\infty)$.
(a) Suppose $g' > 0$ on (a, ∞). If $\lim_{x\to+\infty} g(x) = +\infty$ (it cannot be $-\infty$!), and $\lim_{x\to+\infty} \frac{f'(x)}{g'(x)}$ exists and equals L, $0 < L \leq +\infty$, then there is an $M > 0$ such that $f'(x) > Mg'(x)$ for sufficiently large x. Show that this implies $\lim_{x\to+\infty} f(x) = +\infty$ [$f - Mg$ is eventually increasing]. Modify the argument for the case $g' < 0$ on (a, ∞) and/or $-\infty \leq L < 0$.
(b) Modify the argument to apply to Version 2a′, and similarly for Version 2b′ and 2c′.

V.9.4.9. (a) Suppose f and g are differentiable on $(a, +\infty)$ and g' is never zero, and that $\lim_{x\to+\infty} \frac{f'(x)}{g'(x)}$ exists and is finite. If $\lim_{x\to+\infty} g(x) = \pm\infty$ is false, show that $\lim_{x\to+\infty} g(x)$ and $\lim_{x\to+\infty} f(x)$ exist and are finite, and $\lim_{x\to+\infty} \frac{f(x)}{g(x)}$ exists (possibly in the extended sense). [Note that g must be strictly monotone. For some M, $M|g| \pm f$ are eventually increasing.] If $\lim_{x\to+\infty} g(x) \neq 0$, the condition that $\lim_{x\to+\infty} \frac{f'(x)}{g'(x)}$ exists can be relaxed to just a requirement that $\frac{f'(x)}{g'(x)}$ be bounded.

(b) Show that $\lim_{x\to+\infty} \frac{f(x)}{g(x)}$ is not necessarily the same as $\lim_{x\to+\infty} \frac{f'(x)}{g'(x)}$ (which need not even exist if $\lim_{x\to+\infty} g(x) \neq 0$), by considering

$$\lim_{x\to+\infty} \frac{1}{1-\frac{1}{x}} \quad \text{or} \quad \lim_{x\to+\infty} \frac{1-\frac{1}{x^2}}{1-\frac{1}{x}} \quad \text{or} \quad \lim_{x\to+\infty} \frac{1-\frac{\sin x}{x^2}}{1-\frac{1}{x}}$$

or by noting that constants can be added to f and/or g without affecting the hypotheses. In particular, the assumption $\lim_{x\to+\infty} g(x) = \pm\infty$ cannot be eliminated from Version 2d′.

(c) Suppose f and g are differentiable on $(a, +\infty)$ and g' is never zero, and that $\lim_{x\to+\infty} \frac{f'(x)}{g'(x)} = \pm\infty$. If $\lim_{x\to+\infty} g(x) = \pm\infty$ is false, show that $\lim_{x\to+\infty} g(x)$ and $\lim_{x\to+\infty} f(x)$ exist, $\lim_{x\to+\infty} g(x)$ is finite, and $\lim_{x\to+\infty} \frac{f(x)}{g(x)}$ exists (possibly, but not necessarily, in the extended sense).

(d) Obtain versions for $\lim_{x\to-\infty}$ and, for an interval (a, b), for $\lim_{x\to a+}$ and $\lim_{x\to b-}$. (There is no good two-sided version.)

V.9.4.10. Show that hypothesis (1) in Version 2c′ can be relaxed to

(1′) $\lim_{x\to a} |g(x)| = +\infty$.

(The same change can be made in Versions 2a′, 2b′, and 2d′, but this is not an actual change in these cases by V.8.2.14..)

V.9.4.11. Let I be an open interval, $c \in I$, and let f be continuous on I and differentiable on $I \setminus \{c\}$. If $\lim_{x\to c} f'(x)$ exists and equals $L \in \mathbb{R}$, show that f is necessarily differentiable at c and that $f'(c) = L$. [Apply l'Hôpital's Rule Version 2c to $\lim_{x\to c} \frac{f(x)-f(c)}{x-c}$.] Give a similar statement for one-sided limits, and a direct proof using the Mean Value Theorem instead of l'Hôpital's Rule.

Compare with V.8.2.15. and V.12.2.5..

V.9.4.12. Let $x_0 \in \mathbb{R}$, and $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ and $\sum_{k=0}^{\infty} b_k(x-x_0)^k$ be power series () centered at x_0 with positive radius of convergence, defining analytic functions f and g in an open interval around x_0. Let a_m be the first nonzero coefficient in the first series, i.e. $a_m \neq 0$ and $a_k = 0$ for $k < m$. Similarly, let b_n be the first nonzero coefficient in the second series.

(a) Use V.15.3.26. and l'Hôpital's Rule to show that if $m \geq n$, then

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \frac{a_n}{b_n}$$

and in particular, the limit always exists (in the usual sense), and is zero if $m > n$ and nonzero if $m = n$.

(b) Obtain the same result from basic limit and power series theorems by first dividing f and g by a suitable power of $x - x_0$, and show additionally that

$$\lim_{x \to x_0} \frac{f(x)}{g(x)}$$

does not exist (in the usual sense) if $m < n$.

(c) Also conclude the result from (), although this argument is less elementary (but shows more).

V.9.4.13. [?, 11.10] Find

$$\lim_{x \to 0} \frac{\sin(\tan(x)) - \tan(\sin(x))}{\arcsin(\arctan(x)) - \arctan(\arcsin(x))} \ .$$

[To apply l'Hôpital's Rule, or Exercise V.9.4.12., try expanding numerator and denominator in Maclaurin series; a computer algebra system can be very helpful.] Does graphing the function near 0 with a computer algebra system help in guessing the answer?

V.9.4.14. [?, 4.4] Let

$$h(x) = \frac{x}{\sqrt{x^2 + 1}}$$

and note that $h(x) > 0$ for $x > 0$. Applying (), (), and l'Hôpital's Rule Version 2d′ to $\lim_{x \to +\infty} h(x)$, the only conclusion that can be drawn is that either the limit is 1 or it does not exist. Use another method to determine which is the case.

V.9.4.15. Let f be a function defined on an open interval around a.

(a) Suppose $f''(a)$ exists. Show that

$$\lim_{h \to 0} \frac{f(a+h) - 2f(a) + f(a-h)}{h^2}$$

exists and equals $f''(a)$.

(b) Show that the limit in (a) can exist even if $f''(a)$ does not exist, and even if f is not continuous at a. [Consider an odd function with $a = 0$.]

V.10. Taylor's Theorem

Taylor's Theorem is one of the most important and useful results in calculus. It can be thought of as a higher-order version of differential approximation or of the Mean Value Theorem, and is crucial in representing functions by power series, which will be discussed later. There are many other applications of both theoretical and practical importance. See III.5.3.8. and V.3.10.9. for purely algebraic versions.

The statement and a "proof" of Taylor's Theorem were first published by Brook Taylor in 1715, although the result was previously known to several mathematicians, and apparently first obtained by Gregory about 1670. Taylor's "proof" fell far short of modern standards of precision and rigor. His statement also did not give error estimates; most modern texts only state later versions of the theorem which include error or remainder formulas. For an account of the history of the result, see [].

V.10.1. Taylor Polynomials

V.10.1.1. Suppose f is a function which is differentiable to some high order near a number x_0. We would like to approximate f near x_0 as closely as possible by a polynomial of fixed degree. It makes intuitive sense that the approximating polynomial should duplicate the values of f and its derivatives at x_0 as closely as possible, since these values should determine the location and shape of the graph of f near x_0. There is a limit to how far this can be done, since if p is a polynomial of degree n, then $p^{(k)}$ is identically 0 for $k > n$; so we can only hope to make the derivatives of a degree n polynomial up through the n'th derivative agree with those of f. In differential approximation, which is the case $n = 1$, we have seen that there is a unique polynomial p_1 of degree ≤ 1 (i.e. linear or affine function) with $p_1(x_0) = f(x_0)$ and $p_1'(x_0) = f'(x_0)$, namely $p_1(x) = f(x_0) + f'(x_0)(x - x_0)$, and p_1 is indeed the polynomial of degree ≤ 1 most closely approximating f near x_0, in the sense that

$$\lim_{x \to x_0} \frac{f(x) - p_1(x)}{x - x_0} = 0$$

and p_1 is the only linear function with this property.

The higher-order case gets its justification from the following result:

V.10.1.2. Theorem. Let g be a real-valued function on an interval I, and x_0 an interior point of I, and $n \geq 1$. Suppose g is n-times differentiable at x_0 (this implies that for all $k < n$, $g^{(k)}$ is defined in a neighborhood of x_0 and is continuous at x_0, and, for $1 \leq r \leq n-1$, $g^{(n-r-1)}$ is r-times differentiable in a neighborhood of x_0). Then

$$\lim_{x \to x_0} \frac{g(x)}{(x - x_0)^n}$$

exists if and only if $g^{(k)}(x_0) = 0$ for $0 \leq k < n$. If the limit exists, it equals $\frac{g^{(n)}(x_0)}{n!}$.

V.10.1.3. Corollary. Let g be a real-valued function on an interval I, and x_0 an interior point of I, and $n \geq 1$. Suppose g is n-times differentiable at x_0 (this implies that for all $k < n$, $g^{(k)}$ is defined in a neighborhood of x_0 and is continuous at x_0, and, for $1 \leq r \leq n-1$, $g^{(n-r-1)}$ is r-times differentiable in a neighborhood of x_0). Then

$$\lim_{x \to x_0} \frac{g(x)}{(x - x_0)^n} = 0$$

if and only if $g(x_0) = g'(x_0) = \cdots = g^{(n)}(x_0) = 0$.

PROOF: By (complete) induction on n. First suppose $n = 1$. By assumption, g is differentiable at x_0 and hence continuous at x_0. If

$$\lim_{x\to x_0} \frac{g(x)}{x - x_0}$$

exists, then $\lim_{x\to x_0} g(x)$ must be 0 since $\lim_{x\to x_0}(x - x_0) = 0$ (). Thus $g(x_0) = 0$. And if $g(x_0) = 0$, then

$$\lim_{x\to x_0} \frac{g(x)}{x - x_0} = \lim_{x\to x_0} \frac{g(x) - g(x_0)}{x - x_0}$$

exists and equals $g'(x_0)$ by definition.

Now let $n > 1$. Assume the result for all k, $1 \le k < n$, and suppose g is n-times differentiable at x_0. Suppose

$$\lim_{x\to x_0} \frac{g(x)}{(x - x_0)^n}$$

exists and equals L. Then, if $0 \le k < n$, we have

$$\lim_{x\to x_0} \frac{g(x)}{(x - x_0)^k} = \left[\lim_{x\to x_0} \frac{g(x)}{(x - x_0)^n}\right] \left[\lim_{x\to x_0} (x - x_0)^{n-k}\right] = L \cdot 0 = 0$$

and thus by the inductive hypothesis $g^{(k)}(x_0) = 0$ (if $k = 0$, we obtain $g(x_0) = 0$ by continuity of g at x_0). Conversely, if $g(x_0) = g'(x_0) = \cdots = g^{(n-1)}(x_0) = 0$, we apply l'Hôpital's rule Version 2c (V.9.2.7.): if

$$\lim_{x\to x_0} \frac{g'(x)}{n(x - x_0)^{n-1}}$$

exists and equals L, then

$$\lim_{x\to x_0} \frac{g(x)}{(x - x_0)^n}$$

also exists and equals L. But the inductive hypothesis applies to g' since it is $(n-1)$-times differentiable at x_0 and $g'(x_0) = (g')'(x_0) = \cdots = (g')^{(n-2)}(x_0) = 0$, so the first limit exists and

$$L = \frac{1}{n} \frac{(g')^{(n-1)}(x_0)}{(n-1)!} = \frac{g^{(n)}(x_0)}{n!} \ .$$



V.10.1.4. In particular, if p is a polynomial of degree $\le n$ and f is n-times differentiable at x_0, then

$$\lim_{x\to x_0} \frac{f(x) - p(x)}{(x - x_0)^n} = 0$$

if and only if $p^{(k)}(x_0) = f^{(k)}(x_0)$ for $0 \le k \le n$. Such a polynomial (if it exists) is unique, and is the "best n'th order polynomial approximation" to f near x_0.

So the problem is to construct a polynomial of degree $\le n$ with specified derivatives at x_0. This is easily done:

V.10.1.5. PROPOSITION. Let $p(x) = a_0 + a_1x + \cdots + a_nx^n$ be a polynomial of degree n. Then, for $0 \leq k \leq n$, $f^{(k)}(0) = k!a_k$.

The proof is a simple calculation. If derivatives at x_0 instead of 0 are desired, a simple trick does the shift: if

$$q(x) = p(x - x_0) = a_0 + a_1(x - x_0) + \cdots + a_n(x - x_0)^n \,,$$

then q is also a polynomial of degree $\leq n$ and $q^{(k)}(x_0) = k!a_k$.

V.10.1.6. Thus, if f is n-times differentiable at x_0 and we want a polynomial p_n of degree $\leq n$ whose derivatives at x_0 up to the n'th are the same as those of f, we may (and must) take

$$p_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

This is the n'th *Taylor Polynomial* of f at x_0 (it will have degree strictly less than n if $f^{(n)}(x_0) = 0$.)

Although our notation does not reflect it, note that p_n depends not only on n (and, of course, f), but also on the choice of x_0, i.e. the function f will generally have different n'th order Taylor polynomials at different points (just as the best linear or affine approximation to f generally varies from point to point.)

We summarize our work in the next theorem (see V.10.1.2. for the last statement):

V.10.1.7. THEOREM. [TAYLOR'S THEOREM] Let f be defined in a neighborhood of x_0, and $n \in \mathbb{N}$. Suppose $f^{(n)}(x_0)$ exists (this implies that for all $k < n$, $f^{(k)}$ is defined in a neighborhood of x_0 and is continuous at x_0, and, for $1 \leq r \leq n-1$, $f^{(n-r-1)}$ is r-times differentiable in a neighborhood of x_0). Then the n'th Taylor polynomial p_n of f at x_0 satisfies

$$\lim_{x \to x_0} \frac{f(x) - p_n(x)}{(x - x_0)^n} = 0$$

and p_n is the unique polynomial of degree $\leq n$ with this property. If f is $(n+1)$-times differentiable at x_0, then

$$\lim_{x \to x_0} \frac{f(x) - p_n(x)}{(x - x_0)^{n+1}} = \frac{f^{(n+1)}(x_0)}{(n+1)!} \,.$$

Note that if $n = 1$, we obtain the differential approximation polynomial p_1. Thus the Taylor polynomials can be regarded as giving "higher-order differential approximation."

V.10.1.8. The Taylor polynomials thus give good approximations to f sufficiently near x_0; rephrasing the limit characterization, for any $\epsilon > 0$ and $n \in \mathbb{N}$ there is a $\delta > 0$ such that

$$|f(x) - p_n(x)| \leq \epsilon |x - x_0|^n \text{ whenever } |x - x_0| < \delta$$

and thus, for x close enough to x_0, the difference between $f(x)$ and $p_n(x)$ is small even compared to the very small number $|x - x_0|^n$.

In general, the δ depends on x_0 and n, as well as on f and ϵ. As a result, for a *fixed* x near x_0 but different from x_0, we should not necessarily expect $p_n(x)$ to be a good approximation to $f(x)$ for any particular n, or

even that $p_n(x)$ becomes a better and better approximation to $f(x)$ as n increases, although fortunately this often happens. In fact, to paraphrase [Act90], the Taylor polynomials give a very, very good approximation close enough to x_0, then get increasingly worse as we move away, and finally (often) explode into complete unusability. These points will be discussed more carefully later.

V.10.2. Vanishing Order

V.10.2.1. DEFINITION. Let f be a function defined in a deleted neighborhood of x_0, and let $\alpha \in [0, \infty)$. Then f *vanishes to order* α *at* x_0 if $f(x) = O(|x - x_0|^\alpha)$ as $x \to x_0$, i.e. there are positive K and δ such that $|f(x)| \le K|x - x_0|^\alpha$ for $0 < |x - x_0| < \delta$.
We say f *vanishes to order* α^+ *at* x_0 if $f(x) = o(|x - x_0|^\alpha)$ as $x \to x_0$, i.e.

$$\lim_{x \to x_0} \frac{f(x)}{|x - x_0|^\alpha} = 0 \ .$$

If f vanishes to order α^+ at x_0, then it vanishes to order α. If it vanishes to order β for some $\beta > \alpha$ at x_0, then it vanishes to order α^+. A function vanishes to order 0^+ at x_0 if and only if it can be extended to be continuous at x_0 by setting $f(x_0) = 0$.

V.10.2.2. EXAMPLES. (i) The function $f(x) = |x - x_0|^\alpha$ vanishes at x_0 to order α but not to order α^+. (ii) The function $f(x) = \frac{|x-x_0|^\alpha}{\log|x-x_0|}$ (with $f(x_0) = 0$) vanishes at x_0 to order α^+, but not to order β for any $\beta > \alpha$.

Taylor's Theorem can be rephrased (the last statement becomes slightly weaker):

V.10.2.3. THEOREM. [TAYLOR'S THEOREM] Let f be n-times differentiable at x_0, and let p_n be the n'th order Taylor polynomial of f at x_0. Then $f - p_n$ vanishes to order n^+ at x_0, and p_n is the unique polynomial of degree $\le n$ with this property. If f is $(n+1)$-times differentiable at x_0, then $f - p_n$ vanishes to order $n + 1$ at x_0.

The special case V.10.1.3. can be similarly rephrased:

V.10.2.4. COROLLARY. Let g be $(n-1)$-times differentiable on an open interval I and n-times differentiable at $x_0 \in I$, with $g(x_0) = g'(x_0) = \cdots = g^{(n)}(x_0) = 0$. Then g vanishes to order n^+ at x_0, and there is a unique function h which is continuous on I (and automatically as differentiable as g on $I \setminus \{x_0\}$), with $h(x_0) = 0$ and $g(x) = (x - x_0)^n h(x)$ for all $x \in I$. If g is $(n+1)$-times differentiable at x_0, then h is differentiable at x_0 and $h'(x_0) = \frac{g^{(n+1)}(x_0)}{(n+1)!}$.

PROOF: Set $h(x) = \frac{g(x)}{(x-x_0)^n}$ for $x \ne x_0$, and $h(x_0) = 0$. The first statement follows immediately from V.10.1.3.. If g is $(n+1)$-times differentiable at x_0, then

$$h'(x_0) = \lim_{x \to x_0} \frac{h(x) - h(x_0)}{x - x_0} = \lim_{x \to x_0} \frac{g(x)}{(x - x_0)^{n+1}}$$

and this limit is $\frac{g^{(n+1)}(x_0)}{(n+1)!}$ by another application of V.10.1.3. with $n+1$ in place of n. 👍

The examples $g(x) = x^n|x|$ and $g(x) = x^{n+1}|x|$ ($x_0 = 0$) show that no higher differentiability of h at x_0 can be concluded in general. But if g is differentiable to order $r > n$ at x_0, then h is differentiable to order $r - n$ at x_0 (V.10.5.4.).

V.10.3. Remainder Formulas

V.10.3.1. It is convenient to think of the situation in terms of the *remainder function* $r_n = f - p_n$. Thus

$$r_n(x) = f(x) - p_n(x) = f(x) - f(x_0) - f'(x_0)(x - x_0) - \cdots - \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

(like p_n, the function r_n depends on x_0 as well as on f and n). The approximation in Taylor's Theorem can be summarized by

$$r_n(x) = o(|x - x_0|^n) \text{ as } x \to x_0 \ (O(|x - x_0|^{n+1}) \text{ if } f^{(n+1)}(x_0) \text{ exists}) .$$

However, for practical applications of Taylor's approximation, as well as for many theoretical applications, it is important to have specific estimates for the size of $r_n(x)$. For example, one might expect that for fixed x near x_0, $r_n(x) \to 0$ as $n \to \infty$, i.e. that the Taylor polynomials for f around x_0 converge to f pointwise near x_0. This turns out to frequently be true, but not always, and has important consequences when it holds. And it goes without saying that when the Taylor polynomials are used to give numerical approximations to function values, it is crucial to have numerical estimates or bounds for the error.

There are many formulas known for $r_n(x)$. The first and still most widely used was found by Joseph Lagrange nearly 100 years after Taylor's work (published 1797). Lagrange's formula and another useful formula due to Cauchy can be derived by the same procedure using the Mean Value Theorem.

Good formulas for the remainder function can only be obtained when f is $(n+1)$-times differentiable in a neighborhood of x_0. Let us assume this, and fix x in such a neighborhood. For specificity we will assume that $x > x_0$, but an essentially identical argument yields the same formulas in the case $x < x_0$.

V.10.3.2. The key trick is to keep x fixed in the formula for r_n, and replace x_0 (which we also regard as fixed) by a variable t. Set

$$\phi(t) = f(x) - f(t) - f'(t)(x - t) - \cdots - \frac{f^{(n)}(t)}{n!}(x - t)^n - r_n(x)\left(\frac{x - t}{x - x_0}\right)^{n+1}$$

for $x_0 \le t \le x$. Then ϕ is differentiable on $[x_0, x]$, $\phi(x_0) = \phi(x) = 0$, and

$$\phi'(t) = -f'(t) - [f''(t)(x-t) - f'(t)] - \cdots - \left[\frac{f^{(n+1)}(t)}{n!}(x - t)^n - \frac{f^{(n)}(t)}{(n-1)!}(x - t)^{n-1}\right] + \frac{r_n(x)}{(x - x_0)^{n+1}}(n+1)(x-t)^n$$

$$= -\frac{f^{(n+1)}(t)}{n!}(x - t)^n + \frac{r_n(x)}{(x - x_0)^{n+1}}(n + 1)(x - t)^n .$$

By Rolle's Theorem, there is a $c \in (x_0, x)$ with $\phi'(c) = 0$. We then have

$$r_n(x) = \frac{(x - x_0)^{n+1}}{(n + 1)(x - c)^n}\frac{f^{(n+1)}(c)}{n!}(x - c)^n = \frac{f^{(n+1)}(c)}{(n + 1)!}(x - x_0)^{n+1} .$$

So we obtain:

V.10.3.3. THEOREM. [TAYLOR'S THEOREM WITH REMAINDER, LAGRANGE'S FORM] Let f be $(n+1)$-times differentiable on an interval I, $x_0 \in I$, p_n the n'th Taylor polynomial for f at x_0, and $x \in I$. Then, for some c between x_0 and x,

$$r_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}$$

or, equivalently,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}\,.$$

V.10.3.4. The case $n = 0$ is just the usual Mean Value Theorem. Thus Taylor's Theorem with the Lagrange Remainder Formula can be regarded as a higher-order generalization of the Mean Value Theorem.

V.10.3.5. We can obtain another formula for the remainder by a more straightforward variation of the argument: set

$$\psi(t) = f(x) - f(t) - f'(t)(x - t) - \cdots - \frac{f^{(n)}(t)}{n!}(x - t)^n$$

for $x_0 \le t \le x$. Then ψ is differentiable on $[x_0, x]$, $\psi(x_0) = r_n(x)$, $\psi(x) = 0$, and

$$\psi'(t) = -f'(t) - [f''(t)(x - t) - f'(t)] - \cdots - [\frac{f^{(n+1)}(t)}{n!}(x - t)^n - \frac{f^{(n)}(t)}{(n-1)!}(x - t)^{n-1}]$$

$$= -\frac{f^{(n+1)}(t)}{n!}(x - t)^n\ .$$

We apply the Mean Value Theorem to ψ to obtain a formula for $r_n(x)$:

V.10.3.6. THEOREM. Let f be $(n+1)$-times differentiable on an interval I, $x_0 \in I$, p_n the n'th Taylor polynomial for f at x_0, and $x \in I$. Then, for some c between x_0 and x,

$$r_n(x) = f(x) - p_n(x) = \phi(x_0) - \phi(x) = \phi'(c)(x_0 - x) = \frac{f^{(n+1)}(c)}{n!}(x - c)^n(x - x_0)\,.$$

This formula for the remainder is usually written in a different form. Since c is between x_0 and x, $c = x_0 + \theta(x - x_0)$ for some θ, $0 < \theta < 1$. Then $x - c = x - x_0 - \theta(x - x_0) = (1 - \theta)(x - x_0)$, so

V.10.3.7. COROLLARY. [TAYLOR'S THEOREM WITH REMAINDER, CAUCHY'S FORM] Let f be $(n+1)$-times differentiable on an interval I, $x_0 \in I$, p_n the n'th Taylor polynomial for f at x_0, and $x \in I$. Then, for some θ between 0 and 1,

$$r_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(x_0 + \theta(x - x_0))}{n!}(1 - \theta)^n(x - x_0)^{n+1}\,.$$

V.10.3.8. Note that the c is not necessarily unique, and depends on n, as well as on f, x, and x_0, and on which formula (Cauchy or Lagrange) is used. There is usually no reasonable way to explicitly find a c that works in a given example. However, as we will see, mere existence of such a c often gives good error estimates and crucial information about the degree of approximation given by the Taylor polynomials.

V.10.3.9. See V.11.6.7. for a generalization of V.10.3.3. to the case where f is approximated at more than one point.

V.10.4. The Integral Remainder Formula

V.10.4.1. There is another formula (first published by S. Lacroix around 1800) for r_n given by an integral. Let us fix f, x_0 and x. Let J be the closed interval from x_0 to x (x may be greater than or less than x_0). If f is differentiable on J and f' is Riemann integrable on J, then

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)\,dt\,.$$

If f'' is defined and Riemann integrable on J, do the integral by parts () the "wrong way" by letting

$$F(t) = f'(t),\ \ G(t) = -(x-t)$$

to obtain

$$f(x) = f(x_0) + [-f'(t)(x-t)]|_{t=x_0}^{t=x} - \int_{x_0}^{x} f''(t)[-(x-t)]\,dt$$

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \int_{x_0}^{x} f''(t)(x-t)\,dt.$$

If $f'''(t)$ exists and is Riemann integrable on J, do this integral again by parts using

$$F(t) = f''(t),\ \ G(t) = -\frac{1}{2}(x-t)^2$$

and continue in the same fashion n steps if f is $(n+1)$-times differentiable on J and $f^{(n+1)}$ is Riemann integrable on J. The formula for $f(x)$ thus obtained consists of the n'th Taylor polynomial p_n for f at x_0 plus an integral:

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n\,dt$$

so we obtain:

V.10.4.2. Theorem. [Taylor's Theorem With Remainder, Integral Form] Let f be $(n+1)$-times differentiable on an interval I and $x, x_0 \in I$. Suppose $f^{(n+1)}$ is Riemann integrable on the interval between x_0 and x. If p_n is the n'th Taylor polynomial of f at x_0, and $r_n = f - p_n$, then

$$r_n(x) = \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n\,dt$$

V.10.4.3. Using more advanced integration (XVIII.4.2.1.), the hypothesis that $f^{(n+1)}$ be Riemann integrable can be eliminated, and we alternately need only require that $f^{(n)}$ be absolutely continuous on the interval between x_0 and x (XIX.5.2.1., XIX.5.2.5.) so $f^{(n+1)}$ is not even necessarily defined everywhere (under this hypothesis it will be defined almost everywhere).

V.10.5. The Morse Lemma

The results of this section do not make essential use of Taylor's Theorem, but are closely related in spirit.

V.10.5.1. If f is $(n-1)$-times differentiable on an interval I around x_0 and n-times differentiable at x_0, and $r_{n-1} = f - p_{n-1}$ is the remainder function, then

$$\lim_{x \to x_0} \frac{r_{n-1}(x)}{(x-x_0)^n} = \frac{f^{(n)}(x_0)}{n!}$$

so if we set

$$g_n(x) = \frac{r_{n-1}(x)}{(x-x_0)^n} \ (x \neq x_0), \quad g_n(x_0) = \frac{f^{(n)}(x_0)}{n!}$$

for $x \in I$, then g_n is continuous on I and $(n-1)$-times differentiable on $I \setminus \{x_0\}$, and

$$f(x) = p_{n-1}(x) + (x-x_0)^n g_n(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n-1)}(x_0)}{(n-1)!}(x-x_0)^{n-1} + (x-x_0)^n g_n(x)$$

for $x \in I$.

V.10.5.2. Proposition. If f is $\mathcal{C}^n$ on I, then for $1 \leq m \leq n$ g_m is $\mathcal{C}^n$ on $I \setminus \{x_0\}$ and continuously $(n-m)$-times differentiable at x_0.

Proof: Since $r_{m-1} = f - p_{m-1}$, hence also g_m, is n-times differentiable on $I \setminus \{x_0\}$, differentiability at x_0 is the only issue. First consider the case $m = 1$. Fix $x \in I$, and set $\phi(t) = f(tx + (1-t)x_0)$ for $0 \leq t \leq 1$. Then $\phi'(t) = f'(tx + (1-t)x_0)(x - x_0)$, so

$$(x-x_0)\int_0^1 f'(tx + (1-t)x_0)\,dt = \int_0^1 \phi'(t)\,dt = \phi(1) - \phi(0) = f(x) - f(x_0) = r_0(x) = (x-x_0)g_1(x)$$

by the Fundamental Theorem of Calculus, so

$$g_1(x) = \int_0^1 f'(tx + (1-t)x_0)\,dt\ .$$

Since f' is continuously $(n-1)$-times differentiable at x_0, by () g_1 is also continuously $(n-1)$-times differentiable at x_0.

Now let $m = 2$ (if $n \geq 2$). We have, for $x \in I$,

$$f(x) = f(x_0) + (x-x_0)g_1(x) = f(x_0) + f'(x_0)(x-x_0) + (x-x_0)^2 g_2(x) = f(x_0) + (x-x_0)[f'(x_0) + (x-x_0)g_2(x)]$$

so

$$g_1(x) = f'(x_0) + (x-x_0)g_2(x) = g_1(x_0) + (x-x_0)g_2(x)$$

for $x \in I$. Thus, by the first part of the proof applied to g_1 in place of f, we obtain that g_2 is continuously $(n-2)$-times differentiable at x_0. Continue inductively. 👍

V.10.5.3. COROLLARY. If f is $\mathcal{C}^\infty$ on I, then g_n is $\mathcal{C}^\infty$ on I for all n.

V.10.5.4. The example $f(x) = x^n|x|$, with $x_0 = 0$, shows that g_m is not $(n-m+1)$-times differentiable at x_0 in general, even if f is $\mathcal{C}^n$ on I and $\mathcal{C}^\infty$ on $I \setminus \{x_0\}$. Note that the conclusion of V.10.5.2. implies that g_m is $\mathcal{C}^{n-m}$ on I.

The hypotheses of V.10.5.2. can be relaxed slightly to only require f to be n-times differentiable at x_0 and $(n-1)$-times differentiable on I, with $f^{(n-1)}$ absolutely continuous on I, but then the conclusion is only that g_m is $(n-1)$-times differentiable on $I \setminus \{x_0\}$ (and n-times differentiable a.e.) and $(n-m)$-times differentiable at x_0.

If f is real analytic, then so are the g_n. The result also holds if f is complex analytic. In these cases, the proofs are simpler, cf. () and ().

The One-Dimensional Morse Lemma

V.10.5.5. If f is more than m-times differentiable at x_0, $f(x_0) = c$, and $f'(x_0) = \cdots = f^{(m-1)}(x_0) = 0$ but $f^{(m)}(x_0) \neq 0$, then $f(x)$ is approximately $p_m(x) = c + d(x-x_0)^m$ for x near x_0, where d is a constant ($d = \frac{f^{(m)}(x_0)}{m!}$). But we can say more: up to a change of coordinates, f is *exactly* of this form, and we can even take $d = \pm 1$.

V.10.5.6. THEOREM. [GENERALIZED ONE-DIMENSIONAL MORSE LEMMA] Let $m, n \in \mathbb{N}$, $m < n$, and let f be $\mathcal{C}^n$ on an interval I around x_0, with $f(x_0) = c$ and $f'(x_0) = \cdots = f^{(m-1)}(x_0) = 0$ but $f^{(m)}(x_0) \neq 0$. Then there is an orientation-preserving diffeomorphism ϕ from an open interval J around 0 onto an open interval $\tilde{I}$ around x_0, with $\phi(0) = x_0$, such that

$$[f \circ \phi](t) = c \pm t^m$$

for all $t \in J$, where the sign is the sign of $f^{(m)}(x_0)$. The function ϕ is $\mathcal{C}^{n-m}$ on J.

PROOF: By V.10.5.2., for $x \in I$ f can be written $f(x) = c + (x-x_0)^m g_m(x)$ with $g_m(x_0) = d \neq 0$. Thus there is an open interval $K \subseteq I$ around x_0 on which g_m is nonzero. For $x \in \tilde{I}$, set

$$\psi(x) = (x - x_0)h(x)$$

where

$$h(x) = |g_m(x)|^{1/m} \ .$$

Then h, hence ψ, is a $\mathcal{C}^{n-m}$ function on K, $\psi(x_0) = 0$, and

$$\psi'(x) = h(x) + (x - x_0)h'(x) \ .$$

Since h' is bounded near x_0 and $h(x_0) > 0$, there is an open interval $\tilde{I}$ around x_0 on which $\psi' > 0$. Thus ψ maps $\tilde{I}$ diffeomorphically onto an open interval J containing 0. Let $\phi : J \to \tilde{I}$ be the inverse function. Then ϕ is also $\mathcal{C}^{n-m}$ by the Inverse Function Theorem (). If $t = \psi(x) \in J$, then

$$[f \circ \phi](t) = f(x) = c + (x-x_0)^m g_m(x) = c \pm [(x-x_0)h(x)]^m = c \pm \psi(x)^m = c \pm t^m \ .$$



The classical one-dimensional Morse Lemma (due to MARSTON MORSE) is the case $m = 2$, and is usually written just for the case f is $\mathcal{C}^\infty$:

V.10.5.7. Corollary. [One-Dimensional Morse Lemma] Let f be $\mathcal{C}^\infty$ on an interval I around x_0, with $f(x_0) = c$ and $f'(x_0) = 0$ but $f''(x_0) \neq 0$. Then there is an orientation-preserving $\mathcal{C}^\infty$-diffeomorphism ϕ from an open interval J around 0 onto an open interval $\tilde{I}$ around x_0, with $\phi(0) = x_0$, such that

$$[f \circ \phi](t) = c \pm t^2$$

for all $t \in J$, where the sign is the sign of $f''(x_0)$.

V.10.5.8. If we allow ϕ to be orientation-reversing, if m is odd we can always take the + sign in the conclusion. But if m is even (in particular, if $m = 2$), the sign does not depend on the orientation of ϕ.

See () for the n-dimensional Morse Lemma. There are also real and complex analytic versions; see V.16.4.4. and XII.6.1.2..

V.10.6. Applications

V.10.7. Exercises

V.10.7.1. In V.10.1.3., the hypothesis that $g^{(n)}(x_0)$ exists is necessary if $n > 1$:

(a) Let $h_n(x) = \begin{cases} x^{n+1} & \text{if} \quad x \in \mathbb{Q} \\ 0 & \text{if} \quad x \notin \mathbb{Q} \end{cases}$. Then $\lim_{x\to 0} \frac{h_n(x)}{x^n} = 0$, but h_n is not even continuous except at 0, so $h_n'', \ldots, h_n^{(n)}$ do not exist at 0 (or anywhere else).

(b) The hypothesis that $g^{(n)}(x_0)$ exists is automatic if $n = 1$: if $g(x_0) = 0$ and $\lim_{x\to 0} \frac{g(x)}{x-x_0} = 0$, then $g'(x_0)$ exists and equals 0.

V.10.7.2. Give an alternate direct proof of Theorem V.10.1.2. not using l'Hôpital's Rule.

V.10.7.3. (a) Mimic the proof of V.10.1.2. to prove the following generalization of V.10.1.3.:

Theorem. [?] Let g be n-times differentiable in an open interval around x_0, with $g^{(n)}$ Hölder continuous () at x_0 with exponent α $(0 < \alpha \leq 1)$ and constant 0, i.e.

$$\lim_{x\to x_0} \frac{g^{(n)}(x) - g^{(n)}(x_0)}{(x - x_0)^\alpha} = 0 \ .$$

Then

$$\lim_{x\to x_0} \frac{g(x)}{(x - x_0)^{n+\alpha}} = 0$$

if and only if $g(x_0) = g'(x_0) = \cdots = g^{(n)}(x_0) = 0$.

(b) Derive V.10.1.3. as a corollary of the case $\alpha = 1$ of (a).

V.10.7.4. Write out the details of the arguments for the Lagrange and Cauchy remainder formulas in the case $x < x_0$.

V.10.7.5. The proof of V.10.3.3. is simple, but requires a nonobvious guess of a function to apply Rolle's theorem to. Phrase the proof in an alternate way not requiring a guess:

(a) Using the hypotheses and the notation of the theorem, show that there is an M such that

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + M(x - x_0)^{n+1} \ .$$

(b) For $x_0 \leq t \leq x$ set

$$\phi(t) = f(x) - f(t) - f'(t)(x - t) - \frac{f''(t)}{2}(x - t)^2 - \cdots - \frac{f^{(n)}(t)}{n!}(x - t)^n - M(x - t)^{n+1} \ .$$

(c) Apply Rolle's Theorem to ϕ on $[x_0, x]$ (check that the hypotheses apply) and deduce Lagrange's form of the remainder. (The M will have to be calculated in terms of $r_n(x)$.)

This version of the argument appeared in a Belgian high-school textbook (*Wiskundig Leerpakket*); cf. https://gowers.wordpress.com/2014/02/11/taylors-theorem-with-the-lagrange-form-of-the-remainder/.

V.10.7.6. (a) Prove the following:

THEOREM. Let f be $(n+1)$-times differentiable on an interval I, $x_0 \in I$, p_n the n'th Taylor polynomial for f at x_0, and $x \in I$. Let g be a differentiable function on I such that $g'(t) \neq 0$ for all t in the open interval between x_0 and x. Then there is a c between x_0 and x for which

$$r_n(x) = -\frac{g(x_0) - g(x)}{g'(c)} \frac{f^{(n+1)}(c)}{n!}(x - c)^n$$

[Set

$$\omega(t) = \frac{\psi(t)}{g(t)}$$

where ψ is as in V.10.3.5., and apply the Cauchy Mean Value Theorem.]

(b) Obtain the Lagrange form of the remainder by taking $g(t) = (x - t)^{n+1}$.

V.10.7.7. Prove the following remainder formula:

THEOREM. [TAYLOR'S THEOREM WITH REMAINDER, SCHLÖMILCH'S FORM] Let f be $(n+1)$-times differentiable on an interval I, $x_0 \in I$, p_n the n'th Taylor polynomial for f at x_0, and $x \in I$. Fix $m \in \mathbb{N}$, $1 \leq m \leq n+1$. Then, for some θ between 0 and 1,

$$r_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(x_0 + \theta(x - x_0))}{n!m}(1 - \theta)^{n+1-m}(x - x_0)^{n+1} \ .$$

The case $m = 1$ is Cauchy's form, and $m = n + 1$ is Lagrange's form. The θ depends on f, x_0, x, n, and m.

V.10.7.8. Let $f(x) = \begin{cases} e^{-1/x^2} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$.

(a) Prove by complete induction that $\lim_{x\to 0} \frac{f(x)}{x^n} = 0$ for all $n \in \mathbb{N} \cup \{0\}$. [Hint: Use l'Hôpital's Rule, but not in the most straightforward way.]

(b) Prove by induction that $f^{(n)}(0)$ exists and equals 0 for all n. Thus f is C^∞ on all of $\mathbb{R}$. [You will need to use part (a) and the definition of the derivative.]

(c) Find the n'th Taylor polynomial p_n for f around 0, and show that $p_n(x)$ does not converge to $f(x)$ for any fixed $x \neq 0$. In fact, although $f(x) > 0$ for all $x \neq 0$, the graph near 0 is so "flat" that the constant function 0 approximates f near zero better than any other polynomial of any degree. See V.17.2.2..

V.10.7.9. Let $f(x) = e^x$, and let p_n be the n'th Taylor polynomial for f around 0. Since $f^{(k)}(x) = e^x$ for all k, $p_n(x) = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!}$.

(a) Show that for any fixed $x \in \mathbb{R}$, $\lim_{n\to\infty} \frac{x^n}{n!} = 0$. [Consider the ratios of successive terms, and use ().]

(b) What is wrong with the following argument? Fix $x \in \mathbb{R}$. Then by Lagrange's remainder formula there is a c between 0 and x such that

$$f(x) - p_n(x) = \frac{e^c}{(n+1)!} x^{n+1} .$$

Since $\lim_{n\to\infty} e^c \frac{x^{n+1}}{(n+1)!} = 0$, $p_n(x) \to f(x)$ as $n \to \infty$.

(c) Modify the argument of (b) to give a valid proof that $p_n(x) \to f(x)$ for any $x \in \mathbb{R}$.

V.10.7.10. Give an alternate proof of the integral form of the remainder for Taylor's Theorem not requiring integration by parts, by computing the iterated integral

$$\int_{x_0}^{x} \int_{x_0}^{x_n} \cdots \int_{x_0}^{x_1} f^{(n+1)}(t)\, dt\, dx_1 \cdots\ dx_n$$

where $x_1, \ldots, x_n$ are variables ranging over the interval between x_0 and x, by successive applications of the Fundamental Theorem of Calculus. Note that the first integral is not necessarily a Riemann integral unless $f^{(n+1)}$ is well-behaved between x_0 and x, but can be interpreted as a Lebesgue integral or an H-K integral in more general cases. The rest of the integrals are always Riemann integrals.

V.10.7.11. Let f be $\mathcal{C}^n$ $(0 \leq n < \infty)$ on an open interval I, and $x_0 \in I$. Suppose there is a sequence (x_k) in $I \setminus \{x_0\}$ with $x_k \to x_0$ and $f(x_k) = 0$ for all k. Show that

$$f(x_0) = f'(x_0) = \cdots f^{(n)}(x_0) = 0$$

and hence

$$\lim_{x\to x_0} \frac{f(x)}{(x - x_0)^n} = 0 .$$

[Apply Rolle's Theorem to the intervals between x_k and x_{k+1} for each k, and iterate.] Show by example that

$$\lim_{x\to x_0} \frac{f(x)}{(x - x_0)^{n+1}} = 0$$

does not necessarily hold.

V.10.7.12. [Extended Second Derivative Test] Let f be defined in a neighborhood of x_0, and $n \in \mathbb{N}$, $n \geq 2$. Suppose $f^{(n)}(x_0)$ exists, and that $f'(x_0) = \cdots = f^{(n-1)}(x_0) = 0$ and $f^{(n)}(x_0) \neq 0$. Using V.10.5.6. (or just V.10.2.4.), show that

(a) If n is even, then f has a strict local extremum at x_0, a local minimum if $f^{(n)}(x_0) > 0$ and a local maximum if $f^{(n)}(x_0) < 0$.

(b) If n is odd, then f is increasing or decreasing at x_0 in the sense of V.3.6.7., increasing if $f^{(n)}(x_0) > 0$ and decreasing if $f^{(n)}(x_0) < 0$, and f has a point of inflection at x_0.

V.10.7.13. Let f be continuously $(n+1)$-times differentiable on an interval I, and $x_0, x \in I$. Derive the Lagrange form for the remainder V.10.3.3. from the integral form V.10.4.2. using the Mean Value Theorem for Integrals (IX.2.6.10.): There is a c between x_0 and x such that

$$\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n \, dt = f^{(n+1)}(c) \int_{x_0}^{x} (x-t)^n \, dt \ .$$

(Note that this is slightly less general than V.10.3.3..)

V.10.7.14. Prove the statements in V.10.4.3. (there are some nontrivial technicalities in addition to the results referenced there).

V.11. Interpolation and Extrapolation

In this section, we discuss some topics normally treated only in numerical analysis texts but which are very helpful and relevant to understanding Taylor's Theorem (our exposition is largely based on the nice treatments in [**?**] and http://www.maths.lancs.ac.uk/~jameson/interpol.pdf). No actual calculus is needed for the basic definitions and formulas, but the analysis is closely related in spirit to differential approximation and Taylor's theorem (indeed, TAYLOR obtained his formula as a limiting case of the Gregory-Newton interpolation formula), and results from calculus are needed for error estimates.

The basic problem is as follows. Suppose we have a function f defined on an interval I, which we presume to be suitably nicely behaved, e.g. differentiable to some high order. We know values of f at some points $x_0, \ldots, x_n$ of I (typically forming a partition of I, but not necessarily), and we want to estimate $f(x)$ for a general $x \in I$. This type of approximation is called *interpolation*, especially if x is in the smallest interval containing $x_0, \ldots, x_n$; if x is not in this interval, the approximation is *extrapolation*. (Extrapolation is generally much more difficult and problematic to do with any accuracy.)

One promising way to proceed is to approximate f by the unique polynomial p of degree $\leq n$ whose values at $x_0, \ldots, x_n$ are the same as f, called the *interpolating polynomial* for f with respect to $\{x_0, \ldots, x_n\}$.

V.11.1. The Interpolating Polynomial

We first consider the problem of the existence and uniqueness of the interpolating polynomial and how to find it.

V.11.1.1. The most direct approach is a simple exercise in linear algebra. Given distinct $x_0, \ldots, x_n$ in $\mathbb{R}$ and real numbers $b_0, \ldots, b_n$ (not necessarily distinct), we want a polynomial

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$

with real coefficients, of degree $\leq n$ (it could have degree less than n since a_n could be zero), such that $p(x_k) = b_k$ for $0 \leq k \leq n$. Such a polynomial, if it exists, must be unique: if $q(x)$ is another such polynomial, then $p(x) - q(x)$ is a polynomial of degree $\leq n$ with at least $n+1$ roots (at $x_0, \ldots, x_n$), hence must be identically zero.

To show the polynomial exists, we must solve the system

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$

for $\{a_0, \ldots, a_n\}$. The matrix on the left is a *Vandermonde matrix* which is invertible (the determinant is $\prod_{j<k}(x_k - x_j)$; cf. Exercise V.11.8.1.); thus the system has a (unique) solution which can be found by Gaussian elimination. (This works over any field with at least $n+1$ elements.)

V.11.1.2. This procedure for finding p works fine in principle, but is regarded as a poor method in numerical analysis, since it is computationally inefficient and the system is often ill-conditioned, especially if the x_k are close together, leading to large roundoff errors.

V.11.1.3. There is another form of the interpolating polynomial which is more computationally efficient. This form is usually attributed to LAGRANGE and is often called the *Lagrange form* of the interpolation polynomial although it was originally due to E. WARING in 1779.

Fix distinct real numbers $x_0, \dots, x_n$. The polynomial

$$p_k(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}$$

of degree n has the property that $p_k(x_j) = \delta_{kj}$ () for $0 \le j \le n$. Thus, if $b_0, \dots, b_n \in \mathbb{R}$, the polynomial p of degree $\le n$ with $p(x_k) = b_k$ for all k is

$$p(x) = \sum_{k=0}^{n} b_k p_k(x) \ .$$

V.11.1.4. There is an even better form of the interpolating polynomial, the Newton form (V.11.3.2.).

V.11.2. Linear Interpolation and Extrapolation

V.11.2.1. If we only know the value of f at one point x_0 of I, the best estimate we can make of $f(x)$ for $x \in I$ is

$$f(x) = f(x_0)$$

i.e. f is constant on I (this is generally a pretty crude estimate, but the best we can do with only one data point, and reasonably accurate for x near x_0 if f is continuous). We will use the notation $f[x_0]$ to denote the number $f(x_0)$; thus we have the 0'th order approximation

$$f(x) = f[x] \approx f[x_0] = f(x_0) \ .$$

Denote by p_0 the constant function with value $f(x_0)$ (we should really write p_{x_0}, but we will think of x_0 as being fixed and adding additional points $x_1, x_2, \dots$ successively).

V.11.2.2. If we know the values of f at two points x_0 and x_1 of I, we can do a little better: we can assume f is linear, with slope

$$f[x_0, x_1] := \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

(i.e. $f[x_0, x_1]$ is the difference quotient (V.3.1.8.), usually called a *divided difference* in numerical analysis.) We then have the approximation

$$f(x) \approx f(x_0) + f[x_0, x_1](x - x_0) = f[x_0] + (x - x_0)f[x_0, x_1]$$

for $x \in I$ and, of course, the exact formula

$$f(x) = f[x_0] + (x - x_0)f[x_0, x]$$

($x \neq x_0$) which is useless for approximation.

V.11.2.3. The linear function

$$p_{0,1}(x) = f[x_0] + (x - x_0)f[x_0, x_1]$$

is the unique linear function with the same values as f at x_0 and x_1.

V.11.2.4. Linear interpolation and extrapolation can be written in the equivalent form

$$f(\lambda x_0 + (1-\lambda)x_1) \approx \lambda f(x_0) + (1-\lambda)f(x_1)$$

(interpolation if $0 < \lambda < 1$, extrapolation if $\lambda < 0$ or $\lambda > 1$).

V.11.3. Higher-Order Interpolation and Extrapolation

V.11.3.1. Now suppose we know the values of f at x_0, x_1, and x_2 (assumed distinct). We want to find the quadratic polynomial $p_{0,1,2}$ whose values are the same as f at x_0, x_1, and x_2. This can be done as in section V.11.1., but we take a different approach. The polynomial will be of the form

$$p_{0,1,2}(x) = f[x_0] + (x - x_0)f[x_0, x_1] + K(x - x_0)(x - x_1)$$

for some constant K. A little algebra gives

$$K = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)} = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} .$$

We thus define

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

so we have, for $x \in I$,

$$f(x) \approx p_{0,1,2}(x) = f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2]$$

and the exact formula

$$f(x) = f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x]$$

(the last expression is not defined if $x = x_0$ or $x = x_1$, but the last term can be taken to be 0 in these cases, which makes the overall formula valid).

V.11.3.2. Continuing inductively in the same manner, if we have $n + 1$ distinct points $x_0, \ldots, x_n \in I$, we have

$$p_{0,1,\ldots,n}(x) = f[x_0] + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2] + \cdots + (x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0, \ldots, x_n]$$

where

$$f[x_0, \ldots, x_n] = \frac{f(x_0)}{(x_0 - x_1)\cdots(x_0 - x_n)} + \cdots + \frac{f(x_k)}{(x_k - x_0)(x_k - x_1)\cdots(x_k - x_{k-1})(x_k - x_{k+1})\cdots(x_k - x_n)} + \cdots + \frac{f(x_n)}{(x_n - x_0)\cdots(x_n - x_{n-1})}$$

are called *higher-order divided differences*. There are many other formulas for these divided differences; see below for some of them.

Thus, for $x \in I$, we have the n'th order approximation

$$f(x) \approx f[x_0] + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2] + \cdots + (x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0, \ldots, x_n]$$

and the exact formula ($x \neq x_0, \ldots, x_{n-1}$)

$$f(x) = f[x_0] + (x-x_0)f[x_0, x_1] + (x-x_0)(x-x_1)f[x_0, x_1, x_2] + \cdots + (x-x_0)(x-x_1)\cdots(x-x_{n-1})f[x_0, \ldots, x_{n-1}, x] .$$

V.11.3.3. Divided differences and the form of the interpolating polynomial in V.11.3.2. are attributed to NEWTON and called the *Newton form*, although they were first obtained by JAMES GREGORY around 1670, at least in the case of evenly-spaced data points.

V.11.3.4. The Newton form of the interpolating polynomial is regarded as the most generally useful and efficient form for numerical analysis purposes. Not only is it computationally efficient, but it also has the nice property that the coefficients do not have to all be recomputed if an additional data point is added.

Properties of Higher-Order Divided Differences

V.11.3.5. In the divided difference $f[x_0, \ldots, x_n]$, the numbers $x_0, \ldots, x_n$ are typically taken in increasing order. However, this is not necessary; in fact, it is obvious from the definition V.11.3.2. that the order in which the x_k are taken is irrelevant: the numerical value is invariant under any permutation of the x_k.

This can be seen in another way. The number $f[x_0, \ldots, x_n]$ is the coefficient of x^n in the interpolating polynomial corresponding to $\{x_0, \ldots, x_n\}$. Since this polynomial obviously depends only on the set $\{x_0, \ldots, x_n\}$ and not on the order of the numbers, so does $f[x_0, \ldots, x_n]$.

V.11.3.6. There is also a recursive formula

$$f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}$$

which justifies the name "higher-order divided difference" (Exercise V.11.8.2.). There are many variations obtained by combining this formula with permutations of the x_k (V.11.3.5.).

V.11.4. The Mean Value Theorem for Divided Differences

We now bring some calculus into the picture.

V.11.4.1. The n'th order divided difference $f[x_0, \ldots, x_n]$ can be regarded as a function $f[\cdot, \ldots, \cdot]$ defined on the subset I_d^{n+1} of I^{n+1} consisting of points with all coordinates distinct. The set I_d^{n+1} is a dense open set in I^{n+1}. If f is continuous on I, this n'th order divided difference is clearly continuous on I_d^{n+1}. We will eventually extend it to all of I^{n+1} if f is sufficiently differentiable on I (V.11.5.7.).

The most important calculus result about divided differences is the following version of the MVT:

V.11.4.2. THEOREM. Let f be continuous on $[a, b]$ and n-times differentiable everywhere on (a, b). Let $x_0, \ldots, x_n$ be $n + 1$ distinct points in $[a, b]$. Then there is at least one $c \in (a, b)$ with

$$f[x_0, \ldots, x_n] = \frac{f^{(n)}(c)}{n!} \ .$$

PROOF: We may assume $(x_0, \ldots, x_n)$ is in increasing order. The function

$$g(x) = f(x) - p_{0,1,\ldots,n}(x)$$

has at least $n+1$ distinct zeroes in $[a,b]$ (at $x_0,\dots,x_n$). Applying Rolle's Theorem repeatedly to the intervals between x_{k-1} and x_k, $g'(x)$ has at least n zeroes in (a,b), $g''(x)$ has at least $n-1$ zeroes, $\dots$, $g^{(n)}(x)$ has at least one zero c. But $p_{0,\dots,n}$ is a polynomial of degree $\le n$ whose degree n term is

$$f[x_0,\dots,x_n]x^n$$

so its n'th derivative is the constant function $n!f[x_0,\dots,x_n]$, and thus

$$0 = g^{(n)}(c) = f^{(n)}(c) - p^{(n)}_{0,\dots,n}(c) = f^{(n)}(c) - n!f[x_0,\dots,x_n] \ .$$



V.11.4.3. Note that the case $n=1$ of this theorem is the usual MVT, so this theorem is a generalization of the MVT. But the MVT, or at least the special case of Rolle's Theorem, is used in the proof.

V.11.4.4. We can use this result to calculate the error in using $p_{0,\dots,n}(x)$ to approximate $f(x)$ for $x \in I$. Set

$$E_n(x) = f(x) - p_{0,\dots,n}(x)$$

for $x \in I$. If x is not one of $x_0,\dots,x_n$, then we have from V.11.3.2. that

$$E_n(x) = (x-x_0)(x-x_1)\cdots(x-x_n)f[x_0,\dots,x_n,x]$$

and combining this with V.11.4.2., we obtain:

V.11.4.5. Corollary. Let f be $(n+1)$-times differentiable on I, and $x_0,\dots,x_n$ distinct points of I. Then, for any $x \in I$, there is a c in the interior of the smallest subinterval of I containing $x_0,\dots,x_n,x$ such that

$$E_n(x) = (x-x_0)(x-x_1)\cdots(x-x_n)\frac{f^{(n+1)}(c)}{(n+1)!} \ .$$

Proof: If x is not one of the x_k, the result follows from V.11.4.2., and if x is one of the x_k both sides are zero no matter what c is chosen.



V.11.5. Repeated Points

V.11.5.1. In defining $f[x_0,x_1]$, the x_0 and x_1 must be distinct. But if (and only if) f is differentiable at x_0, we have, for fixed x_0,

$$\lim_{x_1\to x_0} f[x_0,x_1] = \lim_{x_1\to x_0}\frac{f(x_1)-f(x_0)}{x_1-x_0} = f'(x_0)$$

so we may define

$$f[x_0,x_0] = f'(x_0)$$

if f is differentiable at x_0. This definition is especially suitable if f is strictly differentiable at x_0 (), in which case $f[\cdot,\cdot]$ is continuous at (x_0,x_0). In particular, if f is $\mathcal{C}^1$ on I, then $f[\cdot,\cdot]$ is continuous on I^2.

V.11.5.2. We want to extend this to higher-order divided differences. We first consider the case where all the coordinates are the same. Fix n, and for $x_0 \in I$ write $\boldsymbol{x}_0 = (x_0, \ldots, x_0) \in I^{n+1}$. If $\boldsymbol{x} = (y_0, \ldots, y_n) \in I_d^{n+1}$, write $f[\boldsymbol{x}] = f[y_0, \ldots, y_n]$.

V.11.5.3. PROPOSITION. Let f be n-times differentiable on I, and $x_0 \in I$. If $f^{(n)}$ is continuous at x_0, then

$$\lim_{\substack{\boldsymbol{x} \to \boldsymbol{x}_0 \\ \boldsymbol{x} \in I_d^{n+1}}} f[\boldsymbol{x}] = \frac{f^{(n)}(x_0)}{n!} \ .$$

PROOF: Let $\epsilon > 0$. Then there is a $\delta > 0$ such that, for all $y \in I$, $|y - x_0| < \delta$, we have

$$\left| \frac{f^{(n)}(y)}{n!} - \frac{f^{(n)}(x_0)}{n!} \right| < \epsilon \ .$$

Let $J = I \cap (x_0 - \delta, x_0 + \delta)$. If $\boldsymbol{x} = (y_0, \ldots, y_n) \in J_d^{n+1}$, then by V.11.4.2. there is a $c_{\boldsymbol{x}} \in J$ with

$$f[\boldsymbol{x}] = \frac{f^{(n)}(c_{\boldsymbol{x}})}{n!}$$

so

$$\left| f[\boldsymbol{x}] - \frac{f^{(n)}(x_0)}{n!} \right| < \epsilon \ .$$



V.11.5.4. If f is $\mathcal{C}^n$ on I, we can thus extend the n'th order divided difference to points with all coordinates the same by setting

$$f[x_0, \ldots, x_0] = \frac{f^{(n)}(x_0)}{n!}$$

for $x_0 \in I$. The extended function is continuous on the subset of I^{n+1} consisting of points whose coordinates are either all the same or all different.

V.11.5.5. We can extend higher-order divided differences to points where some, but not all, coordinates coincide. Rather than trying to state things in maximum generality (cf. V.11.5.13.), we will just assume f is $\mathcal{C}^n$ on I. The most efficient approach is to prove an integral formula for the higher-order divided differences which is of general interest:

V.11.5.6. THEOREM. Let f be $\mathcal{C}^n$ on an interval I, and $x_0, \ldots, x_n$ distinct points of I. Then

$$f[x_0, \ldots, x_n] = \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-1}} f^{(n)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1}))\, dt_n \cdots dt_2\, dt_1 \ .$$

PROOF: By induction on n. The case $n = 1$ is essentially the Fundamental Theorem of Calculus in slight disguise: to evaluate the integral

$$\int_0^1 f'(x_0 + t(x_1 - x_0))\,dt$$

for f $\mathcal{C}^1$, make the substitution

$$u = x_0 + t(x_1 - x_0)$$
$$du = (x_1 - x_0)\,dt$$

to obtain

$$\int_0^1 f'(x_0 + t(x_1 - x_0))\,dt = \frac{1}{x_1 - x_0}\int_{x_0}^{x_1} f'(u)\,du = \frac{f(x_1) - f(x_0)}{x_1 - x_0}\ .$$

Thus $f[x_0, x_1]$ is the average value of f' over the interval $[x_0, x_1]$ (or $[x_1, x_0]$ if $x_1 < x_0$).

Now suppose the formula holds for $n - 1$. Compute the first integral

$$\int_0^{t_{n-1}} f^{(n)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1}))\,dt_n$$

making the substitution

$$u = x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1})$$
$$du = (x_n - x_{n-1})\,dt_n$$

so the first integral evaluates to

$$\frac{1}{x_n - x_{n-1}}[f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1}))]\Big|_0^{t_{n-1}}$$

$$= \frac{1}{x_n - x_{n-1}}[f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}) + t_{n-1}(x_n - x_{n-1}))$$
$$-f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]$$

$$= \frac{1}{x_n - x_{n-1}}[f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_n - x_{n-2}))$$
$$-f^{(n-1)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))]$$

so we have

$$\int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-1}} f^{(n)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_n(x_n - x_{n-1}))\,dt_n \cdots dt_2\,dt_1$$

$$= \frac{1}{x_n - x_{n-1}}\left[\int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-2}} f^{(n)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_n - x_{n-2}))\,dt_{n-1} \cdots dt_2\,dt_1\right.$$
$$\left. - \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{n-2}} f^{(n)}(x_0 + t_1(x_1 - x_0) + t_2(x_2 - x_1) + \cdots + t_{n-1}(x_{n-1} - x_{n-2}))\,dt_{n-1} \cdots dt_2\,dt_1\right]$$

$$= \frac{1}{x_n - x_{n-1}}(f[x_0, \ldots, x_{n-2}, x_n] - f[x_0, \ldots, x_{n-2}, x_{n-1}]) = f[x_0, \ldots, x_n]$$

by the inductive hypothesis and V.11.3.6..

V.11.5.7. The formula in V.11.5.6. makes perfect sense even if $x_0, \ldots, x_n$ are not distinct, and clearly gives a continuous function of $(x_0, \ldots, x_n)$ on I^{n+1}. We denote this extended function also by $f[x_0, \ldots, x_n]$. The extended function agrees with the extension given in V.11.5.4. in the case all the coordinates are equal. Thus we obtain

V.11.5.8. COROLLARY. If f is $\mathcal{C}^n$ on an interval I, then $f[\cdot, \ldots, \cdot]$ ($n+1$ arguments), extended as above to points with repeated coordinates, is continuous on I^{n+1}.

V.11.5.9. If f is $\mathcal{C}^n$ on $[a, b]$, then the conclusion of Theorem V.11.4.2. holds also (with $c \in [a, b]$) if not all the x_k are distinct, as can be seen by a straightforward limiting argument using V.11.5.8. and the Bolzano-Weierstrass Theorem (). In particular, if f is $\mathcal{C}^n$, then for any x_0 and any $[a, b]$ containing x_0 there is a $c \in [a, b]$ for which $f[x_0, \ldots, x_0] = \frac{f^{(n)}(c)}{n!}$. Taking the limit as $a, b \to x_0$, we obtain an alternate proof of V.11.5.3..

To extend $f[x_0, \ldots, x_n]$ to points with some repetitions in somewhat greater generality, we use the following result, which is proved inductively using V.11.5.8. and the recursion formula for divided differences with distinct entries:

V.11.5.10. THEOREM. Let f be a function on an interval I, $J_0, \ldots, J_r$ disjoint subintervals of I, and $m_0, \ldots, m_r \in \mathbb{N}$. Suppose f is $\mathcal{C}^{m_k - 1}$ on J_k for each k. Then

$$f[x_{(0,0)}, \ldots, x_{(0,m_0-1)}, x_{(1,0)}, \ldots, x_{(1,m_1-1)}, \ldots, x_{(r,m_r-1)}]$$

extends to a continuous function on $J_0^{m_0} \times \cdots \times J_r^{m_r}$.

V.11.5.11. We can obtain an explicit formula for $f[x_0, \ldots, x_n]$ when the points are not all distinct. Since the function values are independent of the order of the coordinates on a dense set, by continuity this holds everywhere. Thus to evaluate the function we may put the entries in increasing order, with the repeated coordinates grouped together. We write $x_0, \ldots, x_r$ for the distinct values, and suppose x_k occurs m_k times, so $m_0 + \cdots + m_r = n + 1$.

V.11.5.12. THEOREM. Suppose f is $\mathcal{C}^n$ on I. Then $f[x_0, \ldots, x_r]$ is $\mathcal{C}^{n-r}$ on I_d^{r+1}, and

$$f[x_0, \ldots, x_0, x_1, \ldots, x_1, \ldots, x_r] = \frac{1}{(m_0 - 1)!(m_1 - 1)! \cdots (m_r - 1)!} \frac{\partial^{n-r}}{\partial x_0^{m_0-1} \partial x_1^{m_1-1} \cdots \partial x_r^{m_r-1}} (f[x_0, \ldots, x_r])$$

where x_k occurs m_k times on the left and once on the right.

V.11.5.13. Actually, for this formula to hold and define $f[x_0, \ldots, x_0, x_1, \ldots, x_1, \ldots, x_r]$ continuously at $(x_0, \ldots, x_0, x_1, \ldots, x_1, \ldots, x_r)$, it suffices that f be $\mathcal{C}^{m_k - 1}$ in an interval around x_k for each k (although the $\mathcal{C}^{n-r}$ statement does not hold in general). See [**?**, p. 72-73] for an outline of the proof. Note that the formula agrees with the previous cases where $r = n$ (all $m_k = 1$) or $r = 0$ ($m_0 = n + 1$).

V.11.6. Osculating Polynomials

V.11.6.1. We want to extend the notion of interpolating polynomial to the case where the points $x_0, \dots, x_n$ are not all distinct. The first thing we must decide is what this should mean: if x_0 is repeated m times, how should the approximating polynomial p agree with f at x_0?

The approximation should be of "order m" in the sense that $f - p$ should have a zero of "order m" at x_0. The broadest sense in which this can be interpreted is to demand that

$$f(x) - p(x) = o(|x - x_0|^{m-1})$$

as $x \to x_0$ in I. By V.10.1.3., if f is $(m-1)$-times differentiable at x_0, this will happen if and only if

$$p(x_0) = f(x_0), p'(x_0) = f'(x_0), \dots, p^{(m-1)}(x_0) = f^{(m-1)}(x_0) \ .$$

So, if f is differentiable to high enough order, we need to find a polynomial with specified values and derivatives up to a certain level at each of the points. This is possible because of the following extension of V.11.1.1.:

V.11.6.2. THEOREM. Let $n \in \mathbb{N}$, $x_0, \dots, x_r$ distinct real numbers, $m_0, \dots, m_r \in \mathbb{N}$ with $m_0 + \cdots + m_r = n + 1$ (so necessarily $r \le n$), and

$$b_0^{(0)}, b_0^{(1)}, \dots, b_0^{(m_0-1)}, b_1^{(0)}, \dots, b_r^{(m_r-1)}$$

real numbers (not necessarily distinct). Then there is a unique polynomial p with real coefficients, of degree $\le n$, for which $p^{(j)}(x_k) = b_k^{(j)}$ for $0 \le k \le r$, $0 \le j \le m_k - 1$.

PROOF: To find

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n$$

we need to solve the system

$$\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & \cdots & x_0^n \\ 0 & 1 & 2x_0 & 3x_0^2 & \cdots & nx_0^{n-1} \\ 0 & 0 & 2 & 6x_0 & \cdots & n(n-1)x_0^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & n(n-1)\cdots(n-m_r+2)x_r^{n-m_r+1} \end{bmatrix} \begin{bmatrix} a_0 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} b_0^{(0)} \\ b_0^{(1)} \\ b_0^{(2)} \\ \vdots \\ b_1^{(0)} \\ \vdots \\ b_r^{(m_r-1)} \end{bmatrix}$$

where there are m_k rows for each x_k. It can be shown (Exercise V.11.8.1.) that the matrix on the left is invertible if the x_k are distinct, so the system has a unique solution. 👍

V.11.6.3. COROLLARY. Let f be a function on an interval I. Let $n \in \mathbb{N}$, $x_0, \dots, x_r$ distinct numbers in I, and $m_0, \dots, m_r \in \mathbb{N}$ with $m_0 + \cdots + m_r = n + 1$ (so necessarily $r \le n$). Suppose f is $(m_k - 1)$-times differentiable at x_k for each k. Then there is a unique polynomial p of degree $\le n$, with real coefficients, such that $p^{(j)}(x_k) = f^{(j)}(x_k)$ for $0 \le k \le r$, $0 \le j \le m_k - 1$.

V.11.6.4. DEFINITION. The polynomial p is called the *osculating polynomial* of f with respect to $(x_0, \dots, x_r; m_0, \dots, m_r)$.

V.11.6.5. The osculating polynomial generalizes both the interpolating polynomial (the case $r = n$ and all $m_k = 1$) and the Taylor polynomial (the case $r = 0$, $m_0 = n + 1$). If f is sufficiently differentiable, the osculating polynomial is a limiting case of interpolation polynomials for distinct points approaching the repeated points (this is essentially how TAYLOR found the Taylor polynomials).

V.11.6.6. In fact, there is a formula for the osculating polynomial using the extended definitions of higher-order divided differences which is just an extension of the Newton form of the interpolation polynomial:

$$p(x) = f[x_0] + (x - x_0)f[x_0, x_0] + \cdots + (x - x_0)^{m_0-1} f[x_0, \dots, x_0] + (x - x_0)^{m_0} f[x_0, \dots, x_0, x_1]$$

$$+(x-x_0)^{m_0}(x-x_1)f[x_0, \dots, x_0, x_1, x_1]+\cdots+(x-x_0)^{m_0}(x-x_1)^{m_1}\cdots(x-x_r)^{m_r-1}f[x_0, \dots, x_0, x_1, \dots, x_1, \dots, x_r]$$

where x_{k+1} starts appearing when x_k has been repeated m_k times. There is also an exact formula

$$f(x) = f[x_0] + (x - x_0)f[x_0, x_0] + \cdots + (x - x_0)^{m_0-1} f[x_0, \dots, x_0] + (x - x_0)^{m_0} f[x_0, \dots, x_0, x_1]$$

$$+(x-x_0)^{m_0}(x-x_1)f[x_0, \dots, x_0, x_1, x_1]+\cdots+(x-x_0)^{m_0}(x-x_1)^{m_1}\cdots(x-x_r)^{m_r}f[x_0, \dots, x_0, x_1, \dots, x_1, \dots, x_r, x]$$

where x_k is repeated m_k times, which is valid for $x \neq x_0, \dots, x_r$, and even also for $x = x_k$ if f is m_k-times differentiable at x_k.

We thus get a formula for the error from approximation by the osculating polynomial: if $E(x) = f(x) - p(x)$, we have

$$E(x) = (x - x_0)^{m_0}(x - x_1)^{m_1} \cdots (x - x_r)^{m_r} f[x_0, \dots, x_0, x_1, \dots, x_1, \dots, x_r, x]$$

and, combining this with V.11.5.9., we obtain:

V.11.6.7. COROLLARY. Let f be $(n + 1)$-times differentiable on I, $x_0, \dots, x_r$ distinct points of I, and $m_0, \dots, m_r \in \mathbb{N}$ with $m_0 + \cdots + m_r = n + 1$ (so necessarily $r \leq n$). Let p be the osculating polynomial of f with respect to $(x_0, \dots, x_r; m_0, \dots, m_r)$. Then, for any $x \in I$, there is a c in the smallest subinterval of I containing $x_0, \dots, x_r, x$ such that

$$f(x) - p(x) = (x - x_0)^{m_0}(x - x_1)^{m_1} \cdots (x - x_r)^{m_r} \frac{f^{(n+1)}(c)}{(n + 1)!} .$$

Note that both V.11.4.5. and the Lagrange form of the error in Taylor's Theorem (V.10.3.3.) are special cases.

V.11.6.8. An osculating polynomial where all m_k are 2 (n is then necessarily odd) is called a *Hermite polynomial*, and approximation by such a polynomial is called *Hermite interpolation*. (Approximation with general $m_k \geq 2$ is also sometimes called Hermite interpolation.)

V.11.6.9. In practical approximation problems, it can happen, although it is somewhat unusual, that one has data allowing approximation of an unknown function by an osculating polynomial, especially a Hermite polynomial. Such an approximation, when possible, will generally (but not always) be more accurate than approximation by the interpolation polynomial for the same data points.

V.11.7. How Good are the Approximations?

It is beyond the scope of this book to give a comprehensive discussion of the accuracy (not to mention the efficiency) of the various ways to approximate functions by polynomials. But we will briefly survey some of the aspects of accuracy in this section. Almost any book on numerical analysis contains a much more thorough treatment. The article http://people.maths.ox.ac.uk/trefethen/mythspaper.pdf has some important observations about polynomial approximation. See also [**?**].

Approximation With Evenly-Spaced Nodes

V.11.7.1. If $I = [a, b]$, the most obvious choice of $x_0, \dots, x_n$ is as a partition of I into n subintervals of equal length $h = \frac{b-a}{n}$, i.e. $x_k = x_0 + kh$. In this case, the denominators in the recursive formula (V.11.3.6.) for the divided differences are all equal to h, so we may instead consider just the ordinary differences

$$\Delta f(x_j) = \Delta^1 f(x_j) = f(x_{j+1}) - f(x_j)$$

$$\Delta^{k+1} f(x_j) = \Delta^k f(x_{j+1}) - \Delta^k f(x_j)$$

(these are usually called *forward differences*), giving the formula (easily proved by induction)

$$f[x_0, \dots, x_k] = \frac{1}{k! h^k} \Delta^k f(x_0) \ .$$

We also use the notation

$$\pi_k(t) = t(t-1)\cdots(t-k+1)$$

for $k \geq 1$ (we may set $\pi_0(t) = 1$; π_k is called the k'th *factorial polynomial*); cf. V.3.11.14.. The Newton formula for the interpolation polynomial then becomes, for $0 \leq t \leq n$,

$$p_{0,\dots,n}(x_0 + th) = f(x_0) + \sum_{k=1}^{n} \frac{\pi_k(t)}{k!} \Delta^k f(x_0) = \sum_{k=0}^{n} \frac{\pi_k(t)}{k!} \Delta^k f(x_0)$$

if we set $\Delta^0 f(x_0) = f(x_0)$. This is the *Gregory-Newton interpolation formula*. We may further rewrite this using the generalized binomial coefficient

$$\binom{t}{k} = \frac{\pi_k(t)}{k!}$$

(cf. V.12.5.1.), so

$$p_{0,\dots,n}(x_0 + th) = \sum_{k=0}^{n} \binom{t}{k} \Delta^k f(x_0) \ .$$

We may get similar formulas for backward differences or centered differences; see books on numerical analysis for details.

V.11.7.2. If f is continuous on $[a,b]$ and $(n+1)$-times differentiable on (a,b), we get a corresponding formula for the error term for $0 \le t \le n$:

$$E_n(x_0+th) = \pi_n(t)h^{n+1}\frac{f^{(n+1)}(c)}{(n+1)!}$$

for some $c \in (a,b)$.

V.11.7.3. We also get a formula for the Lagrange form of the interpolation formula using evenly-spaced nodes. Using the same notation as above, we have

$$p_{0,\dots,n}(x_0+th) = \frac{\pi_n(t)}{n!}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\frac{f(x_k)}{t-k} \ .$$

The Runge Phenomenon

V.11.7.4. Suppose f is a function on $[a,b]$. One might expect that as $n \to \infty$, the interpolation polynomials $p_{0,\dots,n}$ for f corresponding to a subdivision of $[a,b]$ into n subintervals of equal length should converge to f pointwise or even uniformly on $[a,b]$ if f is sufficiently smooth. But this fails dramatically in general, even if f is analytic. This phenomenon was found by C. Runge in 1901, although related work had previously been done by C. Meray.

V.11.7.5. Runge's classic example is the innocent-looking function

$$g(x) = \frac{1}{x^2+1}$$

on the interval $[-5,5]$, or equivalently by rescaling

$$f(x) = \frac{1}{25x^2+1}$$

on $[-1,1]$. These functions are real-analytic on all of $\mathbb{R}$, and complex-analytic on $\mathbb{C}$ except for two poles. Figure V.10 (from http://www.math.utk.edu/~ccollins/M471/runge.jpg) shows some approximants of small order, which suggest there is an approximation problem near the ends of the interval.

V.11.7.6. In fact, the approximants for g converge pointwise for

$$|x| < x_c \approx 3.63$$

and u.c. on $(-x_c, x_c)$, but diverge pointwise for $x_c < |x| < 5$. The demonstration http://demonstrations.wolfram.com/RungesPhenomenon shows just how wildly the approximations diverge near the endpoints.

V.11.7.7. It should not be too surprising that there are problems near the endpoints, since interpolation near the endpoints is almost extrapolation, for which one would expect problems. In fact, the polynomial π_n has large values near the ends of the interval $[0,n]$, so the error in V.11.7.2. might tend to blow up (and often does) toward the ends of the interval as n increases.

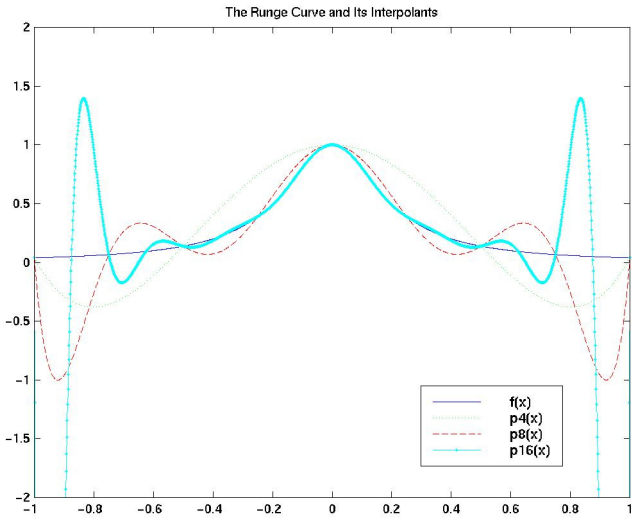


Figure V.10: The Runge Phenomenon

V.11.7.8. The real "explanation" of why approximation fails in this example is that the poles of the function in $\mathbb{C}$ are too close to the interval of approximation (cf. Exercise V.11.8.10.). A detailed discussion can be found in [**?**], (), ().

V.11.7.9. A nice discussion of the Runge phenomenon can be found in http://www.maa.org/sites/default/files/images/upload_library/22/Ford/Epperson329-341.pdf.

Chebyshev Nodes

V.11.7.10. It would appear that one gets a better approximation using nodes which are not evenly spaced, but are more concentrated near the endpoints. This is true in reasonable cases. We will work on the interval $[-1, 1]$ for simplicity; other intervals can be handled via an affine transformation.

V.11.7.11. If f is smooth, the error in approximating by $p_{0,\dots,n}$ for $\{x_0, \dots, x_n\}$ is given by

$$(x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(c)}{(n+1)!}$$

for some $c \in (-1, 1)$. To try to minimize this, we can choose the x_k so that the polynomial

$$q(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$$

is uniformly as small as possible on $[-1, 1]$. Chebyshev showed that the minimum occurs when

$$x_k = \cos\left(\frac{2k+1}{2(n+1)}\pi\right) .$$

These are called the *Chebyshev nodes* (note that they do not form a partition of $[-1,1]$; the endpoints are not nodes). If the x_k are the Chebyshev nodes, the maximum value of $q(x)$ is 2^{-n}, so the error in the approximation can be bounded:

$$|f(x) - p_{0,\dots,n}(x)| \leq \frac{1}{2^n(n+1)!} \sup_{-1<c<1} |f^{(n+1)}(c)|$$

for $x \in [-1,1]$.

V.11.7.12. In fact, the Chebyshev approximations converge uniformly to f if f is Hölder continuous, and in particular if f is $\mathcal{C}^1$ or even Lipschitz (the minimal requirement to insure convergence is somewhat weaker than Hölder continuity; cf. ()). They converge rapidly if f is differentiable to high order. There are, however, examples of continuous f where the Chebyshev approximations do not converge even pointwise. In fact, for any scheme of choosing interpolation points an example of a continuous f can be concocted for which the scheme does not converge (Faber's theorem).

Extrapolation

V.11.7.13. Because the polynomial

$$q(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$$

grows rapidly outside the smallest interval containing $x_0, \dots, x_n$, the error estimate from V.11.4.5. or V.11.6.6. quickly blows up outside this interval. In fact, extrapolation using an interpolating or osculating polynomial is not generally too accurate outside the interval of nodes.

V.11.7.14. Using a Taylor polynomial for f at x_0 to approximate $f(x)$ for $x \neq x_0$ is really always extrapolation, so it is not surprising that the Taylor polynomials do not always give good approximations to $f(x)$ for a fixed $x \neq x_0$. It is actually more surprising (from the interpolation-extrapolation point of view) that they really do give good approximations in so many important cases.

Other Types of Approximation

V.11.7.15. The Chebyshev polynomials are not usually the best uniform approximants to a smooth function f. It can be shown that for any f which is continuous on $[a,b]$, there is a unique polynomial p of degree $\leq n$ which approximates f uniformly most closely on $[-1,1]$ (*minimax approximation*), i.e. for which $\|f - p\|_\infty$ is minimum [existence of a minimizing polynomial is a simple compactness argument, but uniqueness is not so obvious since $\|\cdot\|_\infty$ is not strictly convex], and p is an interpolation polynomial for f for some choice of $x_0, \dots, x_n$ (*Chebyshev's equioscillation theorem*, cf. Exercise V.11.8.9.). But there is no reasonable procedure for finding the best $x_0, \dots, x_n$ in general, even if f is smooth, although there is an iterative procedure (the *Remez algorithm*) which converges rapidly in good cases. In reasonable cases, the Chebyshev polynomials come close to realizing the minimum. See [May06] for a nice discussion of Chebyshev's equioscillation theorem with illustrative applets.

V.11.7.16. Another way of uniformly approximating a continuous f is by Bernstein polynomials (XX.10.1.2.). These are not interpolation polynomials and converge uniformly to f relatively slowly even if f is smooth, but they are guaranteed to converge uniformly to f for any continuous f.

V.11.7.17. One can also try for approximation of f in other norms besides the uniform norm. One alternative which can be efficiently computed is using the 2-norm (*least-square approximation*). See ().

V.11.7.18. One can also use other functions besides polynomials to approximate, e.g. rational functions or trigonometric functions. There is an extensive theory of such approximations, discussed in numerical analysis texts.

V.11.8. Exercises

V.11.8.1. (a) Verify the formula for the determinant of the Vandermonde matrix (V.11.1.1.). [Regard $x_0, \dots, x_n$ successively as variables. The determinant is a polynomial in these variables which vanishes when $x_k = x_j$ for $j \neq k$, so $x_k - x_j$ is a factor.]
(b) Similarly compute the determinant of the matrix in the proof of V.11.6.2. and show it is nonzero. [Experiments with small-size examples using a computer algebra system can be helpful in guessing the correct formula.]
(c) Give an alternate proof that the system of V.11.6.2. (and the special case V.11.1.1.) always has a unique solution without computing determinants, as follows. First show that the homogeneous system (all $b_k^{(j)} = 0$) has a unique solution [a solution q is a polynomial of degree $\leq n$ with at least $n+1$ roots, counted with multiplicity]. This implies that the matrix is invertible, so the system is uniquely solvable for any $b_k^{(j)}$'s.

V.11.8.2. (a) Verify the formula for K in V.11.3.1. and the formula in V.11.3.2..
(b) Verify the recursive formula for $f[x_0, \dots, x_n]$ in V.11.3.6.. [Use that

$$f[x_0] + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] + \cdots + (x - x_0)(x - x_1)\cdots(x - x_{n-1})f[x_0, \dots, x_n]$$

$$= f[x_n] + (x - x_n)f[x_n, x_{n-1}] + (x - x_n)(x - x_{n-1})f[x_n, x_{n-1}, x_{n-2}] + \cdots + (x - x_n)(x - x_{n-1})\cdots(x - x_1)f[x_n, \dots, x_0]$$

(both are the interpolating polynomial) and equate the coefficients of x^{n-1}.]

V.11.8.3. Suppose f is $\mathcal{C}^r$ on an interval I ($r \geq 1$), and $x_0 \in I$. Set

$$g(x) = \begin{cases} f[x_0, x] & \text{if} \quad x \neq x_0 \\ f'(x_0) & \text{if} \quad x = x_0 \end{cases} .$$

Show that g is $\mathcal{C}^{r-1}$ on I.

V.11.8.4. (a) Suppose f is a function on an interval I, and $x_0, \dots, x_n$ are distinct points of I. Show that, for $x \in I$, x not equal to any x_k,

$$f[x_0, \dots, x_n, x] = \sum_{k=0}^{n} \frac{f[x_k, x]}{\prod_{j \neq k}(x_k - x_j)} .$$

(b) Suppose f is $\mathcal{C}^r$ on I ($r \geq 1$). Set

$$g(x) = f[x_0, \dots, x_n, x] = \sum_{k=0}^{n} \frac{f[x_k, x]}{\prod_{j \neq k}(x_k - x_j)}$$

where $f[x_k, x]$ is extended to $x = x_k$ as in V.11.8.3.. Show that g is $\mathcal{C}^{r-1}$ on I.

V.11.8.5. Write out the details of the proof of V.11.5.10. and V.11.5.13..

V.11.8.6. Show that the formula in V.11.6.6. gives the osculating polynomial.

V.11.8.7. Derive the formula for $p_{0,\dots,n}$ in V.11.7.1. from V.11.3.2., and the Lagrange form in V.11.7.3. from V.11.1.3..

V.11.8.8. For $n \geq 0$, define

$$T_{n+1}(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$$

where $x_0, \dots, x_n$ are the Chebyshev nodes (V.11.7.11.). T_{n+1} is called the $(n+1)$'st *monic Chebyshev polynomial of the first kind*. (References differ on whether to multiply by a conventional factor of 2^n, which is why we use the term "monic Chebyshev polynomial." We will not discuss Chebyshev polynomials of other kinds.)

(a) Show that

$$T_n(x) = 2^{-n+1} \cos(n \cos^{-1} x)$$

for $n \geq 1$ and $-1 \leq x \leq 1$. (This formula can also be used to define T_n for $n = 0$.)

(b) Show that $\max\limits_{|x| \leq 1} |T_n(x)| = 2^{-n+1}$, and that the maximum occurs precisely at the $n+1$ points $\cos \frac{k}{n}$, $0 \leq k \leq n$.

(c) Show that T_n $(n \geq 1)$ has the smallest uniform norm of all monic polynomials of degree $\leq n$, with real coefficients, with n roots in $[-1, 1]$ counted with multiplicity.

(d) Prove the recurrence formula $T_{n+1}(x) = xT_n(x) - \frac{1}{4}T_{n-1}(x)$ for $n \geq 1$.

(d) Show that T_n is a solution to the differential equation

$$(1 - x^2)y'' = xy' - n^2 y \ .$$

(e) Show that

$$T_n(x) = \frac{1}{2}\left[(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n\right] \ .$$

The Chebyshev polynomials also arise in least-squares approximation ().

V.11.8.9. Let f be a continuous function on a closed bounded interval $[a, b]$, and $n \in \mathbb{N}$.

(a) Prove using the Heine-Borel theorem that there is a polynomial p of degree $\leq n$ for which $\|f - p\|_\infty$ is minimum among all polynomials of degree $\leq n$.

(b) If p is a polynomial of degree $\leq n$ and there are $n+2$ distinct points $y_0, \dots, y_{n+1}$ in $[a, b]$ (in increasing order) such that

$$f(y_k) - p(y_k) = \pm \|f - p\|_\infty$$

with alternating signs, then $\|f - p\|_\infty$ is minimum among all polynomials of degree $\leq n$. [If q is a polynomial of degree $\leq n$ with $\|f - q\|_\infty < \|f - p\|_\infty$, then $q(y_k)$ is alternately greater than and less than $p(y_k)$. Thus $p - q$ is a nonzero polynomial of degree $\leq n$ with at least $n+1$ roots by the Intermediate Value Theorem.]

(c) Chebyshev's Equioscillation Theorem states the converse to (b): a polynomial p for which $\|f - p\|_\infty$ is minimum among all polynomials of degree $\leq n$ has the equioscillation property of (b).

(d) Show that the result of (c) implies that the polynomial of degree $\leq n$ of minimum uniform distance from f is unique. [If p and q minimize distance from f, then so does $\phi = \frac{1}{2}(p + q)$, and thus ϕ has the equioscillation property.]

(e) Show by the Intermediate Value Theorem that the closest approximant p is an interpolation polynomial for f with nodes $x_0, \ldots, x_n$ with $y_k < x_k < y_{k+1}$ for all k.

V.11.8.10. Let $f(x) = \frac{1}{x^2+1}$. Sketch the graphs of the Taylor polynomials of f around 0. On what interval do they converge to f? What is the radius of convergence of the Taylor series of f around 0? What about the Taylor series around a general $x_0 \in \mathbb{R}$? Justify this geometrically by considering f as an analytic function with poles on $\mathbb{C}$. This suggests some connection between complex poles and the Runge phenomenon; one would expect interpolating polynomials to approximate somewhat better over an interval than Taylor polynomials, but perhaps not to an unlimited extent.

V.12. Miscellaneous Topics in Calculus

V.12.1. How Discontinuous Can a Derivative Be?

If f is a differentiable function (on an interval, or all of $\mathbb{R}$), what can be said about the properties of the function f'?

V.12.1.1. EXAMPLE. If f is differentiable, the function f' need not itself be differentiable everywhere. For perhaps the simplest example, let

$$f(x) = \begin{cases} x^2 & \text{if } x \geq 0 \\ -x^2 & \text{if } x < 0 \end{cases} .$$

Then we have

$$f'(x) = \begin{cases} 2x & \text{if } x \geq 0 \\ -2x & \text{if } x < 0 \end{cases} .$$

(It must be shown directly from the definition of derivative that $f'(0)$ exists and equals 0.) Thus f is differentiable everywhere, but f' is not differentiable at 0.

V.12.1.2. In this example, f' is continuous. In fact, every continuous function occurs as the derivative of a differentiable function: if g is continuous on an interval I, and $x_0 \in I$, for $x \in I$ define

$$f(x) = \int_{x_0}^{x} g(t)\,dt .$$

Then, by the Fundamental Theorem of Calculus (), $f'(x) = g(x)$.

However, a derivative need not even be continuous:

V.12.1.3. EXAMPLE. Let $f(x) = x^2 \sin \frac{1}{x}$ for $x \neq 0$, and $f(0) = 0$. Then, for $x \neq 0$,

$$f'(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x} .$$

It can be checked from the definition (Exercise ()) that $f'(0)$ exists and equals 0. Thus f is differentiable everywhere, but f' is not continuous at 0 since $\lim_{x\to 0} f'(x)$ does not exist.

In this example, although f' is not continuous, it is bounded on bounded intervals, and measurable, hence Lebesgue integrable on any bounded interval, and f can be recovered from f' by

$$f(x) = \int_{0}^{x} f'(t)\,dt$$

where the integral is a Lebesgue integral. (In fact, since f' is bounded and continuous except at one point, it is Riemann integrable on any bounded interval, so the integral can even be taken to be a Riemann integral in this case.)

But f' need not be even locally Lebesgue integrable:

V.12.1.4. EXAMPLE. Let $f(x) = x^2 \sin \frac{1}{x^2}$ for $x \neq 0$, and $f(0) = 0$. Then, for $x \neq 0$,

$$f'(x) = 2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2} \ .$$

It can again be checked from the definition (Exercise ()) that $f'(0)$ exists and equals 0, so f is differentiable, but f' is unbounded and not Lebesgue integrable in any neighborhood of 0.

V.12.1.5. However, if f is any differentiable function, then f' is always locally H-K integrable to f (). In fact, the function f' of V.12.1.4. is one of the standard examples of a function which is H-K integrable but not Lebesgue integrable.

V.12.1.6. The above examples of derivatives are continuous except at one point. A differentiable function can be easily constructed whose derivative has a dense set of discontinuities (Exercise V.12.4.2.). Much more extreme examples can be constructed; see Exercise V.12.4.3. for an example of a differentiable function whose derivative is discontinuous almost everywhere. One of the most extreme examples is Example () of KATZNELSON and STROMBERG ([KS74]; cf. [vRS82, 13.2]):

V.12.1.7. EXAMPLE. There is a differentiable function f with the following properties:

(i) $|f'(x)| \leq 1$ for all x.

(ii) Each of the following sets is dense in $\mathbb{R}$:

$$A = \{x : f'(x) > 0\}$$
$$B = \{x : f'(x) = 0\}$$
$$C = \{x : f'(x) < 0\}$$

(iii) f is not monotone on any interval.

In fact, "many" differentiable functions satisfy (ii) and (iii) (XIII.12.5.7., XIII.12.5.8.). It follows from (ii) that the discontinuities of f' are dense; in fact, f' must be discontinuous wherever it is nonzero.

By (), f' is not Riemann integrable on any interval, hence f' is not continuous a.e. on any interval (). In fact, the sets A and C both have positive Lebesgue measure.

But the next result, combined with (), shows that the derivative of a differentiable function cannot be "too discontinuous."

V.12.1.8. PROPOSITION. If f is differentiable, then f' is of Baire Class 1 (), and in particular must be Lebesgue measurable, and continuous except on a meager set (which may, however, have positive Lebesgue measure (V.12.1.7., V.12.4.3.)).

PROOF: Define

$$g_n(x) = \frac{f\left(x + \frac{1}{n}\right) - f(x)}{1/n} = n\left[f\left(x + \frac{1}{n}\right) - f(x)\right]$$

(The definition may need slight modification if working on a closed interval; see Exercise V.12.4.1.). Then each g_n is continuous, and $g_n \to f'$ pointwise. 👍

V.12.2. Other Properties of Derivatives

V.12.2.1. Derivatives have another property characteristic of continuous functions: they have the *Intermediate Value Property*, or *Darboux Property*, i.e. they satisfy the conclusion of the Intermediate Value Theorem (V.8.2.13.).

V.12.2.2. So which functions are derivatives, i.e. if g is a function on an interval I, when is there a differentiable function f on I with $f' = g$? This is a difficult question, and no reasonable necessary and sufficient conditions are known. It is *sufficient* that g be continuous, but not necessary. And it is *necessary* that g be a Baire Class 1 function with the Intermediate Value Property, but this is not sufficient either:

V.12.2.3. EXAMPLE. Let $g_t(x) = 2x \sin \frac{1}{x} - \cos \frac{1}{x}$ for $x \neq 0$, and $g_t(0) = t$ (cf. V.12.1.3.). Then g_t is Baire Class 1 for any t (), and has the Intermediate Value Property if $-1 \leq t \leq 1$. But g_t is a derivative if and only if $t = 0$. (The only differentiable function f with $f(0) = 0$ for which f' could be g_t is $f(x) = x^2 \sin \frac{1}{x}$; the derivative of this f is g_0.)

V.12.2.4. A derivative must also be H-K integrable over every closed bounded subinterval of I, which is a quite mild restriction, although not automatic for a Baire Class 1 function with the Intermediate Value Property (e.g. the function $g(x) = \frac{1}{x} \sin^2 \frac{1}{x}$, $g(0) = 0$.)

V.12.2.5. One *can* give one necessary and sufficient condition: we must have that

$$\lim_{h \to 0+} \frac{1}{h} \int_{x_0}^{x_0+h} g(t)\, dt \quad \text{and} \quad \lim_{h \to 0+} \frac{1}{h} \int_{x_0-h}^{x_0} g(t)\, dt$$

must both exist and be equal, for every x_0, and the value of g at x_0 must be equal to this common value, i.e. for every x_0, the value of g at x_0 must be approximately equal to the average value of g over short intervals to the left and right of x_0. However, this condition is not a useful one to check in most cases. Compare with V.8.2.15. and V.9.4.11..

V.12.2.6. EXAMPLE. Let $g(t) = 1$ for $t > 0$ except for thin "spikes" down to 0 around $1/n$, and $g(-t) = -g(t)$ for $t > 0$. (See Fig. ().) Then g is bounded, continuous except at 0 (hence Riemann-integrable over every interval), but $\frac{1}{h} \int_0^h g(t)\, dt$ is approximately 1 and $\frac{1}{h} \int_{-h}^0 g(t)\, dt$ is approximately -1 if the spikes are thin enough. So g is not a derivative no matter how $g(0)$ is defined, i.e. g is not equal a.e. to the derivative of an everywhere differentiable function. If $-1 \leq g(0) \leq 1$, then g has the Intermediate Value Property.

See the article [] for more information on which functions are derivatives.

V.12.3. Analysis on $\mathbb{Q}$

In understanding the crucial role of completeness in real analysis, it is instructive to consider what happens if we just use $\mathbb{Q}$ (or a similar ordered field) as our number system. In this section (up to V.12.3.6.), "number" will always mean "rational number," and we will pretend we don't know about anything else. In fact, when Calculus was first developed, and up to the nineteenth century, completeness was not really understood or used, except sometimes implicitly and somewhat vaguely in assumptions that certain numbers (e.g. sums of certain infinite series) existed.

V.12.3.1. In our rational world, we can develop Calculus mechanically in the usual way: the definitions of limits of sequences and functions can be made identically to the ones in IV.1.3.6. and V.1.1.3., and the properties described in IV.1.3.21. and () hold, including the Sequential Criterion. Continuity can also be defined in the usual way, and the usual functions are continuous, although not many standard functions besides polynomials and rational functions are defined on entire intervals (cf. V.12.3.6.). We can define differentiability and the derivative as in V.3.1.9., and all the usual rules of differentiation hold, and in particular polynomials and rational functions are differentiable everywhere with their usual derivatives. Some properties of derivatives such as V.3.6.3. persist. We can also define Riemann integrability and the Riemann integral using Definitions 1A (IX.2.3.1.) or Definition 1B (IX.2.3.2.), which are equivalent (but not equivalent to the other definitions! And many standard functions are not Riemann-integrable on many intervals.) To verify these facts, note that the proofs do not use any form of completeness and are equally valid in our restricted number system.

V.12.3.2. However, almost all of the basic theorems of Calculus fail to hold. Most obviously, the Completeness Axiom fails, as do the Monotone Sequence Theorem (), the Cauchy Criterion (), and *a fortiori* the Bolzano-Weierstrass Theorem (). For example, the sequence

$$3, 3.1, 3.14, 3.141, 3.1415, 3.14159, 3.141492, 3.1415926, \ldots$$

is a bounded strictly increasing sequence which is a Cauchy sequence, but does not converge and does not even have a convergent subsequence (note that it is still true that a Cauchy sequence with a convergent subsequence must converge ()).

V.12.3.3. The Nested Intervals Property also does not hold. For example, the sequence

$$[3, 4], [3.1, 3.2], [3.14, 3.15], [3.141, 3.142], [3.1415, 3.1416], [3.14159, 3.1416], \ldots$$

is a decreasing sequence of closed bounded intervals whose intersection is empty.

V.12.3.4. In fact, it is a little ambiguous what a "closed bounded interval" should be. Should it be an interval of the form $[a, b]$, or should it be a bounded interval which is topologically closed? Specifically, is the interval

$$\{x : x^2 < 2\}$$

closed, open, both, or neither? Is it even an interval? (It is by the definition in III.1.9.12., and is topologically both open and closed.) Note that although this interval is bounded, it does not have a length.

V.12.3.5. The following functions provide counterexamples to other basic theorems of Calculus.

EXAMPLES. (i) Let

$$f(x) = \frac{1}{x^2 - 2} \ .$$

Then f is defined everywhere [on $\mathbb{Q}$!], and is continuous and even differentiable everywhere. But f is not bounded above or below on $[1, 2]$, and is not uniformly continuous on $[1, 2]$ (cf. ()). It is also not Riemann integrable on $[1, 2]$, in contrast to ().

(ii) Let

$$f(x) = \begin{cases} x & \text{if} \quad x^2 < 2 \\ 0 & \text{if} \quad x^2 > 2 \end{cases} \ .$$

Then f is defined everywhere, and is continuous and even differentiable everywhere. And f is bounded on $[1, 2]$, but does not attain a maximum on this interval.

By (i) and (ii), both conclusions of the Extreme Value Theorem fail to hold.

(iii) Let

$$f(x) = \begin{cases} x & \text{if} \quad x^2 < 2 \\ 3 - x & \text{if} \quad x^2 > 2 \end{cases} \ .$$

Then f is defined everywhere, and is continuous and even differentiable everywhere. We have $f(1) = f(2) = 1$, but there is no $c \in (1, 2)$ with $f'(c) = 0$. Thus Rolle's Theorem, and *a fortiori* the Mean Value Theorem, fail to hold on $[1, 2]$.

(iv) Let

$$f(x) = \begin{cases} 1 & \text{if} \quad x^2 < 2 \\ 0 & \text{if} \quad x^2 > 2 \end{cases} \ .$$

Then f is defined everywhere, and is continuous and even differentiable everywhere, and $f'(x) = 0$ for all x. But f is not constant, contradicting V.8.2.3. in our setting. And f does not satisfy the Intermediate Value Theorem on $[1, 2]$. The Fundamental Theorem of Calculus, Version 2 also does not hold for this f: f' is Riemann integrable on $[1, 2]$, but

$$\int_1^2 f'(x)\, dx \neq f(2) - f(1) \ .$$

Note that f itself is not Riemann integrable on $[1, 2]$, in contrast to IX.2.8.5..

One can do much worse. Let X be a closed subset of $\mathbb{R}$ contained in $\mathbb{J}$ (for example, there is a Cantor set contained in $\mathbb{J}$ (XIII.10.6.9.)). Then $U = \mathbb{R} \setminus X$ is an open set containing $\mathbb{Q}$. U is a finite or countable disjoint union $\cup I_n$ of open intervals; we may let f be constant on each $I_n \cap \mathbb{Q}$, with the constant values chosen arbitrarily (in $\mathbb{Q}$) and independently on the intervals. Such an f is differentiable everywhere and f' is identically 0. Thus a function with an antiderivative has a huge number of antiderivatives, which do not differ by constants.

Perhaps we cheated in the *construction* of these functions since we had to go outside $\mathbb{Q}$, but the *existence* of these functions is an incontestable fact. The underlying reason these pathological functions exist, and the other examples too, is that $\mathbb{Q}$ is totally disconnected (zero-dimensional) as a topological space (XIII.13.6.1.).

V.12.3.6. Using $\mathbb{Q}$ as our number system has other inadequacies independent of the completeness problem: many functions we would like to define have no reasonable definition. For example, e^x makes no sense (cf. III.9.3.5.), and $\cos x$ and $\sin x$ are defined only for $x = 0$ (III.9.3.6.). Even $\sqrt{x}$ has a small domain (although the domain is dense). And the definite integrals of even some rational functions over some intervals, such as

$$\int_1^2 \frac{1}{x}\,dx \quad \text{or} \quad \int_0^1 \frac{1}{1+x^2}\,dx$$

are not defined. [Note that a limited version of the Fundamental Theorem of Calculus, Version 1 holds: if f is a continuous function and *if*

$$A(x) = \int_a^x f(t)\,dt$$

exists (i.e. is in $\mathbb{Q}$) for all $x \in [a, b]$, then A is differentiable on $[a, b]$ and $A' = f$.]

But these inadequacies can be remedied without completeness. It is routine (using Countable AC) to construct a countable subfield $\mathbb{F}$ of $\mathbb{R}$ (necessarily containing $\mathbb{Q}$) such that

> $\mathbb{F}$ is real-closed (), and in particular contains all real algebraic numbers.
>
> $\mathbb{F}$ contains e, π, and all other standard mathematical constants.
>
> If $x, y \in \mathbb{F}$ and $y > 0$, then $\cos x$ (hence $\sin x = \pm\sqrt{1 - \cos^2 x}$, $\tan x$, etc.), $\log y$, y^x, $\Gamma(y)$, $\tan^{-1} x$, and (if $-1 \le x \le 1$) $\sin^{-1} x$ are all in $\mathbb{F}$.
>
> More generally, if $\{f_n : n \in \mathbb{N}\}$ is a countable collection of functions from $\mathbb{R}$ to $\mathbb{R}$, then $f_n(x) \in \mathbb{F}$ for all n and all $x \in \mathbb{F}$ for which $f_n(x)$ is defined. If each f_n is smooth, we may take $\{f_n\}$ to be closed under derivatives and antiderivatives of arbitrary order. We may also obtain that
>
> $$\int_a^b f_n(x)\,dx \in \mathbb{F}$$
>
> for every n and every $a, b \in \mathbb{F}$ for which f_n is defined on $[a, b]$.

If $\mathbb{F}$ is such a field, then all mechanical parts of Calculus involving standard functions can be carried out in $\mathbb{F}$, and existence of antiderivatives and definite integrals, at least of standard functions, is assured (we apparently cannot construct $\mathbb{F}$ so that every continuous function has an antiderivative; we cannot have such an $\mathbb{F}$ so that even indicator functions of clopen intervals have definite integrals over larger intervals, cf. V.12.3.5.(iv)). But counterexamples to most of the theorems of Calculus can be constructed as in V.12.3.5., although it is more delicate to describe them explicitly; one way to do it is to fix a bounded increasing sequence (c_n) in $\mathbb{F}$ with no limit in $\mathbb{F}$ [i.e. whose limit is in $\mathbb{R} \setminus \mathbb{F}$], and define functions like

$$f(x) = \begin{cases} 1 & \text{if } x < c_n \text{ for some } n \\ 0 & \text{otherwise} \end{cases} .$$

V.12.4. Exercises

V.12.4.1. Let f be a differentiable function on a closed bounded interval $[a, b]$. Find a sequence of continuous functions on $[a, b]$ converging pointwise to f'. [Modify the functions g_n in the proof of V.12.1.8. on $[b - \frac{1}{n}, b]$.]

V.12.4.2. Let $\phi(x) = x^2 \sin\left(\frac{1}{x}\right)$ for $x \neq 0$ and $\phi(0) = 0$. Then ϕ is differentiable everywhere ().

(a) Show that $|\phi(x)| \leq 1$ and $|\phi'(x)| \leq 3$ for $x \in [-1, 1]$.

(b) Let (q_k) be an enumeration of the rationals in $[0, 1]$, and for each k set $f_k(x) = 2^{-k}\phi(x - q_k)$. Then f_k is differentiable everywhere, $|f_k(x)| \leq \frac{1}{2^k}$ and $|f'(x)| \leq \frac{3}{2^k}$ for $x \in [0, 1]$.

(c) $\sum_{k=1}^{\infty} f_k$ and $\sum_{k=1}^{\infty} f_k'$ converge uniformly on $[0, 1]$ to functions f and g respectively. Apply V.8.5.5. (the hypotheses are more than fulfilled by the partial sums) to show that f is differentiable on $[0, 1]$ and that $g = f'$.

(d) Show that f' is discontinuous at each q_n (and only at these numbers). [For each n, write $f = f_n + \tilde{f}_n$, where $\tilde{f}_n = \sum_{k \neq n} f_k$; repeat the earlier arguments to show that $\tilde{f}_n$ is differentiable on $[0, 1]$ and that $\tilde{f}_n' = \sum_{k \neq n} f_k'$. Since each f_k' for $k \neq n$ is continuous at q_n, $\tilde{f}_n'$ is continuous at q_n. But f_n' is discontinuous at q_n.]

V.12.4.3. [vRS82, 14.3–14.5] If $I = [a, b]$ is a closed bounded interval, let g_I be the function with graph as in Figure (), with $L_I = \frac{b-a}{4}$.

(a) Let f_I be an antiderivative of g_I with $f_I(a) = 0$. Show that $f(b) = 0$ and $0 \leq f(x) \leq L_I^2$ for all $x \in [a, b]$.

(b) Let K be a Cantor set of positive measure in $[0, 1]$, and (I_n) an enumeration of the closures of the open intervals in $[0, 1] \setminus K$. Let f be the function on $[0, 1]$ which is 0 on K and f_{I_n} on I_n, and similarly g the function which is 0 on K and g_{I_n} on I_n. Show that f is differentiable on $[0, 1]$ and $f' = g$. [It is only necessary to show that $f'(a) = 0$ for $a \in K$. If $f(x) \neq 0$, then $x \in I_n$ for some n, so

$$\left|\frac{f(x) - f(a)}{x - a}\right| = \frac{f_{I_n}(x)}{|x - a|} \leq \frac{L_{I_n}^2}{|x - a|} \leq |x - a|$$

since $|x - a| \geq L_{I_n}$.]

(c) Show that for $a \in K$, $\limsup_{x \to a} g(x) = 1$, so g is discontinuous at each point of K, i.e. f' is discontinuous on a set of positive measure.

(d) Show that for $a \in K$, $\liminf_{x \to a} g(x) = 0$, so g is lower semicontinuous.

(e) Let K_n be a Cantor set in $[0, 1]$ of measure $1 - \frac{1}{n}$, and h_n the g constructed as above for K_n. Set $h = \sum_{n=1}^{\infty} 2^{-n} h_n$. Then h has an antiderivative ϕ on $[0, 1]$ ().

(f) Since each h_n is lower semicontinuous, h is discontinuous wherever one of the h_n is discontinuous (). Thus h is discontinuous at all points of $\cup K_n$, i.e. ϕ' is discontinuous a.e.

V.12.4.4. [vRS82, 14.J] Let g be a Darboux continuous function on an interval I which is of Baire Class 1 (e.g. $g = f'$ for a differentiable function f on I). Show that the graph of g is connected.

When is a function f from $\mathbb{R}$ (or $[a,b]$) to $\mathbb{R}$ the derivative of an everywhere differentiable function? Necessary conditions:

(1) f is Baire class 1, i.e. a pointwise limit of a sequence of continuous functions [if $f = F'$, set $f_n(t) = n[F(t+1/n) - F(t)]$; then f_n is continuous and $f_n \to f$.] This implies:

  (a) f is Lebesgue measurable.

  (b) f is continuous except on a set of first category [**?**].

(2) f is Henstock-Kurzweil integrable on every bounded interval [**?**].

(3) f has the Intermediate Value Property.

These properties are not sufficient:

EXAMPLES.

(i) Let $f(t) = \sin(1/t)$ for $t \neq 0$. For any α, $-1 \leq \alpha \leq 1$, define $f(0) = \alpha$. Then f has the above properties, but is a derivative if and only if $\alpha = 0$.

(ii) Let $f(t) = 1$ for $t > 0$ except for thin "spikes" down to 0 around $1/n$, and $f(-t) = -f(t)$ for $t > 0$, $f(0) = 0$. Then f has the above properties, is bounded, continuous except at 0, but not equal a.e. to the derivative of an everywhere differentiable function.

V.12.5. The Binomial Theorem

The Binomial Theorem in algebra gives a formula for positive integral powers of a sum of two numbers. The Binomial Theorem can be extended to apply also to negative and nonintegral powers, but the extended version involves infinite series and there are convergence questions which must be addressed.

Recall that the *binomial coefficients* are defined by the following formula:

V.12.5.1. DEFINITION. If $n \in \mathbb{N}$, $0 \leq k \leq n$, then

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}$$

V.12.5.2. The last expression makes sense for any $n \in \mathbb{R}$ if $k \in \mathbb{N}$, and hence we can define the generalized binomial coefficient

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-k+1)}{k!}$$

for $\alpha \in \mathbb{R}$, $k \in \mathbb{N}$. We set $\binom{\alpha}{0} = 1$ for all α (even $\alpha = 0$). Note that if $n \in \mathbb{N} \cup \{0\}$, then $\binom{n}{k} = 0$ for $k > n$; but if $\alpha \notin \mathbb{N} \cup \{0\}$, then $\binom{\alpha}{k} \neq 0$ for all k. $\binom{\alpha}{k}$ is read "α choose k," since if $n \in \mathbb{N}$, $\binom{n}{k}$ is just the number of ways of choosing k elements out of an n-element set.

Here is the standard Binomial Theorem of algebra:

V.12.5.3. THEOREM. Let $x, y \in \mathbb{C}$, $n \in \mathbb{N}$. Then

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

(with the convention that $t^0 = 1$ for all t, even $t = 0$.) The formula actually holds (suitably interpreted) in any commutative ring.

This theorem can easily be proved by induction on n using the following relation among binomial coefficients, which can be proved by induction on k:

V.12.5.4. PROPOSITION. Let $\alpha, \beta \in \mathbb{R}$, $k \in \mathbb{N} \cup \{0\}$. Then

$$\binom{\alpha+\beta}{k} = \sum_{j=0}^{k} \binom{\alpha}{k-j}\binom{\beta}{j}$$

In particular, if $k \in \mathbb{N}$,

$$\binom{\alpha+1}{k} = \binom{\alpha}{k} + \binom{\alpha}{k-1}$$

The fact that the binomial coefficient $\binom{n}{k}$ is always a natural number if n and k are natural numbers with $k \leq n$ has an interesting consequence in number theory, which is not entirely obvious:

V.12.5.5. COROLLARY. If $k \in \mathbb{N}$, then the product of any k consecutive integers is divisible by $k!$.

PROOF: Let n be the largest number in the sequence. Then $n \geq k$, and

$$\frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \binom{n}{k} \in \mathbb{N} .$$

V.12.5.6. COROLLARY. Let g be a polynomial with integer coefficients, and $k \in \mathbb{N}$. Then all coefficients of the k'th derivative $g^{(k)}$ are divisible by $k!$.

PROOF: It suffices to prove this for $g(x) = x^n$. If $n \geq k$, then $g^{(k)}(x) = n(n-1)(n-2)\cdots(n-k+1)x^{n-k}$, and $g^{(k)}$ is identically 0 if $k > n$.

This result is also an easy corollary of Taylor's Theorem.

V.12.5.7. To prove a version of the Binomial Theorem when n is not a positive integer, we restrict to the case where $y = 1$ and regard $(1+x)^n$ as a function of x. This is not an essential restriction, since if $y \neq 0$ we have $(x+y)^n = y^n \left(\frac{x}{y} + 1\right)^n$.

Fix $\alpha \in \mathbb{R}$ and consider the function $f(x) = (1+x)^\alpha$. This function is infinitely differentiable on $(-1, \infty)$ and extends to a continuous function on $[-1, \infty)$ if $\alpha \geq 0$. For certain special α it also extends to numbers less than -1.

V.12.5.8. PROPOSITION. For any $k \in \mathbb{N} \cup \{0\}$,

$$\frac{f^{(k)}(0)}{k!} = \binom{\alpha}{k} .$$

Thus the Taylor series for f around 0 is

$$\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k .$$

The proof of the first formula is an easy calculation, and the second formula follows immediately.

V.12.5.9. THEOREM. [BINOMIAL THEOREM] For any $\alpha \in \mathbb{R}$, the series

$$\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$$

converges on $(-1, 1)$ to $(1+x)^\alpha$. If $\alpha \geq 0$, the series converges uniformly on $[-1, 1]$ to $(1+x)^\alpha$.

PROOF: The proof has three steps. (i) Show that the series converges on $(-1,1)$. (ii) Show that the series represents the function on $(-1,1)$. (iii) If $\alpha \geq 0$, show the series converges uniformly on $[-1,1]$. Since the results are obvious when $\alpha \in \mathbb{N} \cup \{0\}$ (since then the sum is finite), we will assume $\alpha \notin \mathbb{N} \cup \{0\}$. We set $a_k = \left|\binom{\alpha}{k}\right|$; then $a_k > 0$ for all k.
(i) Using the Ratio Test, the radius of convergence of the series is

$$\lim_{k\to\infty} \frac{a_k}{a_{k+1}} = \lim_{k\to\infty} \frac{k+1}{k-\alpha} = 1$$

so the series converges for $|x| < 1$.
(ii) For $x \in (-1,1)$, let $f(x) = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$. Then f is differentiable on $(-1,1)$, and

$$f'(x) = \sum_{k=1}^{\infty} k\binom{\alpha}{k} x^{k-1} = \sum_{k=0}^{\infty} (k+1)\binom{\alpha}{k+1} x^k$$

on $(-1,1)$. We have

$$(1+x)f'(x) = \sum_{k=0}^{\infty} \left[(k+1)\binom{\alpha}{k+1} + k\binom{\alpha}{k}\right] x^k \ .$$

Since

$$(k+1)\binom{\alpha}{k+1} = \frac{\alpha(\alpha-1)\cdots(\alpha-k)}{k!} = (\alpha-k)\binom{\alpha}{k}$$

we have

$$(k+1)\binom{\alpha}{k+1} + k\binom{\alpha}{k} = \alpha\binom{\alpha}{k}$$

for any k and hence, since $\alpha \neq 0$,

$$f(x) = \frac{1}{\alpha}(1+x)f'(x) \ .$$

Since $g(x) = (1+x)^{-\alpha}$ is differentiable on $(-1,\infty)$ and $g'(x) = -\alpha(1+x)^{-\alpha-1}$ (), by the product rule

$$\frac{d}{dx}\left((1+x)^{-\alpha} f(x)\right) = (1+x)^{-\alpha} f'(x) - \alpha(1+x)^{-\alpha-1} f(x) = 0 \ .$$

Thus $(1+x)^{-\alpha} f(x)$ is constant, so $f(x) = C(1+x)^{\alpha}$ for some C. Since $f(0) = 1$, $C = 1$.
(iii) Suppose $\alpha > 0$. We will show that $\sum_{k=0}^{\infty} a_k$ converges. If $k > \alpha$, then $\frac{a_{k+1}}{a_k} = \frac{k-\alpha}{k+1}$, so $(k+1)a_{k+1} = (k-\alpha)a_k$, $ka_k - (k+1)a_{k+1} = \alpha a_k > 0$. Thus the sequence (ka_k) is nonnegative and decreasing for $k > \alpha$, and hence $\lambda = \lim_{k\to\infty} ka_k$ exists. We have

$$\sum_{k=0}^{n} (ka_k - (k+1)a_{k+1}) = -(n+1)a_{n+1} \to -\lambda$$

so $\sum_{k=0}^{\infty}(ka_k - (k+1)a_{k+1})$ converges. Since $a_k = \alpha^{-1}(ka_k - (k+1)a_{k+1})$ for $k > \alpha$, $\sum_{k=0}^{\infty} a_k$ converges. Thus, by the Weierstrass M-test, $\sum_{k=0}^{\infty} \binom{\alpha}{k} x^k$ converges absolutely and uniformly on $[-1,1]$. Since by (ii) the series represents $(1+x)^{\alpha}$ on $(-1,1)$, it represents $(1+x)^{\alpha}$ on $[-1,1]$ by continuity. 👍

V.12.6. Exercises

V.12.6.1. (a) Prove the second formula in V.12.5.4. by induction on k. [Hint: show using the inductive step that

$$\binom{\alpha+1}{k+1} = \binom{\alpha}{k}\left[\frac{\alpha-k}{k+1} + \frac{1}{k+1}\right] + \binom{\alpha}{k-1}\frac{\alpha-k+1}{k+1}$$

and consider three terms by multiplying. Show that the first term is $\binom{\alpha}{k+1}$ and the last is $\frac{k}{k+1}\binom{\alpha}{k}$.]

(b) Give a combinatorial argument for this formula if $\alpha \in \mathbb{N}$, interpreting $\binom{n}{k}$ as the number of ways of choosing k elements from an n-element set.

(c) Use this formula to prove V.12.5.3. by induction on n.

(d) Generalize the argument in (a) to prove the first formula in V.12.5.4..

V.12.6.2. Here is an alternate proof of V.12.5.9.(ii) (this was essentially CAUCHY'S original argument). Fix x with $|x| < 1$. Set $F_n(\alpha) = \sum_{k=0}^{n}\binom{\alpha}{k}x^k$.

(a) Mimic the proof of the Ratio test to show that the infinite series $\sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$ converges absolutely and uniformly on $\mathbb{R}$ to a function $F(\alpha)$, i.e. $F_n(\alpha) \to F(\alpha)$ uniformly in α. Thus $F(\alpha)$ is continuous in α.

(b) For fixed $\alpha, \beta \in \mathbb{R}$, use the product formula for power series () to show that

$$F(\alpha)F(\beta) = \sum_{k=0}^{\infty}\left[\sum_{j=0}^{k}\binom{\alpha}{k-j}\binom{\beta}{j}\right]x^k \ .$$

(c) Conclude from V.12.5.4. that $F(\alpha)F(\beta) = F(\alpha+\beta)$.

(d) Apply () to conclude that $F(\alpha) = b^\alpha$ for some $b > 0$. Since $F(1) = 1 + x$, $b = 1 + x$.

V.12.6.3. (a) Show that for any $n \in \mathbb{N}$ and $x \in \mathbb{R}$,

$$(1-x)(1+x+x^2+\cdots+x^n) = 1 - x^{n+1} \ .$$

(b) Show that for all $x \in \mathbb{R}$, $x \neq -1$,

$$\frac{1}{1+x} = 1 - x + x^2 - \cdots + (-1)^n x^n + \frac{(-1)^{n+1}x^{n+1}}{1+x} \ .$$

(c) Show from (b) that if $|x| < 1$, then the series $\sum_{n=0}^{\infty}(-1)^n x^n$ converges to $\frac{1}{1+x}$. Compare with the Binomial Theorem for $\alpha = -1$.

V.12.6.4. Here is yet another proof of V.12.5.9.(ii), based on the Cauchy Remainder Formula for Taylor's Theorem (V.10.7.7.).

(a) Fix x with $|x| < 1$. By the Cauchy remainder formula, for any n there is a $\theta = \theta_n$, $0 < \theta < 1$, such that

$$R_n(x) = (1+x)^\alpha - \sum_{k=0}^{n}\binom{\alpha}{k}x^k = \frac{f^{(n+1)}(\theta x)}{n!}(1-\theta)^n x^{n+1}$$

$$= \frac{\alpha(\alpha-1)\cdots(\alpha-n)(1+\theta x)^{\alpha-n-1}}{n!}(1-\theta)^n x^{n+1}$$
$$= \left[\frac{(\alpha-1)(\alpha-2)\cdots(\alpha-n)}{n!}x^n\right]\left[\alpha x(1+\theta x)^{\alpha-1}\left(\frac{1-\theta}{1+\theta x}\right)^n\right]$$

(b) If $M = \max((1+x)^{\alpha-1}, (1-x)^{\alpha-1})$, the second factor is bounded in absolute value by $M|\alpha x|$ independent of n since $\left|\frac{1-\theta}{1+\theta x}\right| < 1$.

(c) The first factor is $\binom{\alpha-1}{n}x^n$ which goes to 0 as $n \to \infty$ since $\sum \binom{\alpha-1}{n}x^n$ converges by V.12.5.9.(i).

(d) Thus $R_n(x) \to 0$ as $n \to \infty$ and the series converges to $(1+x)^\alpha$.

V.13. The Gamma Function

The Gamma function, while not as well known as trigonometric, exponential, and logarithm functions, is of comparable importance in both theoretical and applied mathematics, and notably in probability and statistics. In this section, we will only develop the most basic aspects of the theory of this function.

The origin of the Gamma function was an attempt to find an analytic expression for the factorial function and extend its definition from $\mathbb{N}$ to $\mathbb{R}$. EULER found both infinite product and improper integral formulas for $n!$, both of which also make sense for nonintegral n:

$$n! = \prod_{k=1}^{\infty} \frac{\left(1+\frac{1}{k}\right)^n}{1+\frac{n}{k}} = \int_0^1 (-\log s)^n \, ds \ .$$

V.13.1. Definition of the Gamma Function

LEGENDRE (1811) used the integral expression of EULER, modified by the substitution $t = -\log s$, to give the first formal definition of the Gamma function:

V.13.1.5. DEFINITION. If $x > 0$, define

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} \, dt \ .$$

Note that LEGENDRE'S definition, for unknown reasons, has a shift of 1 in the argument, so that $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. It has remained the convention in mathematics to this day to include this shift, although it has dubious practical merit.

V.13.1.6. PROPOSITION. The improper Riemann integral $\displaystyle\int_0^{\infty} t^{x-1} e^{-t} \, dt$ converges for $x > 0$ (and only for $x > 0$).

PROOF: The integrand is continuous and positive on $(0, \infty)$, hence the Riemann integral over any closed bounded subinterval is defined. The integral $\int_0^{\infty}$ converges if and only if $I_1 = \int_0^1$ and $I_2 = \int_1^{\infty}$ both converge. For I_1, note that $t^{x-1}e^{-t} < t^{x-1}$ for $0 < t \leq 1$, and $\int_0^1 t^{x-1} \, dt$ converges for $x > 0$. (For the parenthetical assertion, $t^{x-1}e^{-t} \geq e^{-1}t^{x-1}$ for $0 < t \leq 1$, and $\int_0^1 t^{x-1} \, dt$ diverges for $x \leq 0$.)

For I_2, note that, for any $x \in \mathbb{R}$,

$$\lim_{t \to +\infty} \frac{t^{x-1}}{e^{t/2}} = 0$$

by (), and hence $t^{x-1}e^{-t} < e^{-t/2}$ for all sufficiently large t. Since $\int_1^{\infty} e^{-t/2} \, dt$ converges, so does I_2 for any $x \in \mathbb{R}$. 👍

The next result gives a crucial factorial-like recurrence property of the Gamma function.

V.13.1.7. PROPOSITION. If $x > 0$, then $\Gamma(x+1) = x\Gamma(x)$.

PROOF: Let $0 < a < b < \infty$. Compute the integral $\displaystyle\int_a^b t^x e^{-t}\,dt$ by parts:

$$\int_a^b t^x e^{-t}\,dt = \left[-t^x e^{-t}\right]_a^b - \int_a^b xt^{x-1}(-e^{-t})\,dt$$

$$= -a^x e^{-a} + b^x e^{-b} + x\int_a^b t^{x-1}e^{-t}\,dt\ .$$

As $a \to 0$ and $b \to +\infty$, the original integral approaches $\Gamma(x+1)$ and the final expression approaches $x\Gamma(x)$. 👍

V.13.1.8. COROLLARY. If $n \in \mathbb{N} \cup \{0\}$, then $\Gamma(n+1) = n!$. [Recall that $0! = 1$ by definition.]

PROOF: This is an easy proof by induction using the recurrence formula and the direct computation

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = \lim_{b\to+\infty}\int_0^b e^{-t}\,dt = \lim_{b\to+\infty}[-e^{-t}]_0^b = \lim_{b\to+\infty}[-e^{-b}+1] = 1 = 0!\ .$$

👍

So the Gamma function extends the (shifted) factorial function to $(0,\infty)$. This extension has the desired analytic property:

V.13.1.9. THEOREM. The Gamma function is real analytic () (and in particular C^∞) on $(0,\infty)$, and

$$\Gamma^{(n)}(x) = \int_0^\infty t^{x-1}e^{-t}(\log t)^n\,dt\ .$$

This is a special case of the general result () concerning differentiation under the integral sign. It is routine to show that the integral giving $\Gamma^{(n)}(x)$ converges for $x > 0$, and this follows anyway from the general theory.

V.13.1.10. Not many values of $\Gamma(x)$ for noninteger x can be computed exactly (of course, modern computers can easily calculate values to great numerical accuracy). One important value which can be computed () is

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\ .$$

V.13.2. Log-Convexity and a Characterization of the Gamma Function

V.13.2.11. The Gamma function is far from the only function on $(0,\infty)$ extending the (shifted) factorial: an extension can be defined arbitrarily for nonintegers. Γ is not even the only real-analytic function extending the factorial: if f is any real-analytic function vanishing at the positive integers, e.g. $f(x) = \sin \pi x$, then $\Gamma + f$ is an extension.

Satisfying the functional equation $g(x+1) = xg(x)$ for all $x > 0$ is only slightly more difficult to achieve. In fact, if a function g is defined arbitrarily on $(0,1)$, set $g(1) = 1$ and extend g by repeatedly applying the functional equation. It is not even hard to find another real-analytic extension satisfying the functional equation, e.g. $g(x) = \Gamma(x)(1 + \sin 2\pi x)$ ($1 + \sin 2\pi x$ can be replaced by any periodic real-analytic function f with period 1 and $f(1) = 1$).

V.13.2.12. However, Γ seems to be the "simplest" and/or "most natural" extension, due to the multitude of formulas for it and its appearance in many contexts. Many leading mathematicians over almost a 200-year period tried to make this precise with a natural characterization, with varied success. The best characterization (to date) was obtained in 1922 by H. Bohr and J. Mollerup using the notion of log-convexity.

V.13.2.13. It is immediate from V.13.1.9. that $\Gamma''(x) > 0$ for $x \in (0,\infty)$, and thus that Γ is a convex function on $(0,\infty)$. But more is true.

V.13.2.14. Definition. If I is an interval, a function $g : I \to (0,\infty)$ is *log-convex* on I if $\log g$ is convex on I.

V.13.2.15. Alternatively, a positive function g on an interval I is log-convex on I if, for all $a, b \in I$, $0 \le \lambda \le 1$,

$$g(\lambda a + (1-\lambda)b) \le g(a)^{\lambda} g(b)^{1-\lambda} .$$

The right-hand side is a sort of weighted geometric mean of $g(a)$ and $g(b)$. If g is continuous, then by () g is log-convex if and only if, for every $a, b \in I$,

$$g\left(\frac{a+b}{2}\right) \le \sqrt{g(a)g(b)} .$$

V.13.2.16. Proposition. A log-convex function is convex.

Proof: Let g be log-convex on I, and let $a, b \in I$, and $0 \le \lambda \le 1$. Since the log function is concave, we have

$$\log(\lambda g(a) + (1-\lambda)g(b)) \ge \lambda \log g(a) + (1-\lambda) \log g(b) .$$

On the other hand, since g is log-convex, we have

$$\log g(\lambda a + (1-\lambda)b) \le \lambda \log g(a) + (1-\lambda) \log g(b) .$$

Combining these, we get

$$\log g(\lambda a + (1-\lambda)b) \le \log(\lambda g(a) + (1-\lambda)g(b))$$

and it follows that $g(\lambda a + (1-\lambda)b) \le \lambda g(a) + (1-\lambda)g(b)$ since the log function is increasing. 👍

V.13.2.17. The converse is false. $g(x) = x^2$ is convex on $(0, \infty)$, but not log-convex on any subinterval. In fact, an increasing log-convex function on an interval (a, ∞) must grow at least exponentially as $x \to +\infty$.

V.13.2.18. Since a sum of convex functions is convex, it is obvious that a product of log-convex functions is log-convex. It is also clear that a positive scalar multiple of a log-convex function is log-convex. It is less obvious, but true, that a sum of log-convex functions is log-convex (XIX.3.5.7.).

V.13.2.19. LEMMA. Let $a_1, a_2, b_1, b_2, c_1, c_2$ be positive real numbers. If $a_1c_1 - b_1^2 \geq 0$ and $a_2c_2 - b_2^2 \geq 0$, then

$$(a_1 + a_2)(c_1 + c_2) - (b_1 + b_2)^2 \geq 0 \ .$$

PROOF: If a, b, c are positive real numbers, then the polynomial $f(x) = ax^2 + 2bx + c$ takes only nonnegative real values (for x real) if and only if $ac - b^2 \geq 0$. By assumption, the polynomials $f_1(x) = a_1x^2 + 2b_1x + c_1$ and $f_2(x) = a_2x^2 + 2b_2x + c_2$ take only nonnegative real values, and hence

$$f_1(x) + f_2(x) = (a_1 + a_2)x^2 + 2(b_1 + b_2)x + (c_1 + c_2)$$

also takes only nonnegative real values. 👍

V.13.2.20. PROPOSITION. A sum of log-convex functions on an interval I is log-convex on I.

PROOF: Since log-convex functions are continuous, by XIX.3.5.2. it suffices to show that if f and g are log-convex on I and $x, y \in I$, then

$$\left[f\left(\frac{x+y}{2}\right) + g\left(\frac{x+y}{2}\right)\right]^2 \leq (f(x) + g(x))(f(y) + g(y)) \ .$$

Apply XIX.3.5.6. with $a_1 = f(x)$, $a_2 = g(x)$, $b_1 = f\left(\frac{x+y}{2}\right)$, $b_2 = g\left(\frac{x+y}{2}\right)$, $c_1 = f(y)$, $c_2 = g(y)$. 👍

V.13.2.21. The conclusion of XIX.3.5.7. extends to arbitrary finite sums. Since a positive pointwise limit of log-convex functions is obviously log-convex, it also extends to a positive pointwise limit of finite sums of log-convex functions, e.g. a pointwise-convergent infinite series. By approximating the integrals by Riemann sums, we also obtain:

V.13.2.22. PROPOSITION. Let f be a function on $I \times [a, b]$ for some interval I and $a, b \in \mathbb{R}$, $a < b$. If $f(x, t)$ is log-convex on I for each fixed t, and $f(x, t)$ is continuous on $[a, b]$ for each fixed x, then

$$g(x) = \int_a^b f(x, t)\, dt$$

is log-convex on I.

PROOF: Let $n \in \mathbb{N}$ and $h = \frac{b-a}{n}$. Then $g_n(x) = h(f(x, a) + f(x, a+h) + \cdots + f(x, a+(n-1)h)$ is log-convex, and $g_n \to g$ pointwise. Since f is strictly positive, it is obvious that g is also strictly positive. 👍

This result extends to improper integrals:

V.13.2.23. Corollary. Let f be a function on $I \times (a,b)$ for some interval I and $a, b \in [-\infty, \infty]$, $a < b$. If $f(x,t)$ is log-convex on I for each fixed t, $f(x,t)$ is continuous on (a,b) for each fixed x, and the improper Riemann integral $\int_a^b f(x,t)\,dt$ converges for all $x \in I$, then

$$g(x) = \int_a^b f(x,t)\,dt$$

is log-convex on I.

Proof: Let (a_n) be a decreasing sequence with $a_n \to a$, (b_n) an increasing sequence with $b_n \to b$, $a_n < b_n$ for all n. Set

$$g_n(x) = \int_{a_n}^{b_n} f(x,t)\,dt\ .$$

Then g_n is log-convex on I for all n by XIX.3.5.9., and $g_n \to g$ pointwise on I. It is also clear that g is strictly positive. 👍

V.13.2.24. Theorem. The Gamma function is log-convex on $(0,\infty)$.

Proof: We have that $\Gamma(x) \in (0,\infty)$ for $x > 0$. Let $a, b > 0$ and $0 \le \lambda \le 1$.

$$\log\Gamma(\lambda a + (1-\lambda)b) = \log\int_0^\infty e^{-t}t^{\lambda a+(1-\lambda)b-1}\,dt = \log\int_0^\infty (e^{-t}t^{a-1})^\lambda (e^{-t}t^{b-1})^{1-\lambda}\,dt\ .$$

Apply Hölder's inequality () with $p = \frac{1}{\lambda}$ and $q = \frac{1}{1-\lambda}$ to obtain

$$\log\int_0^\infty (e^{-t}t^{a-1})^\lambda (e^{-t}t^{b-1})^{1-\lambda}\,dt \le \log\left[\left(\int_0^\infty e^{-t}t^{a-1}\,dt\right)^\lambda \left(\int_0^\infty e^{-t}t^{b-1}\,dt\right)^{1-\lambda}\right]$$

$$= \lambda\log\Gamma(a) + (1-\lambda)\log\Gamma(b)\ .$$

👍

The next theorem is the main result of this section. Our proof is adapted from the Planet Math website http://planetmath.org/?op=getobj&from=objects&id=3808.

V.13.2.25. Theorem. [Bohr-Mollerup] The Gamma function is the unique function g on $(0,\infty)$ which is log-convex and satisfies $g(1) = 1$ and $g(x+1) = xg(x)$ for all $x > 0$.

Proof: Since Γ satisfies the conditions, we need only prove uniqueness. In fact, it suffices to show that the conditions uniquely determine a function on $(0,1)$, since the rest of the values are then uniquely determined by the recurrence formula.

So let g be a function satisfying the conditions, $0 < x < 1$, and $n \in \mathbb{N}$. Using $n + x = (1-x)n + x(n+1)$, we have

$$g(n+x) \le g(n)^{1-x}g(n+1)^x = g(n)^{1-x}g(n)^x n^x = (n-1)!n^x\ .$$

Similarly, $n+1 = x(n+x) + (1-x)(n+1+x)$, so $n! \le g(n+x)(n+x)^{1-x}$.

Combining these two we get

$$n!(n+x)^{x-1} \le g(n+x) \le (n-1)!n^x$$

and by using the recurrence relation to obtain $g(n+x) = x(x+1)(x+2)\cdots(x+n-1)$,

$$a_n := \frac{n!(n+x)^{x-1}}{x(x+1)\cdots(x+n-1)} \le g(x) \le \frac{(n-1)!n^x}{x(x+1)\cdots(x+n-1)} =: b_n \ .$$

These inequalities hold for every positive integer n. We have

$$\lim_{n\to\infty} \frac{b_n}{a_n} = \lim_{n\to\infty} \left(\frac{n}{n+x}\right)^{x-1} = 1$$

so

$$1 \le \frac{g(x)}{a_n} \le \frac{b_n}{a_n} \to 1$$

and thus $g(x) = \lim_{n\to\infty} a_n$, so g is uniquely determined. 👍

As a corollary of the proof, we get a formula for the unique solution, which must be Γ. This formula is due to Gauss, and is equivalent to Euler's infinite product formula from the introduction to the section.

V.13.2.26. Corollary. For $0 < x \le 1$,

$$\Gamma(x) = \lim_{n\to\infty} \frac{n!n^x}{x(x+1)\cdots(x+n)} \ .$$

V.13.3. Miscellaneous Additional Results

We give only a few additional sample results about the Gamma function. More are pursued in the Exercises.

V.13.3.27. The integral definition of $\Gamma(x)$ makes sense if x is a complex number with positive real part, and defines a complex analytic function on the right half of the complex plane. This function can be analytically continued to the entire complex plane, with simple poles at the nonpositive integers. In fact, Gauss's formula (V.13.2.26.) holds for any complex number x which is not a nonpositive integer. It turns out that this complex Γ has no zeros, so the reciprocal is an entire function.

V.13.3.28. Unlike the trigonometric, exponential, and logarithm functions, the Gamma function does not satisfy any *algebraic differential equation* (roughly, a differential equation whose coefficient functions are algebraic functions). This theorem was proved by Hölder in 1887. In particular, there is no formula for the derivative or antiderivative of the Gamma function in terms of standard functions, including the Gamma function itself.

V.13.3.29. There is an almost bewildering array of other formulas, identities, and applications related to the Gamma function. There are entire books (e.g. [Art64]) written on the subject. See [Dav59] for a good survey of the history and importance of this function. In DAVIS'S words, "[E]ach generation has found something of interest about the gamma function. Perhaps the next generation will also."

V.13.4. Exercises

V.13.4.1. (a) Use Gauss's formula V.13.2.26. to prove the product formula

$$\Gamma(x) = \frac{1}{x} \prod_{k=1}^{\infty} \frac{\left(1 + \frac{1}{k}\right)^x}{1 + \frac{x}{k}}$$

if x is not a nonpositive integer. Discuss convergence of the infinite product.

(b) Derive from this WEIERSTRASS'S product formula

$$\frac{1}{\Gamma(x)} = x e^{\gamma x} \prod_{k=1}^{\infty} \left[\left(1 + \frac{x}{k}\right) e^{x/k} \right]$$

where γ is the Euler-Mascheroni constant (IV.2.16.26.). Show that the product converges for all x, and the convergence is u.c.

(c) Deduce from (b) that Γ has simple poles at the nonpositive integers.

V.13.4.2. (a) Use the product formula to prove that if x is not an integer, then

$$\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin \pi x} \ .$$

(b) Use this to give an alternate proof that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

V.13.4.3. Show that $\Gamma'(1) = -\gamma$, where γ is the Euler-Mascheroni constant (IV.2.16.26.).

V.13.5. Stirling's Formula

In this section, we establish an important and useful approximation to the factorial and gamma functions, attributed to James Stirling (1730), but primarily due to A. de Moivre.

V.13.5.1. Theorem. [Stirling's Formula] For any $n \in \mathbb{N}$, we have

$$\sqrt{2\pi n}\, n^n e^{-n} < n! < \sqrt{2\pi n}\, n^n e^{-n} e^{\frac{1}{12n}} \ .$$

In particular,

$$\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\, n^n e^{-n}} = 1 \ .$$

More generally,

$$\lim_{x\to+\infty} \frac{\Gamma(x)}{\sqrt{\frac{2\pi}{x}}\, x^x e^{-x}} = 1 \ .$$

Informally, we have the approximation

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n} = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}}\, x^x e^{-x} = \sqrt{\frac{2\pi}{x}} \left(\frac{x}{e}\right)^x$$

for $n \in \mathbb{N}$, $x > 0$. The theorem gives precise error estimates. Better error estimates can be obtained by making more careful estimates along the way in the proof; see Exercise ().

V.13.5.2. Stirling's formula can be written as

$$n\log n - n + \frac{1}{2}\log(2\pi n) < \log(n!) < n\log n - n + \frac{1}{2}\log(2\pi n) + \frac{1}{12n} \ .$$

A crude version of this formula can be obtained by approximating

$$\log(n!) = \sum_{k=1}^{n} \log k = \sum_{k=2}^{n} \log k$$

by integrals. We have

$$n\log n - n + 1 = \int_1^n \log x\, dx < \log(n!)$$

$$< \int_2^{n+1} \log x\, dx = (n+1)\log(n+1) - (n+1) - (2\log 2 - 2) = n\log(n+1) - n + \log(n+1) + (1 - 2\log 2) \ .$$

This gives

$$e\left(\frac{n}{e}\right)^n = n^n e^{-n+1} \ < \ n! \ < \ \frac{e}{4}(n+1)^{n+1} e^{-n} = \frac{e(n+1)}{4}\left(\frac{n+1}{e}\right)^n \ .$$

Better approximations can be made using the Euler-Maclaurin formula () to compute the difference between $\log(n!)$ and $\displaystyle\int_1^n \log x\,dx$, and the theorem can be proved this way. See ().

We will proceed differently. The proof we give consists of a sequence of calculations, some of which yield other formulas of interest.

V.13.5.3. Recall that if $|x| < 1$, we have

$$\log(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k \ .$$

Thus, we have, for $|x| < 1$, combining terms and using some reindexing,

$$\log\left(\frac{1+x}{1-x}\right) = \log(1+x) - \log(1-x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k - \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} (-x)^k$$

$$= \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k + \sum_{k=1}^{\infty} \frac{1}{k} x^k = \sum_{k=0}^{\infty} \frac{2}{2k+1} x^{2k+1} = 2x \sum_{k=0}^{\infty} \frac{1}{2k+1} x^{2k} \ .$$

This is in itself a useful formula.

V.13.5.4. Now set $x = \frac{1}{2n+1}$ for $n \in \mathbb{N}$. Then $\frac{1+x}{1-x} = \frac{n+1}{n} = 1 + \frac{1}{n}$, so we obtain

$$\log\left(1 + \frac{1}{n}\right) = \frac{2}{2n+1} \sum_{k=0}^{\infty} \frac{1}{2k+1} \frac{1}{(2n+1)^{2k}}$$

$$\left(n + \frac{1}{2}\right) \log\left(1 + \frac{1}{n}\right) = \sum_{k=0}^{\infty} \frac{1}{2k+1} \frac{1}{(2n+1)^{2k}} \ .$$

Therefore

$$\left(n + \frac{1}{2}\right) \log\left(1 + \frac{1}{n}\right) < 1 + \frac{1}{3} \sum_{k=1}^{\infty} \frac{1}{(2n+1)^{2k}}$$

and the last sum is a geometric series with sum

$$\frac{\frac{1}{(2n+1)^2}}{1 - \frac{1}{(2n+1)^2}} = \frac{1}{(2n+1)^2 - 1} = \frac{1}{4n(n+1)}$$

so we obtain the estimates

$$1 < \left(n + \frac{1}{2}\right) \log\left(1 + \frac{1}{n}\right) < 1 + \frac{1}{12n(n+1)}$$

$$e < \left(1 + \frac{1}{n}\right)^{n+\frac{1}{2}} < e^{1 + \frac{1}{12n(n+1)}}$$

$$1 < \frac{1}{e} \left(1 + \frac{1}{n}\right)^{n+\frac{1}{2}} < e^{\frac{1}{12n(n+1)}}$$

for $n \in \mathbb{N}$.

V.13.5.5. Now set

$$a_n = n!e^n n^{-n-\frac{1}{2}} \ , \quad y_n = e^{-\frac{1}{12n}} \ .$$

We have

$$1 < \frac{a_n}{a_{n+1}} = \frac{1}{e}\left(1+\frac{1}{n}\right)^{n+\frac{1}{2}} < e^{\frac{1}{12n(n+1)}} = \frac{y_{n+1}}{y_n} \ .$$

Thus the sequence (a_n) is nonnegative and decreasing, hence converges to a limit a. The sequence $(a_n y_n)$ is increasing, and since $y_n \to 1$, $a_n y_n \to a$, and $a_n y_n < a < a_n$ for all n. Thus

$$a < a_n < a e^{\frac{1}{12n}}$$

for all n. Since

$$n! = a_n \sqrt{n}\, n^n e^{-n}$$

we obtain that

$$a\sqrt{n}\, n^n e^{-n} < n! < a\sqrt{n}\, n^n e^{-n} e^{\frac{1}{12n}}$$

for $n \in \mathbb{N}$. (This version is due to DE MOIVRE.)

It remains to show that $a = \sqrt{2\pi}$ (this was STIRLING'S contribution to the formula). We first derive WALLIS'S beautiful infinite product formula for π:

V.13.5.6. THEOREM.

$$\lim_{n\to\infty}\left[\frac{1}{2}\cdot\frac{3}{2}\cdot\frac{3}{4}\cdot\frac{5}{4}\cdot\frac{5}{6}\cdot\frac{7}{6}\cdots\frac{2n-1}{2n}\cdot\frac{2n+1}{2n}\right] = \frac{2}{\pi} \ .$$

PROOF: We will prove the equivalent statement

$$\lim_{n\to\infty}\left[\frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdots\frac{2n}{2n-1}\cdot\frac{2n}{2n+1}\right] = \frac{\pi}{2} \ .$$

Use the formula obtained by integration by parts:

$$\int \sin^n x\, dx = \int \sin^{n-1} x \sin x\, dx = -\sin^{n-1} x \cos x + (n-1)\int \sin^{n-2} x \cos^2 x\, dx$$

$$= -sin^{n-1}x\cos x + (n-1)\int \sin^{n-2} x(1-\sin^2 x)\, dx = -sin^{n-1}x\cos x + (n-1)\int \sin^{n-2} x\, dx - (n-1)\int \sin^n x\, dx$$

$$\int \sin^n x\, dx = -\frac{1}{n}\sin^{n-1} x\cos x + \frac{n-1}{n}\int \sin^{n-2} x\, dx$$

$$\int_0^{\pi/2} \sin^n x\, dx = \frac{n-1}{n}\int_0^{\pi/2} \sin^{n-2} x\, dx \ .$$

Iterating, we obtain

$$\int_0^{\pi/2} \sin^{2n} x\, dx = \frac{2n-1}{2n}\cdot\frac{2n-3}{2n-2}\cdots\frac{1}{2}\cdot\int_0^{\pi/2} dx = \frac{2n-1}{2n}\cdot\frac{2n-3}{2n-2}\cdots\frac{1}{2}\cdot\frac{\pi}{2}$$

$$\int_0^{\pi/2} \sin^{2n+1} x\,dx = \frac{2n}{2n+1}\cdot\frac{2n-2}{2n-1}\cdots\frac{2}{3}\cdot\int_0^{\pi/2}\sin x\,dx = \frac{2n}{2n+1}\cdot\frac{2n-2}{2n-1}\cdots\frac{2}{3}\ .$$

We thus obtain, for any n,

$$\frac{\pi}{2} = \frac{2\cdot 2}{1\cdot 3}\cdot\frac{4\cdot 4}{3\cdot 5}\cdot\frac{6\cdot 6}{5\cdot 7}\cdots\frac{(2n)^2}{(2n-1)(2n+1)}\cdot\frac{\int_0^{\pi/2}\sin^{2n} x\,dx}{\int_0^{\pi/2}\sin^{2n+1} x\,dx}\ .$$

Thus to prove the formula it suffices to show

$$\lim_{n\to\infty}\frac{\int_0^{\pi/2}\sin^{2n} x\,dx}{\int_0^{\pi/2}\sin^{2n+1} x\,dx} = 1\ .$$

We have

$$0 \le \sin^{2n+1} x \le \sin^{2n} x \le \sin^{2n-1} x$$

for $0 \le x \le \frac{\pi}{2}$ and thus

$$1 \le \frac{\int_0^{\pi/2}\sin^{2n} x\,dx}{\int_0^{\pi/2}\sin^{2n+1} x\,dx} \le \frac{\int_0^{\pi/2}\sin^{2n-1} x\,dx}{\int_0^{\pi/2}\sin^{2n+1} x\,dx} = \frac{2n+1}{2n}$$

so the result follows from the Squeeze Theorem ().

V.13.5.7. Corollary.

$$\lim_{n\to\infty}\frac{(n!)^2\cdot 2^{2n}}{(2n)!\cdot\sqrt{n}} = \sqrt{\pi}\ .$$

Proof: We have

$$\frac{\pi}{2} = \lim_{n\to\infty}\left[\frac{2^2\cdot 4^2\cdot 6^2\cdots(2n-2)^2}{3^2\cdot 5^2\cdot 7^2\cdots(2n-1)^2}\cdot\frac{(2n)^2}{2n+1}\right] = \lim_{n\to\infty}\left[\frac{2^2\cdot 4^2\cdot 6^2\cdots(2n-2)^2}{3^2\cdot 5^2\cdot 7^2\cdots(2n-1)^2}\cdot 2n\right]$$

since $\lim_{n\to\infty}\frac{2n}{2n+1} = 1$. Thus, taking the square root and multiplying numerator and denominator by $2\cdot 4\cdot 6\cdots(2n-2)$, we have

$$\sqrt{\frac{\pi}{2}} = \lim_{n\to\infty}\left[\frac{2^2\cdot 4^2\cdot 6^2\cdots(2n-2)^2}{(2n-1)!}\cdot\sqrt{2n}\right] = \lim_{n\to\infty}\left[\frac{2^2\cdot 4^2\cdot 6^2\cdots(2n)^2}{(2n)!}\cdot\frac{\sqrt{2n}}{2n}\right]\ .$$

V.13.5.8. To calculate a, note that that

$$\sqrt{\pi} = \lim_{n\to\infty}\left[\frac{(n!)^2\cdot 2^{2n}}{(2n)!\cdot\sqrt{n}}\right] = \lim_{n\to\infty}\left[\frac{(a_n e^{-n} n^{n+1/2})^2\cdot 2^{2n}}{a_{2n}e^{-2n}(2n)^{2n+1/2}\cdot\sqrt{n}}\right] = \lim_{n\to\infty}\frac{(a_n)^2}{a_{2n}\sqrt{2}} = \frac{a}{\sqrt{2}}\ .$$

This completes the proof of Stirling's Formula.

V.13.5.9. To obtain the formula for the Gamma function, note that

$$\lim_{n\to\infty} \frac{n\Gamma(n)}{\sqrt{2\pi n}\, n^n e^{-n}} = \lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\, n^n e^{-n}} = 1$$

$$\lim_{n\to\infty} \frac{\Gamma(n)}{\sqrt{\frac{2\pi}{n}}\, n^n e^{-n}} = 1 \ .$$

To transition to the limit as $x \to +\infty$ through positive real numbers, use XIX.3.7.10. and log convexity of the Gamma function:

The proof using the Euler-Maclaurin formula works directly for the Gamma function for $x > 0$ (and even for complex numbers away from the negative real axis). See, for example, [Str81, p. 468].

V.13.5.10. There is a more precise series giving an approximation to the factorial and the gamma function:

$$n! \approx \sqrt{2\pi n}\, n^n e^{-n} \left[1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} - \frac{571}{2488320n^4} + \cdots\right]$$

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}}\, x^x e^{-x} \left[1 + \frac{1}{12x} + \frac{1}{288x^2} - \frac{139}{51840x^3} - \frac{571}{2488320x^4} + \cdots\right]$$

(the coefficients have a closed formula which is complicated; see e.g. http://mathworld.wolfram.com/StirlingsSeries.html.)

This is an asymptotic expansion () and not a convergent series: for fixed n or x adding more terms eventually gives a worse approximation. The relative error after taking finitely many terms is always smaller than the next term.

V.13.5.11. By approximating the terms in the asymptotic expansion, we obtain a better asymptotic formula for $n!$:

$$n! \approx \sqrt{\left(2n + \frac{1}{3}\right)\pi} \left(\frac{n}{e}\right)^n \ .$$

This approximation is due to R. W. Gosper.

V.13.5.12. Another numerically efficient approximation, due to G. Nemes [**?**], is

$$\Gamma(x) \approx \sqrt{\frac{2\pi}{x}} \left[\frac{1}{e}\left(x + \frac{1}{12x - \frac{1}{10x}}\right)\right]^x \ .$$

V.13.6. Exercises

V.14. The Volume of the Unit Ball in $\mathbb{R}^n$

Throughout this section, we will use the Euclidean norm

$$\|(x_1, \ldots, x_n)\| = \sqrt{x_1^2 + \cdots + x_n^2} \ .$$

Recall that the closed unit ball and unit sphere in $\mathbb{R}^n$ are

$$B^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq 1\}$$

$$S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\} \ .$$

Our objective in this section is to compute the Lebesgue measure $\beta_n = \lambda^{(n)}(B^n)$ of the unit ball in $\mathbb{R}^n$. These numbers will be needed for a number of purposes, such as the development of Hausdorff measure, and are intimately tied up with the measure theory of hypersurfaces and other subsets of Euclidean space (such as the $(n-1)$-dimensional "hypersurface area" of S^{n-1}).

V.14.0.1. We can compute β_n for small values of n by elementary or calculus methods:

We have $\beta_1 = \lambda([-1, 1]) = 2$.

We have that β_2 is the area of the unit disk in $\mathbb{R}^2$. This can be computed as in () or () to obtain $\beta_2 = \pi$.

Using spherical coordinates (), we may obtain $\beta_3 = \frac{4}{3}\pi$.

Actually, β_n can be calculated for any n just using calculus, either directly (Exercise V.14.4.1.) or using "hyperspherical" coordinates (Exercise V.14.4.2.). However, the formula obtained in this way becomes increasingly complicated for large n. (Try to guess the right formula!)

Instead, we will use some machinery from measure theory to rather painlessly compute a simple formula for β_n. This argument is adapted from [Fol99, 2.7].

V.14.1. Radial Borel Sets and Functions

Fix an n. We will examine radially symmetric sets in $\mathbb{R}^n$.

V.14.1.2. Definition. A subset S of $\mathbb{R}^n$ is *radially symmetric* if $\mathbf{x} \in S$, $\|\mathbf{y}\| = \|\mathbf{x}\|$ imply $\mathbf{y} \in S$.
B^n and S^{n-1} are radially symmetric sets in $\mathbb{R}^n$. A set is radially symmetric if and only if it is a union of spheres of various radii around the origin.

V.14.1.3. If $A \subseteq [0, \infty)$, the *symmetrization* of A in $\mathbb{R}^n$ is

$$s_n(A) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \in A\} \ .$$

It is clear that $s_n(A)$ is radially symmetric. Conversely, identify $[0, \infty)$ with the ray

$$R_+^{(n)} = \{(t, 0, \ldots, 0) : t \in [0, \infty)\} \subseteq \mathbb{R}^n \ .$$

If $S \subseteq \mathbb{R}^n$ is radially symmetric, set $A = S \cap R_+^{(n)}$; then $S = s_n(A)$.

V.14.1.4. If A is a subset of $[0,\infty)$, then $s_n(A) = f^{-1}(A)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is the norm function $f(\mathbf{x}) = \|\mathbf{x}\|$. Since f is continuous, if $B \subseteq [0,\infty)$ is a Borel set, then $s_n(B)$ is a Borel set in $\mathbb{R}^n$. Conversely, if S is a radially symmetric Borel set in $\mathbb{R}^n$, then $B = S \cap R_+^{(n)}$ is a Borel set in $[0,\infty)$ with $S = s_n(B)$.

V.14.1.5. DEFINITION. A *radial Borel set* is a radially symmetric Borel subset of $\mathbb{R}^n$. Let $\mathcal{R}^{(n)}$ be the set of all radial Borel sets in $\mathbb{R}^n$.

The next proposition is obvious, since the collection of radially symmetric subsets of $\mathbb{R}^n$ is closed under unions, intersections, and complements.

V.14.1.6. PROPOSITION. $\mathcal{R}^{(n)}$ is a sub-σ-algebra of $\mathcal{B}^{(n)}$.

V.14.1.7. DEFINITION. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *radially symmetric* if $\|\mathbf{x}\| = \|\mathbf{y}\|$ implies $f(\mathbf{x}) = f(\mathbf{y})$. If $g : [0,\infty) \to \mathbb{R}$, the *symmetrization* $s_n(g)$ is defined by $[s_n(g)](\mathbf{x}) = g(\|x\|)$. Then $s_n(g)$ is radially symmetric, and every radially symmetric function is of this form.

V.14.1.8. PROPOSITION. (i) A function $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{R}^{(n)}$-measurable if and only if it is Borel measurable and radially symmetric.
(ii) If $g : [0,\infty) \to \mathbb{R}$, then $s_n(g)$ is $\mathcal{R}^{(n)}$-measurable if and only if g is Borel measurable.

PROOF: If f is $\mathcal{R}^{(n)}$-measurable, then f is Borel measurable since $\mathcal{R}^{(n)} \subseteq \mathcal{B}^{(n)}$, and $\{\mathbf{x} : f(\mathbf{x}) = \alpha\} \in \mathcal{R}^{(n)}$ for any $\alpha \in \mathbb{R}$, so f is radially symmetric. The converse and (ii) are obvious from the fact that $\mathcal{R}^{(n)}$ consists precisely of the radially symmetric Borel sets, the symmetrizations of Borel sets in $[0,\infty)$. 👍

The main result of this subsection is:

V.14.1.9. THEOREM. Let g be a nonnegative Borel measurable function on $[0,\infty)$. Then

$$\int_{\mathbb{R}^n} s_n(g)\, d\lambda^{(n)} = n\beta_n \int_0^\infty r^{n-1} g(r)\, d\lambda(r) \ .$$

PROOF: We first prove the result for "annuli." Let $S = s_n((a,b])$, where $0 \le a < b < \infty$. Then S is a set-theoretic difference of two closed balls around the origin, so by the scaling invariance of $\lambda^{(n)}$ we have

$$\int_{\mathbb{R}^n} \chi_S\, d\lambda^{(n)} = \lambda^{(n)}(S) = \beta_n(b^n - a^n) = n\beta_n \int_0^\infty r^{n-1} \chi_{(a,b]}(r)\, d\lambda(r) \ .$$

Now let $\mathcal{S}$ be the collection of all subsets of $\mathbb{R}^n$ of the form $s_n((a,b])$ as above. Then $\mathcal{S}$ is a semialgebra, and the σ-algebra generated by $\mathcal{S}$ is $\mathcal{R}^{(n)}$. Let μ be the measure on $(\mathbb{R}^n, \mathcal{R}^{(n)})$ given by

$$\mu(s_n(B)) = n\beta_n \int_B r^{n-1}\, d\lambda(r) \ .$$

Since we have shown that $\mu|_{\mathcal{S}} = \lambda^{(n)}|_{\mathcal{S}}$, we have $\mu = \lambda^{(n)}|_{\mathcal{R}^{(n)}}$ by (). Thus the formula in the theorem holds for radially symmetric characteristic functions, and hence for radially symmetric simple functions by linearity.

Finally, let (f_k) be an increasing sequence of nonnegative simple functions on $[0,\infty)$ converging pointwise to g. Then $(s_n(f_k))$ is an increasing sequence of radially symmetric simple functions increasing pointwise to $s_n(g)$, so by two applications of the Monotone Convergence Theorem we get

$$\int_{\mathbb{R}^n} s_n(g)\,d\lambda^{(n)} = \lim_{k\to\infty}\int_{\mathbb{R}^n} s_n(f_k)\,d\lambda^{(n)} = \lim_{k\to\infty} n\beta_n\int_0^\infty r^{n-1}f_k(r)\,d\lambda(r) = n\beta_n\int_0^\infty r^{n-1}g(r)\,d\lambda(r)\ .$$

The proof of this theorem is somewhat simpler and more elementary if g is continuous, since the second step can be bypassed by using an improper Riemann integral on the right side. This is actually the only case we will use in the calculation of β_n.

V.14.2. Calculation of β_n

V.14.2.10. The trick we will use to calculate β_n is to consider the function $f_n : \mathbb{R}^n \to \mathbb{R}$ given for $\mathbf{x} = (x_1,\dots,x_n)$ by

$$f_n(\mathbf{x}) = e^{-\|\mathbf{x}\|^2} = e^{-x_1^2-\cdots-x_n^2} = e^{-x_1^2}e^{-x_2^2}\cdots e^{-x_n^2}\ .$$

This function is nonnegative, continuous, and radially symmetric, and is also a product of functions of each coordinate separately; moreover, the integrals can be computed in two separate ways.

V.14.2.11. On the one hand, from V.14.1.9. we have

$$I_n := \int_{\mathbb{R}^n} f_n\,d\lambda^{(n)} = n\beta_n\int_0^\infty r^{n-1}e^{-r^2}\,dr$$

(the last integral may be regarded as an improper Riemann integral).

V.14.2.12. On the other hand, since $f_n(x_1,\dots,x_n) = f_1(x_1)f_1(x_2)\cdots f_1(x_n)$, by Tonelli's theorem (the vector calculus version () suffices) we have

$$\begin{aligned} I_n = \int_{\mathbb{R}^n} f_n\,d\lambda^{(n)} &= \int_{\mathbb{R}} f_1(x_1)\left[\int_{\mathbb{R}} f_1(x_2)\left[\cdots\left[\int_{\mathbb{R}} f_1(x_n)\,d\lambda(x_n)\right]\cdots\right]d\lambda(x_2)\right]d\lambda(x_1)\\ &= \left[\int_{\mathbb{R}} f_1(x)\,d\lambda(x)\right]^n = (I_1)^n\ . \end{aligned}$$

V.14.2.13. The integral which can be computed by elementary means is I_2. Since $\beta_2 = \pi$ (V.14.0.1.), we have

$$I_2 = 2\pi\int_0^\infty re^{-r^2}\,dr = 2\pi\lim_{b\to+\infty}\left[-\frac{1}{2}e^{-r^2}\right]_0^b = \pi$$

and so $I_1 = \sqrt{\pi}$.

V.14.2.14. So we have

$$(I_1)^n = \pi^{n/2} = I_n = n\beta_n \int_0^\infty r^{n-1} e^{-r^2}\, dr\ .$$

Making the substitution $u = r^2$, we have

$$\pi^{n/2} = \frac{n\beta_n}{2} \int_0^\infty u^{\frac{n}{2}-1} e^{-u}\, du = \frac{n\beta_n}{2}\Gamma\left(\frac{n}{2}\right)$$

where Γ is the Gamma function (). Solving for β_n and using $\Gamma(\frac{n}{2}+1) = \frac{n}{2}\Gamma(\frac{n}{2})$, we get the final result:

V.14.2.15. THEOREM.

$$\beta_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}\ .$$

So if $n = 2m$ is even, we have $\beta_n = \frac{\pi^m}{m!}$.

V.14.2.16. Using $\beta_1 = 2$, we can obtain $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2} = \frac{1}{2}\Gamma(\frac{1}{2})$, so $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Thus, if $n = 2m+1$ is odd, we have

$$\Gamma\left(\frac{n}{2}+1\right) = \frac{1}{2}\cdot\frac{3}{2}\cdot\frac{5}{2}\cdots\frac{n}{2}\sqrt{\pi} = \frac{1\cdot 3\cdot 5\cdots n}{2^{m+1}}\sqrt{\pi}$$

so we obtain

$$\beta_n = \frac{2^{m+1}\pi^m}{1\cdot 3\cdot 5\cdots n}\ .$$

V.14.2.17. Thus $\beta_4 = \frac{\pi^2}{2}$, $\beta_5 = \frac{8\pi^2}{15}$, etc. Note that each time n increases by 2, one more factor of π appears. The power of π that occurs in the final formulas is always an integer.

V.14.2.18. We can also define β_s for any $s \in [0,\infty)$ by the formula

$$\beta_s = \frac{\pi^{s/2}}{\Gamma(\frac{s}{2}+1)}\ .$$

This is the theoretical s-dimensional "measure of the closed unit ball in $\mathbb{R}^s$." This definition will be used in the construction of Hausdorff measure.

V.14.3. Surface Area of a Sphere

V.14.3.19. The arguments of this section can be taken much farther. Fix n. We can define "spherical coordinates" (ρ, Θ) in $\mathbb{R}^n \setminus \{\mathbf{0}\}$ by setting

$$\rho(\mathbf{x}) = \|x\|, \qquad \Theta(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|}\ .$$

Then $\rho(\mathbf{x}) \in (0,\infty)$ and $\Theta(\mathbf{x}) \in S^{n-1}$, and $\Phi = (\rho, \Theta)$ is a homeomorphism from $\mathbb{R}^n \setminus \{\mathbf{0}\}$ to $(0,\infty) \times S^{n-1}$; the inverse function Φ^{-1} is given by $\Phi^{-1}(r, \mathbf{x}) = r\mathbf{x}$. Since these functions are homeomorphisms, they send Borel sets to Borel sets.

V.14.3.20. We can also transfer Lebesgue measure $\lambda^{(n)}$ to the product space $(0,\infty)\times S^{n-1}$ via Φ; call the image measure μ. Thus, if E is a Borel subset of $(0,\infty)\times S^{n-1}$, we have $\mu(E)=\lambda^{(n)}(\Phi^{-1}(E))$. Theorem V.14.1.9. can be rephrased to say that if A is a Borel subset of $(0,\infty)$, then

$$\mu(A\times S^{n-1})=n\beta_n\int_A r^{n-1}\,d\lambda(r)\ .$$

Define a measure ν on $(0,\infty)$ by $\nu(A)=\int_A r^{n-1}\,d\lambda(r)$. We will show that there is a finite Borel measure σ on S^{n-1}, with $\sigma(S^{n-1})=n\beta_n$, such that $\mu=\nu\times\sigma$.

V.14.3.21. If B is a Borel subset of S^{n-1} (in either the abstract sense regarding S^{n-1} as a topological space, or a subset of S^{n-1} which is a Borel set in $\mathbb{R}^n$; since S^{n-1} is a closed, hence Borel, subset of $\mathbb{R}^{n-1}$, the two notions are equivalent), then $\Phi^{-1}((0,1]\times B)$ is a Borel subset of $\mathbb{R}^n$. Define

$$\sigma(B)=n\lambda^{(n)}((0,1]\times B)\ .$$

It is almost immediate that σ is a Borel measure on S^{n-1}. The main result is:

V.14.3.22. THEOREM. The measures μ and $\nu\times\sigma$ coincide on the Borel sets in $(0,\infty)\times S^{n-1}$.

PROOF: In a manner almost identical to the proof of V.14.1.9., for a fixed B we have $\mu(A\times B)=\nu(A)\sigma(B)$ for every Borel subset A of $(0,\infty)$ (first show this for $A=(a,b]$ by homogeneity, and then bootstrap to all Borel sets). Thus, if $\mathcal{A}$ is the σ-algebra of Borel sets in $(0,\infty)$ and $\mathcal{B}$ the Borel sets in S^{n-1}, we have $\mu=\nu\times\sigma$ on $\mathcal{A}\times\mathcal{B}$. So by () μ and $\nu\times\sigma$ agree on $\mathcal{A}\otimes\mathcal{B}$, which is the σ-algebra of Borel sets in $(0,\infty)\times S^{n-1}$. 👍

V.14.3.23. The measure σ is called the *surface measure* of S^{n-1}. In the cases $n=2$ and $n=3$ this surface measure represents arc length and surface area respectively. This measure can be defined in several other ways: it is the restriction of $(n-1)$-dimensional Hausdorff measure () on $\mathbb{R}^n$ to S^{n-1}, it is the Riemannian measure on S^{n-1} as a smooth submanifold of $\mathbb{R}^n$ (), and it can be defined via a parametrization ().

V.14.3.24. The total surface measure $\sigma(S^{n-1})$ is just $n\beta_n$. Thus $\sigma(S^1)=2\pi$, the circumference of the unit circle, and $\sigma(S^2)=4\pi$, the surface area of the unit sphere in $\mathbb{R}^3$. By scaling, the surface measure of the sphere of radius r in $\mathbb{R}^n$ is $r^{n-1}\sigma(S^{n-1})$. This is the derivative of the function $V(r)=\beta_n r^n$ giving the Lebesgue measure of a ball in $\mathbb{R}^n$ of radius r. An interpretation of this is: if S_h is the spherical shell of inside radius 1 and outside radius $1+h$ in $\mathbb{R}^n$, we have:

$$\sigma(S^{n-1})=\lim_{h\to 0}\frac{\lambda^{(n)}(S_h)}{h}\ .$$

A similar formula holds for any Borel subset of S^{n-1}.

V.14.4. Exercises

V.14.4.1. Let $n \geq 2$. Use iterated integration to compute β_n as follows.

(a) Write $\beta_n = \displaystyle\int_{x_1^2+\cdots+x_n^2\leq 1} 1\, d(x_1,\ldots,x_n) = \int_{-1}^{1} \left[\int_{x_2^2+\cdots+x_n^2\leq 1-x_1^2} 1\, d(x_2,\ldots,x_n) \right] dx_1$.

(b) The quantity in brackets is the $(n-1)$-dimensional measure of a ball of radius $\sqrt{1-x_1^2}$ in $\mathbb{R}^{n-1}$. By scaling invariance, this integral equals $\beta_{n-1}(1-x_1^2)^{\frac{n-1}{2}}$.

(c) Obtain a recurrence formula $\beta_n = \beta_{n-1} \displaystyle\int_{-1}^{1} (1-x^2)^{\frac{n-1}{2}}\, dx = 2\beta_{n-1} \int_0^1 (1-x^2)^{\frac{n-1}{2}}\, dx$.

(d) Make the change of variables $x = \cos\theta$ to convert the recurrence formula to

$$\beta_n = \beta_{n-1} \int_0^{\pi} \sin^n\theta\, d\theta = 2\beta_{n-1} \int_0^{\pi/2} \sin^n\theta\, d\theta \ .$$

(e) Evaluate the integral using () and obtain the formula of V.14.2.15. using $\beta_1 = 2$.

V.14.4.2. Let $n \geq 3$. Compute the n-dimensional Lebesgue measure of the ball $B_r^{(n)}$ of radius r around the origin in $\mathbb{R}^n$ as follows.

(a) Introduce "hyperspherical" coordinates $(\rho, \theta_1, \ldots, \theta_{n-1})$ in $\mathbb{R}^n$, where

$$\begin{aligned} x_1 &= \rho\cos\theta_1 \\ x_2 &= \rho\sin\theta_1\cos\theta_2 \\ &\cdots \\ x_{n-1} &= \rho\sin\theta_1\cdots\sin\theta_{n-2}\cos\theta_{n-1} \\ x_n &= \rho\sin\theta_1\cdots\sin\theta_{n-2}\sin\theta_{n-1} \end{aligned}$$

(for $n = 3$, the usual spherical coordinates (ρ, ϕ, θ) have $\theta_1 = \phi$, $\theta_2 = \theta$, and the coordinates come out in the order (z, x, y)).

(b) Show that the n-dimensional volume $\lambda^{(n)}(B_r^{(n)})$ is given by

$$\begin{aligned} &\int_0^r \int_0^{\pi} \cdots \int_0^{\pi} \int_0^{2\pi} \rho^{n-1} \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-2}\, d\theta_1\, d\theta_2 \cdots d\theta_{n-1}\, d\rho \\ &= \left(\int_0^r \rho^{n-1}\, d\rho \right) \left(\int_0^{2\pi} d\theta_{n-1} \right) \prod_{k=1}^{n-2} \left(\int_0^{\pi} \sin^k\theta_{n-k-1}\, d\theta_{n-k-1} \right) \\ &= \frac{2\pi}{n} r^n \prod_{k=1}^{n-2} \left(\int_0^{\pi} \sin^k\theta\, d\theta \right) \ . \end{aligned}$$

(This formula also holds with a simplified argument if $n = 2$ if the last product is taken to be the "empty product" 1.)

(c) Conclude that if $n \geq 2$, then

$$\beta_n = \frac{2\pi}{n} \prod_{k=1}^{n-2} \left(\int_0^{\pi} \sin^k \theta \, d\theta \right)$$

(with the last product equal to 1 if $n = 2$). Show that this formula coincides with the formula of V.14.2.15. (cf. ()).

V.14.4.3. (a) Find for which $n \in \mathbb{N}$ we have β_n maximum.

(b) Find for which $s \in [0, \infty)$ we have β_s maximum.

(c) Show that $\lim_{s \to +\infty} \beta_s = 0$.

V.14.4.4. [BB92] Show that

$$\left(10^{-5} \sum_{-\infty}^{\infty} e^{-\frac{n^2}{10^{10}}} \right)^2$$

agrees with π to at least 42 billion decimal places, but the numbers are not equal. Which is larger? [Show that

$$\pi = \left(10^{-5} \int_{-\infty}^{\infty} e^{-\frac{x^2}{10^{10}}} \, dx \right)^2$$

(cf. V.14.2.13.).]

V.15. Infinite Series of Functions

V.15.1. Infinite Series of Functions

Pointwise, Uniform, and Absolute Convergence of Series

Normal Convergence of Series and the Weierstrass M-Test

The most useful basic result about convergence of series of functions is the following test, due to Weierstrass, a version of IV.2.5.2.:

V.15.1.1. Theorem. [Weierstrass M-Test] Let

$$\sum_{k=1}^{\infty} f_k$$

be an infinite series of real- or complex-valued functions on a set X. Suppose there is a sequence (M_k) of nonnegative real numbers such that $|f_k(x)| \leq M_k$ for all k and all $x \in X$, and that

$$\sum_{k=1}^{\infty} M_k$$

converges. Then the series of functions converges absolutely and uniformly on X.

Proof: Let

$$s_n = \sum_{k=1}^{n} f_k$$

be the n'th partial sum of the function series. Fix $\epsilon > 0$. There is then an $N \in \mathbb{N}$ such that

$$\sum_{k=n+1}^{m} M_k < \epsilon$$

for $N \leq n < m$. Then, by the triangle inequality, we have, for $N \leq n < m$,

$$|s_m(x) - s_n(x)| = \left| \sum_{k=n+1}^{m} f_k(x) \right| \leq \sum_{k=n+1}^{m} |f_k(x)| \leq \sum_{k=n+1}^{m} M_k < \epsilon$$

for all $x \in X$. Thus the sequence (s_n) is a uniform Cauchy sequence, hence converges () uniformly on X. And, for each $x \in X$,

$$\sum_{k=1}^{\infty} |f_k(x)|$$

converges by comparison with $\sum_{k=1}^{\infty} M_k$, so $\sum_{k=1}^{\infty} f_k(x)$ converges absolutely. 👍

There is an alternate phrasing of this result.

V.15.1.2. Definition. Let X be a set, and f a bounded real-or-complex-valued function on X. Set

$$\|f\|_\infty = \sup_{x\in X} |f(x)| \ .$$

Then $\|f\|$ is called the *supremum norm* of f on X.

This norm is indeed a norm () on the set of bounded functions on X, and is complete ().

V.15.1.3. Corollary. Let X be a set, and (f_k) a sequence of bounded real- or complex-valued functions on X. If

$$\sum_{k=1}^{\infty} \|f_k\|_\infty$$

converges, then the infinite series

$$\sum_{k=1}^{\infty} f_k$$

converges absolutely and uniformly on X.

V.15.1.4. An infinite series of bounded functions for which

$$\sum_{k=1}^{\infty} \|f_k\|_\infty$$

converges is said to *converge normally*. By V.15.1.3. an infinite series of functions which converges normally converges absolutely and uniformly. Note, though, that normal convergence of an infinite series is not the same thing as normal (i.e. u.c.) convergence of the sequence of partial sums in the sense of IV.3.2.2., an unfortunate ambiguity in established terminology. Normal convergence generalizes to Banach spaces ().

Term-by-Term Differentiation

V.15.1.5. It is important to understand that even a uniformly convergent infinite series of differentiable functions cannot be differentiated term-by-term in general. For a simple example, the infinite series

$$\sum_{k=1}^{\infty} 2^{-k} \sin 2^k x$$

converges uniformly on $\mathbb{R}$ to a (continuous) function f. But the series of term-by-term derivatives

$$\sum_{k=1}^{\infty} \cos 2^k x$$

does not even converge pointwise at 0 or most other points. (In fact, f is not even differentiable at most places.) See I.6.1.18. for a heuristic explanation.

We do have the following series version of V.8.5.5.:

V.15.1.6. THEOREM. Let (F_n) and (f_n) be sequences of functions on an interval I, and A a countable subset of I. Suppose

(a) Each F_n is continuous on I.

(b) For every $x \in I \setminus A$, each F_n is differentiable at x and $F_n'(x) = f_n(x)$.

(c) The series $\sum_{n=1}^{\infty} f_n$ converges u.c. on I to a function f.

(d) There is an $x_0 \in I$ such that the series $\sum_{n=1}^{\infty} F_n(x_0)$ converges.

Then:

(i) The series $\sum_{n=1}^{\infty} F_n$ converges u.c. on I to a (continuous) function F.

(ii) For each $x \in I \setminus A$, F is differentiable at x and $F'(x) = f(x)$.

As usual, the proof is just an application of V.8.5.5. to the sequence of partial sums.

V.15.1.7. The important special case where $A = \emptyset$ and each f_n is continuous follows from IX.2.9.15..

V.15.2. Exercises

V.15.2.1. (a) Prove the *sophomore's dream*:

$$\int_0^1 x^{-x}\,dx = \int_0^1 \frac{1}{x^x}\,dx = \sum_{k=1}^{\infty} \frac{1}{k^k} \quad \text{and} \quad \int_0^1 x^x\,dx = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k^k}\ .$$

[Write $x^x = e^{x \log x}$, expand in an infinite series using the Maclaurin series for e^x, and integrate term-by-term.]

The name "sophomore's dream" was coined in [BBG04].

(b) Another sophomore's dream:

$$\int_0^{\infty} \frac{x}{x^3+1}\,dx = \int_0^{\infty} \frac{1}{x^3+1}\,dx\ .$$

V.15.2.2. Does the iterated integral

$$\int_0^8 \int_0^{\sqrt{4-x}} 2y\,dy\,dx$$

exist? If so, what is its value? Interpret the integral geometrically.

V.15.2.3. Define

$$f(x) = \sum_{k=0}^{\infty} \frac{x^2}{(1+x^2)^k} \ .$$

(a) Show that the series converges (absolutely) for all $x \in \mathbb{R}$. [For each x, it is a multiple of a geometric series.]

(b) Show that the series does not converge uniformly on any interval containing 0.

(c) Show that $f(0) = 0$ and $f(x) = 1 + x^2$ for $x \neq 0$. Thus f is not continuous at 0.

V.15.3. Power Series

Power series are probably the most important kinds of infinite series of functions (although trigonometric series are also very important). A great many of the most important functions can be represented as power series, and power series have some special properties which make them relatively easy to use and work with.

> "Mathematicians of the 18th century mostly regarded power series as infinite polynomials. To a large extent, polynomials and power series behave in calculations in the same way. It would have appeared artificial to regard them as essentially different entities."
>
> *Hans Niels Jahnke*[16]

This attitude is largely justifiable (although the justifications are nontrivial!), *provided* one carefully stays within the open interval of convergence (or disk in the complex case), and persists to this day:

> "A power series is, after all, a rather long polynomial."
>
> *Forman S. Acton*[17]

V.15.3.1. DEFINITION. Let x_0 be a real number. A *(real) power series centered at* x_0 is an expression of the form

$$\sum_{k=0}^{\infty} a_k (x - x_0)^k$$

where the a_k are real numbers. Complex power series can be defined similarly: if z_0 is a complex number, a *complex power series centered at* z_0 is an expression of the form

$$\sum_{k=0}^{\infty} a_k (z - z_0)^k$$

where the a_k are complex numbers.

While we will (except in Chapter XII) generally only consider real power series, the theory of power series is most naturally done for complex power series as part of complex analysis, and there are places in the real theory where the complex point of view is very helpful. Any real power series can, of course, be regarded as a complex power series by replacing x by z.

V.15.3.2. A power series is regarded as an infinite series of functions. It is conventional when working with power series to regard $(x - x_0)^0 = 1$ no matter how $x - x_0$ is interpreted; thus the $k = 0$ term of the power series is just the constant function a_0. The same convention holds for complex power series.

V.15.3.3. A power series is, at first, only a formal expression, but it can be regarded as defining a function f of x, whose domain is the set of all x for which the series converges. Because of our convention, x_0 is always in the domain of the defined function f, with $f(x_0) = a_0$. It can happen that the domain of f consists only of $\{x_0\}$ (such power series are not very interesting as functions).

The most important basic property of power series is V.15.3.5., which says that the domain of the defined function is an interval, or degenerate interval, centered at x_0.

[16] [Jah03, p. 108].
[17] [Act90, p. 13-14].

V.15.3.4. DEFINITION. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, and define

$$L = \limsup_{k\to\infty} |a_k|^{1/k} = \limsup_{k\to\infty} \sqrt[k]{|a_k|}$$

and let $R = \frac{1}{L}$ if $0 < L < +\infty$, $R = 0$ if $L = +\infty$, and $R = +\infty$ if $L = 0$. R is called the *radius of convergence* of the power series.

An identical definition can be made for complex power series. Note that the radius of convergence of a power series depends only on the coefficients a_k, not on the center x_0, of the power series. The slower the coefficients grow in size, or the faster they approach zero, the larger the radius of convergence.

The term "radius of convergence" is justified by the next result:

V.15.3.5. THEOREM. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with radius of convergence R. Then

(i) If $x \in \mathbb{R}$ and $|x-x_0| < R$, then the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ converges absolutely.

(ii) If $x \in \mathbb{R}$ and $|x-x_0| > R$, then the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ diverges.

(iii) If $r \in \mathbb{R}$ and $0 < r < R$, then the series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ converges uniformly on the interval $[x_0-r, x_0+r]$.

(Of course, (i) and (iii) are vacuous if $R = 0$ and (ii) is vacuous if $R = +\infty$.)

PROOF: In cases (i) and (ii), apply the Root Test () to the series $\sum_{k=0}^{\infty} |a_k(x-x_0)^k|$. We have

$$\limsup |a_k(x-x_0)^k|^{1/k} = |x-x_0| \limsup |a_k|^{1/k} = \frac{|x-x_0|}{R}$$

(it is $+\infty$ if $x \neq x_0$ and $R = 0$, and 0 if $R = +\infty$ or $x = x_0$). Thus the series converges absolutely if $|x-x_0| < R$ or $x = x_0$ and diverges since the terms do not go to zero if $|x-x_0| > R$.
For (iii), use the Weierstrass M-test (). Set $M_k = |a_k|r^k$. Then $|a_k(x-x_0)^k| \leq M_k$ for all k and for all x, $|x-x_0| \leq r$. We have

$$\limsup M_k^{1/k} = r \limsup |a_k|^{1/k} = \frac{r}{R} < 1$$

if $0 < R < \infty$, and $\limsup M_k^{1/k} = 0$ if $R = +\infty$. Thus $\sum_{k=0}^{\infty} M_k$ converges by the Root Test. ☝

The statement and proof of the theorem work identically for complex power series (the uniform convergence in (iii) is on the closed disk of radius r centered at x_0).

V.15.3.6. Corollary. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with radius of convergence $R>0$. Then the series converges u.c. on the open interval (x_0-R, x_0+R), and the sum is continuous on this interval.

Proof: If $[a,b] \subseteq (x_0-R, x_0+R)$, then $[a,b] \subseteq [x_0-r, x_0+r]$ for some $r<R$. Continuity of the sum follows from ().

The sum is actually far more than just continuous on this open interval ().

V.15.3.7. Thus the set on which a power series centered at x_0 converges is an interval centered at x_0, possibly of length zero, i.e. the degenerate interval $\{x_0\}$, or of infinite length, i.e. all of $\mathbb{R}$, possibly including one or both endpoints, called the *interval of convergence* of the power series. (If $0<R<\infty$, the theorem has nothing to say about behavior at the points where $x-x_0=\pm R$, the endpoints of the interval of convergence; endpoint behavior will be discussed below.) Similarly, the set on which a complex power series centered at z_0 converges is a disk in $\mathbb{C}$ centered at z_0, possibly degenerate or all of $\mathbb{C}$, and possibly including some or all points on the boundary circle of the disk.

V.15.3.8. The radius of convergence of a power series $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ can always be computed as

$$R = \frac{1}{\limsup |a_k|^{1/k}}$$

(with the usual conventions if the lim sup is 0 or $+\infty$), but it is not always easy to evaluate this expression. Sometimes the radius of convergence can be computed more simply by the following:

V.15.3.9. Proposition. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with all a_k nonzero. If

$$\lim_{k\to\infty} \left| \frac{a_k}{a_{k+1}} \right|$$

exists (in the usual or extended sense), it equals the radius of convergence of the power series.

Proof: Let S be this limit. Apply the Ratio Test () to show that $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ converges absolutely if $|x-x_0|<S$ and diverges if $|x-x_0|>S$. Thus S must be exactly the radius of convergence.

Note that in the limit of V.15.3.9., the a_k and a_{k+1} are reversed from the way they appear in the Ratio Test ().

Another convenient characterization of the radius of convergence is the following:

V.15.3.10. PROPOSITION. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series. The power series has radius of convergence $R>0$ if and only if, for all r, $0<r<R$, there is a constant M such that $|a_k| \le \frac{M}{r^k}$ for all k, and there is no such M if $r>R$.

PROOF: Suppose the series has radius of convergence $R>0$, and $0<r<R$. Then

$$\limsup |a_k|^{1/k} < \frac{1}{r}$$

so there is a k_0 such that $|a_k| < \frac{1}{r^k}$ for all $k \ge k_0$. There is an $M \ge 1$ such that $|a_k| \le \frac{M}{r^k}$ for $0 \le k < k_0$. This M works for all k.

Conversely, if the condition holds for all r with $0<r<R$ and no $r>R$, we have, for each such r with $0<r<R$,

$$\limsup |a_k|^{1/k} \le \limsup \frac{M^{1/k}}{r} = \frac{1}{r}$$

for some constant M, so the radius of convergence R' is $\ge R$. If $R'>R$, choose r so that $R<r<R'$; then the condition holds for r by the first half of the proof, a contradiction. 👍

V.15.3.11. EXAMPLES. (i) Let $c \in \mathbb{R}$, $c \ne 0$. The power series $\sum_{k=0}^{\infty} c^k(x-x_0)^k$ has radius of convergence $\frac{1}{|c|}$, and interval of convergence $\left(x_0 - \frac{1}{|c|}, x_0 + \frac{1}{|c|}\right)$.

(ii) The power series $\sum_{k=0}^{\infty} \frac{1}{k!}(x-x_0)^k$ has radius of convergence $+\infty$ and interval of convergence $\mathbb{R}$.

(iii) The power series $\sum_{k=0}^{\infty} k!(x-x_0)^k$ has radius of convergence 0 and (degenerate) interval of convergence $\{x_0\}$.

V.15.3.12. The convergence behavior of an infinite series does not depend on x_0, and if x_0 is changed, leaving the a_k the same, the function defined is simply changed by a translation. In particular, if f is the function defined by $\sum_{k=0}^{\infty} a_k x^k$ on its interval of convergence I, then the function defined by $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ is simply the function $g(x) = f(x-x_0)$ on $x_0 + I$. Thus in analyzing behavior of functions defined by power series, we generally may (and often will) take $x_0 = 0$ to simplify notation.

Endpoint Behavior

Almost any kind of behavior can occur at endpoints of the interval of convergence, as the following examples show.

V.15.3.13. EXAMPLES. (i) The power series $\sum_{k=0}^{\infty} x^k$ has radius of convergence 1. The series diverges if $x = \pm 1$ since the terms do not go to zero, so the interval of convergence is $(-1, 1)$.

(ii) The power series $\sum_{k=1}^{\infty} \frac{1}{k} x^k$ (a power series with constant term 0) has radius of convergence 1. The series diverges if $x = 1$ (harmonic series), but converges conditionally if $x = -1$, so the interval of convergence is $[-1, 1)$. The power series $\sum_{k=1}^{\infty} \frac{(-1)^k}{k} x^k$ has interval of convergence $(-1, 1]$.

(iii) The power series $\sum_{k=1}^{\infty} \frac{1}{k^2} x^k$ has radius of convergence 1. The series converges absolutely if $x = \pm 1$, so the interval of convergence is $[-1, 1]$.

(iv) The series $\sum_{k=1}^{\infty} \frac{(-1)^k}{k} x^{2k}$ (this is a power series with odd coefficients 0) has radius of convergence 1. The series converges conditionally if $x = \pm 1$, so the interval of convergence is $[-1, 1]$.

V.15.3.14. If a power series converges absolutely at one endpoint, it converges absolutely at the other endpoint too, since the absolute series $\sum_{k=0}^{\infty} |a_k (x - x_0)^k|$ is the same at both endpoints. This is the only restriction on endpoint behavior for real infinite series.

V.15.3.15. Convergence behavior of a complex series on the boundary circle of the disk of convergence is much more subtle. The series can diverge everywhere on this circle (e.g. $\sum_{k=0}^{\infty} (z - z_0)^k$) or it can converge absolutely at all points of the circle (e.g. $\sum_{k=1}^{\infty} \frac{1}{k^2} (z - z_0)^k$), or it can converge conditionally on some subset S of the boundary circle. But S cannot be an arbitrary subset of the circle (it must be an $F_{\sigma\delta}$ (), and not even all of these occur; see [Kör83]). The study of boundary behavior of analytic functions is a big topic in complex analysis, with important applications.

We will primarily be interested in using power series on open intervals. But we note one important result about the behavior of a power series near an endpoint or boundary point where it converges:

V.15.3.16. THEOREM. [ABEL'S THEOREM] Let $\sum_{k=0}^{\infty} a_k (x - x_0)^k$ be a power series. If $[a, b]$ is a closed bounded interval contained in the interval of convergence of the series, then the series converges uniformly on $[a, b]$. (In other words, the series converges u.c. on its entire interval of convergence.)

If $[a, b]$ is contained in the interior of the interval of convergence, the result follows immediately from V.15.3.6., so Abel's Theorem only gives additional information in the case where $[a, b]$ contains one or both endpoints of the interval of convergence, and this is the case it is designed for.

V.15.3.17. COROLLARY. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with interval of convergence I. Then the function on I defined by the series is continuous on I, including at any endpoints of I.

PROOF: Combine V.15.3.16. with (). 👍

V.15.3.18. Abel's Theorem also has a complex version, which we do not state precisely. But even the real version is enough to almost immediately show that if $\sum_{k=0}^{\infty} a_k(z-z_0)^k$ is a complex power series, and it converges at some z_1 on the boundary circle of the disk of convergence, then the value at z_1 is the radial limit of the value at z as $z \to z_1$ along the ray from z_0 to z_1. See Exercise ().

We now give the proof of Abel's Theorem.

PROOF: By applying V.15.3.6. to expand the interval if necessary, we may assume $a \leq x_0 \leq b$. We prove that the series converges uniformly on $[x_0, b]$; the proof that it converges uniformly on $[a, x_0]$ is essentially identical.

Let $\epsilon > 0$. Since $\sum_{k=1}^{\infty} a_k(b-x_0)^k$ converges, by () there is an N such that

$$\left| \sum_{k=n+1}^{n+m} a_k(b-x_0)^k \right| < \epsilon$$

whenever $n \geq N$ and $m \in \mathbb{N}$. If $x \in [x_0, b]$, set

$$c_k = \left(\frac{x-x_0}{b-x_0} \right)^k .$$

Then by Abel's Partial Summation Formula (IV.2.7.5.) with $r = n$ we have, for any $n \geq N$ and $m \in \mathbb{N}$,

$$\sum_{k=n+1}^{n+m} a_k(x-x_0)^k = \sum_{k=n+1}^{n+m} c_k a_k(b-x_0)^k = \sum_{k=n+1}^{n+m} (c_k - c_{k+1}) s_{k,n} + c_{n+m+1} s_{n+m,n}$$

where $s_{k,n} = \sum_{j=n+1}^{k} a_j(b-x_0)^j$. Since $|s_{k,n}| < \epsilon$ for all $k \geq n$, and $0 \leq c_{k+1} \leq c_k \leq 1$ for all k, we have

$$\left| \sum_{k=n+1}^{n+m} a_k(x-x_0)^k \right| \leq \sum_{k=n+1}^{n+m} (c_k - c_{k+1})|s_{k,n}| + c_{n+m+1}|s_{n+m,n}|$$

$$\leq \epsilon \left[\sum_{k=n+1}^{n+m} (c_k - c_{k+1}) + c_{n+m+1} \right] = (c_{n+1} - c_{n+m+1})\epsilon \leq \epsilon$$

and so the series converges uniformly on $[x_0, b]$. 👍

Term-by-Term Integration

Because of the u.c. convergence of a power series on its interval of convergence I, we can apply () to integrate a power series term-by-term on any closed bounded interval contained in I:

V.15.3.19. COROLLARY. Let $\sum_{k=0}^{\infty} c_k(x-x_0)^k$ be a power series with interval of convergence I, and $[a,b]$ a closed bounded subinterval of I. Then the function f defined by the power series is integrable on $[a,b]$, and

$$\int_a^b f(x)\,dx = \int_a^b \left[\sum_{k=0}^{\infty} c_k(x-x_0)^k\right] dx = \sum_{k=0}^{\infty}\left[\int_a^b c_k(x-x_0)^k\,dx\right]$$

$$= \sum_{k=0}^{\infty} \frac{c_k}{k+1}\left[(b-x_0)^{k+1} - (a-x_0)^{k+1}\right] = \sum_{k=0}^{\infty} \frac{c_k}{k+1}(b-x_0)^{k+1} - \sum_{k=0}^{\infty} \frac{c_k}{k+1}(a-x_0)^{k+1}\ .$$

In particular, the series on the last line converge. This result is a consequence of V.15.3.6. if $[a,b]$ is contained in the open interval of convergence, but requires Abel's Theorem if a or b is an endpoint of the interval of convergence. If $[a,b]$ is contained in the open interval of convergence, the series in the last expression will automatically converge absolutely by comparison to the power series evaluated at a and b (which converge absolutely by V.15.3.5.(i)). If a or b is an endpoint of the interval of convergence, the original series evaluated at a or b may not converge absolutely, but the next result shows that the integrated series still converge in this case.

V.15.3.20. PROPOSITION. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series. If this power series converges for $x = x_1$, then the integrated series $\sum_{k=0}^{\infty} \frac{a_k}{k+1}(x-x_0)^{k+1}$ also converges for $x = x_1$.

PROOF: Apply Abel's Test (IV.2.7.7.) with $c_k = \frac{x_1-x_0}{k+1}$. 👍

Term-by-Term Differentiation

Term-by-term differentiation of a power series is a dicier matter. One of the beautiful and crucial properties of power series is that they can in fact be differentiated term-by-term, in contrast to more general series of functions.

The key fact making this possible is the next result. Note that if $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ is a power series, the series $\sum_{k=1}^{\infty} ka_k(x-x_0)^{k-1}$ of term-by-term derivatives is also a power series.

V.15.3.21. PROPOSITION. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with radius of convergence R. Then the radius of convergence of the series $\sum_{k=1}^{\infty} ka_k(x-x_0)^{k-1}$ of term-by-term derivatives is also exactly R.

PROOF: First note that the series $\sum_{k=1}^{\infty} ka_k(x-x_0)^{k-1}$ and $\sum_{k=1}^{\infty} ka_k(x-x_0)^k$ converge for exactly the same x's: both converge for $x = x_0$, and for $x \neq x_0$ each series is a constant scalar multiple of the other. Thus it suffices to show that the radius of convergence R' of the series $\sum_{k=1}^{\infty} ka_k(x-x_0)^k$ is equal to R.

Let $L = \limsup |a_k|^{1/k}$. Then $R = \frac{1}{L}$ (with the usual conventions if L is 0 or ∞). Similarly, $R' = \frac{1}{L'}$, where $L' = \limsup |ka_k|^{1/k}$. So it suffices to show that $L' = L$. But $\lim_{k\to\infty} k^{1/k} = 1$ (V.9.3.3.(iii)), so by ()

$$L' = \limsup_{k\to\infty} |ka_k|^{1/k} = \left(\lim_{k\to\infty} k^{1/k}\right)\left(\limsup_{k\to\infty} |a_k|^{1/k}\right) = 1 \cdot L = L \ .$$



This result and proof are also valid for complex power series.

V.15.3.22. Note that this result does not say that the *interval* of convergence of the derived series is the same as the interval of convergence of the original series; in fact, endpoint behavior can be different. In examples V.15.3.13.(ii)–(iv), the interval of convergence of the derived series is strictly smaller than that of the original series; it cannot be larger (cf. V.15.3.20.).

We now apply () to get the main result:

V.15.3.23. THEOREM. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with radius of convergence $R > 0$. Then the function f defined by the power series is differentiable on $(x_0 - R, x_0 + R)$, and

$$f'(x) = \sum_{k=1}^{\infty} ka_k(x-x_0)^{k-1}$$

for all $x \in (x_0 - R, x_0 + R)$.

V.15.3.24. But why stop there? The power series representing f' can also be differentiated term-by-term, and the theorem applies to show that f' is also differentiable on $(x_0 - R, x_0 + R)$, i.e. f has a second derivative on this interval given by taking term-by-term derivatives twice. This argument can be iterated arbitrarily many times. Thus we obtain:

V.15.3.25. Corollary. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series, with radius of convergence $R>0$. Then the function f defined by the power series is C^∞ on (x_0-R, x_0+R), and

$$f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1)a_k(x-x_0)^{k-n} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!}a_k(x-x_0)^{k-n}$$

for every n and every $x \in (x_0-R, x_0+R)$.

Note that V.15.3.23. and hence V.15.3.26. apply also on any open subinterval of the interval of convergence. Thus we can rephrase:

V.15.3.26. Corollary. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series which converges to a function f on an open interval I. Then f is C^∞ on I, and

$$f^{(n)}(x) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1)a_k(x-x_0)^{k-n} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!}a_k(x-x_0)^{k-n} = \sum_{k=0}^{\infty} \frac{(k+n)!}{k!}a_{k+n}(x-x_0)^k$$

for every n and every $x \in I$ (in particular, this series converges for all n and all $x \in I$).

Uniqueness of Power Series Representations

V.15.3.27. V.15.3.26. has another important consequence: if $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ is a power series with positive radius of convergence, then the coefficients a_k can be recovered from the function f defined by the power series. We have that f is infinitely differentiable in an interval around x_0, and $f^{(n)}$ is given by the n'th derived power series. In particular, $f^{(n)}(x_0)$ is the constant term of the n'th derived power series, which is $n!a_n$. Thus we obtain:

V.15.3.28. Corollary. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series with positive radius of convergence, and f the function defined by the series on its open interval of convergence. Then

$$a_n = \frac{f^{(n)}(x_0)}{n!}$$

for each n. In particular, the given power series is the only possible power series centered at x_0 that can represent f on an open interval containing x_0. Thus a function can be represented near x_0 by at most one power series centered at x_0.

V.15.3.29. DEFINITION. Let f be C^∞ on an interval around $x_0 \in \mathbb{R}$. The *Taylor series* for f around x_0 is the infinite series

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \ .$$

Note that the partial sums of the Taylor series for f around x_0 are exactly the Taylor polynomials (V.10.1.6.) of f around x_0. The corollary can be rephrased:

V.15.3.30. COROLLARY. Let f be C^∞ on an interval around $x_0 \in \mathbb{R}$. Then the only power series centered at x_0 which can converge to f in any open neighborhood of x_0 is the Taylor series for f around x_0.

V.15.3.31. Note two important things about the corollary:



(i) The result only applies to power series centered at x_0. It does not imply that f cannot be represented near x_0 by a different power series centered at some $x_1 \neq x_0$. In fact, if f can be represented in some interval by a power series centered at x_0, there will be such a power series centered at any x_1 sufficiently close to x_0 ().

(ii) The result says nothing about *whether* a function defined near x_0 can in fact be represented by a power series centered at x_0, only that *if* it can, the series is unique and is the Taylor series for f around x_0. An obvious necessary condition for f to be representable is that it be C^∞ in an interval around x_0, but this is far from sufficient. The question of which functions can be represented by power series is a substantial one which will be discussed in the next section.

V.16. Analytic Functions

Functions which can be locally represented by their Taylor series, called *analytic functions*, are particularly well-behaved. It turns out that most standard functions, and many of the functions naturally encountered in mathematics, are analytic.

V.16.1. Definitions and Basic Properties

V.16.1.1. DEFINITION. (i) Let U be an open set in $\mathbb{R}$. A function $f : U \to \mathbb{R}$ is *real analytic* on U if for every $x_0 \in U$, f is represented in some neighborhood of x_0 by a (real) power series centered at x_0.

(ii) Let V be an open set in $\mathbb{C}$. A function $f : V \to \mathbb{C}$ is *complex analytic* on V if for every $z_0 \in V$, f is represented in some neighborhood of z_0 by a complex power series centered at z_0.

In mathematics it is most common to use the term *analytic* to mean "complex analytic." However, in this book, except in Chapter XII, it will normally mean "real analytic" or occasionally "real or complex analytic."

V.16.1.2. Some remarks on the definition:

(i) The phrase "represented in some neighborhood of x_0 by a (real) power series centered at x_0" means that there is some $\epsilon > 0$ and some (real) power series centered at x_0 which converges to $f(x)$ for all $x \in (x_0 - \epsilon, x_0 + \epsilon)$. The meaning in (ii) is analogous.

(ii) If f is analytic (in either the real or complex sense), then f is C^∞ and the (unique) power series centered at x_0 representing f in a neighborhood of x_0 is the Taylor series of f around x_0 (V.15.3.29.).

(iii) Since the partial sums of the Taylor series for f around x_0 are the Taylor polynomials (V.10.1.6.) of f around x_0, f is analytic on U if and only if, for each $x_0 \in U$, the Taylor polynomials for f around x_0 converge to f in a neighborhood of x_0, i.e. the remainder terms r_n converge to zero on a neighborhood of x_0 (cf. V.10.1.8.).

(iv) It is important to note that if f is analytic on U and $x_0 \in U$, then the Taylor series of f around x_0 does not converge to f on all of U in general; it only converges to f in *some* open interval around x_0. Thus the property of being analytic on U is strictly a local property. If $f : U \to \mathbb{R}$ is a function, and for every $x \in U$ there is a neighborhood of x in U on which f is analytic, then f is analytic on U. In particular, if f is analytic on every open subinterval of U, then f is analytic on U; thus for convenience we often just work with open intervals when dealing with analytic functions.

A similar comment applies in the complex analytic case. See ().

V.16.1.3. One of the important theorems of complex analysis is a strong form of the converse to V.16.1.2.(ii): if V is an open subset of $\mathbb{C}$ and $f : V \to \mathbb{C}$ is differentiable in the complex sense at every point of V, then f is complex analytic on V. There is nothing resembling an analog of this result for real analytic functions, and this is one of the most fundamental differences in nature between real analysis and complex analysis. More specifically:

> If f is differentiable on I, then f' need not even be continuous on I (), much less differentiable. In particular, a differentiable real function is "usually" not even C^1, much less C^∞.
>
> Even most C^∞ functions are not real analytic; see section () for details.

V.16.1.4. **Notation:** The set of real analytic functions on an open set U is denoted $C^\omega(U)$.

V.16.1.5. EXAMPLES. (i) Let f be a polynomial of degree n. Then, for any $x_0 \in \mathbb{R}$, by Taylor's Theorem the n'th order Taylor polynomial for f around x_0 is f itself, i.e. the Taylor series for f around x_0 (which has only finitely many nonzero terms) represents f on all of $\mathbb{R}$. Thus f is analytic on $\mathbb{R}$.

(ii) Let $f(x) = \sin x$. Then f is C^∞ on $\mathbb{R}$, and $|f^{(k)}(x)| \leq 1$ for all k and x. Fix $x_0 \in \mathbb{R}$, and let p_n be the n'th order Taylor polynomial for f around x_0, and $r_n = f - p_n$. Then, for any $x \in \mathbb{R}$, we have

$$r_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}$$

for some c between x_0 and x, and thus

$$|r_n(x)| \leq \frac{|x - x_0|^{n+1}}{(n+1)!} \to 0$$

as $n \to \infty$, so the Taylor series for f around x_0 converges to $f(x)$ for all x and f is analytic on $\mathbb{R}$. An identical argument shows that $g(x) = \cos x$ is analytic on $\mathbb{R}$.

(iii) Let $f(x) = e^x$. Then f is C^∞ on $\mathbb{R}$, and $f^{(k)}(x) = e^x$ for all k and x. Fix $x_0 \in \mathbb{R}$ and $s > 0$, and let $m = e^{x_0+s}$. Then $|f^{(k)}(x)| < m$ if $|x - x_0| < s$. Let p_n be the n'th order Taylor polynomial for f around x_0, and $r_n = f - p_n$. Then, for any x with $|x - x_0| < s$, we have

$$r_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}$$

for some c between x_0 and x, and thus

$$|r_n(x)| \leq \frac{m|x - x_0|^{n+1}}{(n+1)!} \to 0$$

as $n \to \infty$, so the Taylor series for f around x_0 converges to $f(x)$ for all x, $|x - x_0| < s$. Since s is arbitrary, the Taylor series converges to $f(x)$ for all $x \in \mathbb{R}$. Thus f is analytic on $\mathbb{R}$.

Analyticity of Functions Defined by Power Series

The definition of an analytic function is hard to check directly: to show that f is analytic on U, one must show for *every* $x_0 \in U$ that the Taylor series for f around x_0 converges to f in an interval around x_0, an (uncountably) infinite number of conditions since there is a different Taylor series at each point. The following theorem is one of the most important facts about analytic functions, and saves an enormous amount of work in checking that specific functions are analytic since only one x_0 must be considered.

V.16.1.6. THEOREM. Let $\sum_{k=0}^{\infty} a_k(x - x_0)^k$ be a power series with radius of convergence $R > 0$, and let f be the function defined on $I = (x_0 - R, x_0 + R)$ defined by the power series. Then f is analytic on I.

In fact, the following more precise result holds:

V.16.1.7. THEOREM. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series with radius of convergence $R > 0$, and let f be the function defined on $I = (x_0 - R, x_0 + R)$ defined by the power series. Let $x_1 \in I$, and set $S = R - |x_1 - x_0|$. Then the Taylor series for f around x_1 has radius of convergence at least S and converges to f on $(x_1 - S, x_1 + S)$.

See Figure (). Note that f is C^∞ on I (), so the Taylor series for f around x_1 is defined.

To prove these theorems, we first need a lemma which will also be used later.

V.16.1.8. LEMMA. Let $\sum_{k=0}^{\infty} a_k(x-x_0)^k$ be a power series with radius of convergence $R > 0$, and let f be the function defined on $I = (x_0 - R, x_0 + R)$ defined by the power series. Fix r, $0 < r < R$, and s, $0 < s < R - r$. Then there is a constant m such that

$$|f^{(k)}(x)| \leq \frac{m \cdot k!}{s^k}$$

for all $k \geq 0$ and all x with $|x - x_0| \leq r$.

PROOF: First note that the power series $\sum_{j=0}^{\infty} x^j$ converges to $f(x) = \frac{1}{1-x}$ on $(-1, 1)$, and thus the k'th derived series

$$\sum_{j=k}^{\infty} j(j-1)\cdots(j-k)x^{j-k}$$

converges on $(-1, 1)$ to $f^{(k)}(x) = \frac{k!}{(1-x)^{k+1}}$ for any k ().

Let r and s be as in the statement. Since $r + s < R$, by () there is an M such that $|a_k| \leq \frac{M}{(r+s)^k}$ for all k. If $|x - x_0| \leq r$, then we have, for any k,

$$f^{(k)}(x) = \sum_{j=k}^{\infty} j(j-1)\cdots(j-k)a_j(x-x_0)^{j-k}$$

$$|f^{(k)}(x)| \leq \sum_{j=k}^{\infty} j(j-1)\cdots(j-k)\frac{M}{(r+s)^j}r^{j-k} = \frac{M}{(r+s)^k}\sum_{j=k}^{\infty} j(j-1)\cdots(j-k)\left(\frac{r}{r+s}\right)^{j-k}$$

$$= \frac{M}{(r+s)^k}\frac{k!}{\left(1-\frac{r}{r+s}\right)^{k+1}} = \frac{M}{(r+s)^k}\frac{k!(r+s)^{k+1}}{s^{k+1}} = \frac{m \cdot k!}{s^k}$$

where $m = \frac{M(r+s)}{s}$. 👍

We now give the proof of Theorem V.16.1.6. and most of Theorem V.16.1.7..

PROOF: Let $\sum_{k=0}^{\infty} b_k(x-x_1)^k$ be the Taylor series for f around x_1. We will show that for every r, $|x_1 - x_0| < r < R$, and s, $0 < s < R - r$, this Taylor series has radius of convergence at least s and converges to $f(x)$ whenever $|x - x_1| < \min(s, r - |x_1 - x_0|)$. This will show that this Taylor series has radius of convergence at least S and that it converges to f in a neighborhood of x_1; it follows that f is analytic on I.

Fix such an r and s. By Lemma V.16.1.8., there is an m such that $|f^{(k)}(x)| \leq \frac{m \cdot k!}{s^k}$ for all $k \geq 0$ whenever $|x - x_0| \leq r$. In particular, we have

$$|b_k| = \frac{|f^{(k)}(x_1)|}{k!} \leq \frac{m}{s^k}$$

for all k, so the radius of convergence of $\sum_{k=0}^{\infty} b_k(x-x_1)^k$ is

$$\frac{1}{\limsup |b_k|^{1/k}} \geq \frac{1}{\limsup \left(\frac{m}{s^k}\right)^{1/k}} = s \ .$$

Since s can be made arbitrarily close to S by choosing r close enough to $|x_1 - x_0|$, the radius of convergence is at least S.

Let g be the function on $J = (x_1 - S, x_1 + S)$ defined by $\sum_{k=0}^{\infty} b_k(x-x_1)^k$, and let

$$g_n(x) = \sum_{k=0}^{n} b_k(x-x_1)^k$$

for $n \in \mathbb{N}$. Then $g_n \to g$ on J. Set $\phi_n = f - g_n$. Then ϕ_n is C^∞ on I, $\phi_n^{(k)}(x_1) = 0$ for $0 \leq k \leq n$, and $\phi_n^{(k)} = f^{(k)}$ on I for $k > n$ since g_n is a polynomial of degree $\leq n$.

Set $t = \min(s, r - |x_1 - x_0|)$, and let x satisfy $|x - x_1| < t$. By the Lagrange Remainder Formula (), we have

$$\phi_n(x) = \frac{\phi_n^{(n+1)}(c_n)}{(n+1)!}(x-x_1)^{n+1} = \frac{f^{(n+1)}(c_n)}{(n+1)!}(x-x_1)^{n+1}$$

for some c_n between x_1 and x; since $|c_n - x_0| \leq r$ we have

$$|\phi_n(x)| \leq \frac{m}{s^{n+1}}|x-x_1|^{n+1} = m\left(\frac{|x-x_1|}{s}\right)^{n+1} \ .$$

Since $|x - x_1| < s$, $\phi_n(x) \to 0$, $g_n(x) \to f(x)$, and thus $f = g$ on $(x_1 - t, x_1 + t)$. 👍

V.16.1.9. We have now proved V.16.1.6., and all that remains to finish the proof of V.16.1.7. is to show that $f = g$ on all of J. We have that f and g are both analytic on J by V.16.1.6., and $f = g$ on a neighborhood of x_1. The conclusion then follows by an application of V.16.2.2..

V.16.1.10. An analogous theorem holds for complex analytic functions: if f is defined by a complex power series around z_0 on an open disk D of radius $R > 0$, $z_1 \in D$, and $S = R - |z_1 - z_0|$, then the Taylor series for f around z_1 has radius of convergence at least S and converges to f on the open disk of radius S around z_1. This can be proved by an identical proof to the one above, but there is a simpler proof using other theorems of complex analysis.

Growth of Derivatives of Analytic Functions

We can obtain a closely related characterization of analytic functions: they are exactly the C^∞ functions whose derivatives "do not grow too rapidly" locally. Precisely:

V.16.1.11. THEOREM. Let f be a real-valued function on an open set U in $\mathbb{R}$. Then f is analytic on U if and only if f is C^∞ on U and, for each $x_0 \in U$, there is an open interval J containing x_0 and natural numbers m and n such that

$$|f^{(k)}(x)| \leq m \cdot k! \cdot n^k$$

for all $k \geq 0$ and all $x \in J$.

PROOF: Suppose f is analytic on U and $x_0 \in U$. Then the Taylor series for f around x_0 has positive radius of convergence R, and there is an r, $0 < r < R$, such that this Taylor series represents f on $J = (x_0 - r, x_0 + r)$. Fix s with $0 < S < R - r$. Then by Lemma V.16.1.8., there is an m such that, for any $k \geq 0$ and $x \in J$, $|f^{(k)}(x)| \leq \frac{m \cdot k!}{s^k}$. We may assume $m \in \mathbb{N}$, and we may take n any natural number $\geq \frac{1}{s}$.

Conversely, suppose f is C^∞ on U and the growth condition is satisfied. Fix $x_0 \in U$, and let $\sum_{k=0}^{\infty} a_k(x - x_0)^k$ be the Taylor series of f around x_0. Let J, m, and n be as in the statement, and set $s = \frac{1}{n}$. Then there is an r, $0 < r < s$, such that $(x_0 - r, x_0 + r) \subseteq J$. For each k, let p_k be the k'th Taylor polynomial for f around x_0, and $r_k = f - p_k$ the k'th remainder. If $|x - x_0| < r$, then we have

$$r_k(x) = \frac{f^{(k+1)}(c_k)}{(k+1)!}(x - x_0)^{k+1}$$

for some c_k between x_0 and x; since $c_k \in (x_0 - r, x_0 + r)$ we have

$$|r_k(x)| \leq m \cdot n^{k+1} r^{k+1} = m \left(\frac{r}{s}\right)^{k+1} \to 0$$

as $k \to \infty$; thus the Taylor series for f around x_0 converges to f on $(x_0 - r, x_0 + r)$ and f is analytic on U. 👍

V.16.2. Rigidity of Analytic Functions

One of the remarkable properties of an analytic function f on an open interval I is that the values of f on a small part of I (e.g. a small subinterval of I) completely determine the values of f on all of I. In particular, if an analytic function f on an interval I is identically 0 on a small subinterval of I, then f is identically 0 on all of I, and in fact the zeroes of a nonzero analytic function are (relatively) isolated in I.

V.16.2.1. THEOREM. Let f and g be analytic on an open interval I, and (x_n) a sequence of distinct points in I converging to $x_0 \in I$. If $f(x_n) = g(x_n)$ for all n, then $f(x) = g(x)$ for all $x \in I$.

V.16.2.2. Corollary. Let f and g be analytic on an open interval I, and J a (nondegenerate) subinterval of I. If $f(x) = g(x)$ for all $x \in J$, then $f(x) = g(x)$ for all $x \in I$.

V.16.2.3. Corollary. Let f be analytic and not identically 0 on an open interval I. Then the zeros of f in I are isolated in I: if $x_0 \in I$ and $f(x_0) = 0$, then there is a deleted neighborhood U of x_0 in I such that $f(x) \neq 0$ for all $x \in U$.

V.16.2.4. **Caution:** Two analytic functions on an interval I can agree at a sequence of points in I without being identical, so long as the sequence has no cluster points in I. For example, $f(x) = \sin x$ takes the same values as the constant function 0 at $n\pi$ for all n. The sequence where f and g agree can even have a limit which is not in I: $f(x) = \sin\left(\frac{1}{x}\right)$ is analytic on $(0, +\infty)$ and agrees with the constant function 0 at a sequence of points $\left(\frac{1}{n\pi}\right)$ converging to 0.

To prove Theorem V.16.2.1., we need a lemma which is of independent interest:

V.16.2.5. Lemma. Let f be C^∞ on an open interval I, and (x_n) a sequence of distinct points in I converging to $x_0 \in I$. If $f(x_n) = 0$ for all n, then $f^{(k)}(x_0) = 0$ for all k.

Proof: We have $f(x_0) = \lim_{n\to\infty} f(x_n) = 0$ by continuity, and

$$f'(x_0) = \lim_{n\to\infty} \frac{f(x_n) - f(x_0)}{x_n - x_0} = 0 \ .$$

Suppose $f(x_0) = f'(x_0) = \cdots = f^{(k-1)}(x_0) = 0$ and $f^{(k)}(x_0) \neq 0$, and set $a = \frac{f^{(k)}(x_0)}{k!}$. Then by Taylor's Theorem () we have

$$f(x) = a(x - x_0)^k + o((x - x_0)^k)$$

for all $x \in I$. Thus there is a $\delta > 0$ such that

$$|f(x) - a(x - x_0)^k| < \frac{|a|}{2}|x - x_0|^k$$

$$|f(x)| > \frac{|a|}{2}|x - x_0|^k > 0$$

whenever $0 < |x - x_0| < \delta$. But there is an n with $0 < |x_n - x_0| < \delta$, a contradiction. Thus there is no such k.

See V.10.7.11. for an alternate argument. We now give the proof of Theorem V.16.2.1..

Proof: First suppose g is identically 0, i.e. $f(x_n) = 0$ for all n. Then by V.16.2.5. the Taylor series for f around x_0 is identically 0, i.e. f is identically 0 in a neighborhood of x_0 in I. Let E be the set of all $x \in I$ such that f is identically 0 in a neighborhood of x in I. Then E is nonempty since $x_0 \in E$, and it is obviously open in I. If $c \in I$ is in the closure of E, another application of V.16.2.5. shows that $f^{(k)}(c) = 0$ for all k, and hence f is identically 0 in a neighborhood of c in I, i.e. $c \in E$. Thus E is relatively closed in I. Since I is connected, $E = I$.

Now consider the case of general g. We have that $f - g$ is analytic on I and vanishes at each x_n, so $f - g$ is identically 0 on I.

V.16.2.6. There are (nonanalytic) C^∞ functions on I satisfying the hypotheses of Lemma V.16.2.5. which are not identically zero (V.17.2.2.). Thus C^∞ functions do not generally have the same rigidity properties as analytic functions. In fact, the difference is dramatic; see V.17.2.9..

There are many important consequences of V.16.2.1., for example:

V.16.2.7. COROLLARY. Let f be an analytic function on an open interval I, and $x_0 \in I$. Let J be the interval of convergence of the Taylor series for f around x_0. Then the Taylor series for f around x_0 represents f on all of $I \cap J$, and f extends uniquely to an analytic function on $I \cup J^o$.

PROOF: Let g be the function defined on J by the Taylor series for f around x_0. Then g is analytic on J^o. The functions f and g are analytic on $I \cap J^o$, and agree in a neighborhood of x_0 by V.16.1.7., so they agree on all of $I \cap J^o$ by V.16.2.2.. If x is an endpoint of J contained in I, then x is in the closure of $I \cap J^0$ and both f and g are continuous at x (V.15.3.17.), so $f(x) = g(x)$.

For the last statement, define $h(x) = f(x)$ if $x \in I$ and $h(x) = g(x)$ if $x \in J^o$. Then h is well defined by the first part of the proof, and is obviously analytic on $I \cup J^o$. Since $I \cup J^o$ is an interval, and any extension of f to $I \cup J^o$ agrees with f on I, uniqueness follows from V.16.2.2.. 👍

V.16.2.8. EXAMPLE. Note that the interval I cannot be replaced by an open set U in this result: let $U = \mathbb{R} \setminus \{0\}$, and let $f(x) = 0$ for $x < 0$ and $f(x) = 1$ for $x > 0$. Then f is analytic on U, but V.16.2.7. fails for any $x_0 \in U$.

V.16.2.9. The exact complex analog of V.16.2.1. holds, with essentially identical proof: if f and g are complex analytic on a connected open set V, (z_n) is a sequence in V converging to $z_0 \in V$, and $f(z_n) = g(z_n)$ for all n, then $f(z) = g(z)$ for all $z \in V$. V.16.2.7. also extends to the complex case, with "interval" replaced by "disk." (There are topological obstructions to extending this result to connected open sets in $\mathbb{C}$, even simply connected open sets, but it does extend to convex open sets.)

This is needed to prove the following important fact:

V.16.2.10. THEOREM. Let f be a real analytic function on an open set U in $\mathbb{R}$. Then there is an open set V in $\mathbb{C}$ with $V \cap \mathbb{R} = U$, and a complex analytic function g on V, such that $g|_U = f$.

PROOF: For each $x_0 \in U$, the Taylor series for f around x_0, regarded as a complex power series, defines a complex analytic function g_{x_0} on an open disk in $\mathbb{C}$ centered at x_0 whose intersection with the real axis is contained in U, and on which g_{x_0} agrees with f. It is only necessary to show that these functions agree on the overlaps of their domains; then g can be defined on the union of these disks. But any two of these functions agree on the part of the overlap of their domain disks contained in U (which is an open interval in $\mathbb{R}$), and the overlap is connected, so they agree on the entire overlap. 👍

V.16.3. Combinations of Analytic Functions

In this section, we show that all usual ways of manufacturing new functions out of old preserve analyticity. Specifically:

(i) Linear combinations of analytic functions are analytic.

(ii) Derivatives and antiderivatives of analytic functions are analytic (V.15.3.20., V.15.3.23.).

(iii) Products of analytic functions are analytic.

(iv) Quotients of analytic functions are analytic if the denominator is nonzero.

(v) Compositions of analytic functions are analytic.

(vi) Inverse functions of analytic functions with nonzero derivatives are analytic.

These results, among other things, provide a large supply of standard analytic functions. The direct proofs range from very easy (i) to easy consequences of previous results ((ii)–(iii)) to moderately involved ((iv)–(v)) to quite difficult (vi). The "easy" way to prove (iii)–(vi) is to use V.16.2.10. and theorems from complex analysis, but we will describe direct proofs not using complex analysis.

Linear Combinations of Analytic Functions

V.16.3.1. Proposition. Let f and g be analytic on an open set U, and α a scalar. Then $f+g$ and αf are analytic on U.

Proof: The conclusions are immediate consequences of IV.2.1.16. and IV.2.1.18. respectively (plus the observation that linear combinations of power series centered at the same place are also power series). 👍

Unordered Power Series

V.16.3.2. Definition. An *unordered power series* is an unordered sum (IV.2.9.2.)

$$\sum_{j\in J} a_j X^{n_j}$$

where J is an index set, a_j is a scalar, X is an indeterminate symbol, and $n_j \in \mathbb{N}\cup\{0\}$ for each j. (Thus we may have unordered real power series or unordered complex power series).

The differences between an unordered power series and an ordinary formal power series are:

(i) The order of the terms is not specified.

(ii) There may be many terms, even infinitely many, with the same power of X.

V.16.3.3. An unordered power series $\sum_{j\in J} a_j X^{n_j}$ with coefficients in $\mathbb{F}$ ($= \mathbb{R}$ or $\mathbb{C}$) defines a function by setting

$$f(x) = \sum_{j\in J} a_j x^{n_j}$$

whose domain is the set of x for which the unordered sum converges (i.e. converges absolutely by IV.2.9.9.). The domain may be empty (we adhere to the convention that $x^0 = 1$ even if $x = 0$). We may more generally fix an x_0 and consider the function

$$f(x) = \sum_{j\in J} a_j (x - x_0)^{n_j} \ .$$

V.16.3.4. While the index set J can in principle be uncountable, if it is not countable, or at least if $\{j : a_j \neq 0\}$ is uncountable (terms with $a_j = 0$ can be effectively ignored in the sum), the unordered sum cannot converge for any $x \neq 0$ by IV.2.9.6.. Thus in examining such series expansions of functions there is no essential loss of generality in considering only countable index sets. So from now on we will restrict to the case where J is countable.

V.16.3.5. We can formally rearrange an unordered power series into an ordinary power series. To do this, for each $k \in \mathbb{N} \cup \{0\}$ set $J_k = \{j : n_j = k\}$ (i.e. the set of indices for which the power of X in the term is k). We can then form the ordinary infinite series

$$\sum_{k=0}^{\infty} \left[\left(\sum_{j\in J_k} a_j \right) x^k \right]$$

provided the unordered sum $\sum_{j\in J_k} a_j$ converges for each k. This power series is called the *standard revision* of the unordered power series.

V.16.3.6. THEOREM. Let

$$\sum_{j\in J} a_j X^{n_j}$$

be an unordered power series, and x_0 a scalar. Suppose the unordered sum

$$\sum_{j\in J} a_j (x_1 - x_0)^{n_j}$$

converges for some $x_1 \neq x_0$. Then

(i) The unordered sum

$$f(x) = \sum_{j\in J} a_j (x - x_0)^{n_j}$$

converges for all x, $|x - x_0| \leq |x_1 - x_0|$.

(ii) The unordered sum

$$\sum_{j\in J_k} a_j$$

converges for each k.

(iii) The standard revision

$$\sum_{k=0}^{\infty}\left[\left(\sum_{j\in J_k} a_j\right)(x-x_0)^k\right]$$

of the unordered power series converges absolutely to $f(x)$ for all x, $|x-x_0|\le|x_1-x_0|$. (In particular, this power series has radius of convergence at least $|x_1-x_0|$.)

(If $x_1=x_0$, (i) and (iii) still hold, and (ii) holds for $k=0$, but the result has little content in this case.)

Proof: By IV.2.9.9.,

$$\sum_{j\in J}|a_j(x_1-x_0)^{n_j}|$$

converges, and thus for $|x-x_0|\le|x_1-x_0|$ the unordered sum

$$\sum_{j\in J}|a_j(x-x_0)^{n_j}|$$

converges by comparison, hence $\sum_{j\in J} a_j(x-x_0)^{n_j}$ converges. (Technically we should convert the unordered sums to ordinary infinite series by taking the terms in some order; the order is irrelevant.) Also, for any k, the unordered sum

$$\sum_{j\in J_k}|a_j(x_1-x_0)^{n_j}| = |(x_1-x_0)^k|\sum_{j\in J_k}|a_j|$$

converges since it is a subsum of a convergent nonnegative unordered sum. Thus, since $x_1\ne x_0$,

$$\sum_{j\in J_k} a_j$$

converges for each k. Finally, for any x with $|x-x_0|\le|x_1-x_0|$ the standard revision

$$\sum_{k=0}^{\infty}\left[\left(\sum_{j\in J_k} a_j\right)(x-x_0)^k\right]$$

(with the terms from each J_k ordered in some arbitrary way) is a drastic revision (IV.2.10.14.) of the unordered sum

$$f(x)=\sum_{j\in J} a_j(x-x_0)^{n_j}$$

(with terms ordered in some way). Since the unordered sum converges (absolutely), the standard revision converges to the same sum by IV.2.10.16.. 👍

V.16.3.7. Conversely, any ordinary power series can be regarded as an unordered power series. The unordered power series converges exactly whenever the original power series converges absolutely, i.e. either on the interval (or disk) of convergence or the open interval (or disk) of convergence of the power series.

V.16.3.8. Corollary. Let

$$\sum_{j\in J} a_j X^{n_j}$$

be an unordered power series. If the unordered series

$$\sum_{j\in J} a_j x^{n_j}$$

converges for some $x \neq 0$, then for any x_0 the function

$$f(x) = \sum_{j\in J} a_j (x - x_0)^{n_j}$$

is analytic on an interval (or disk) around x_0. The Taylor series for f around x_0 is the standard revision of the unordered power series.

V.16.3.9. This result can be used to justify many formal manipulations with power series. The rough general principle is:

> If f and g are functions defined by infinite summations, and the summations can be formally manipulated into the same sum, *and the summations converge absolutely in an appropriate sense*, then the functions are equal.

We give several specific examples below (V.16.3.12.,).

Products of Analytic Functions

V.16.3.10. Theorem. Let f and g be analytic functions on an open set U. Then fg is analytic on U.

Proof: Fix $x_0 \in U$, and let

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k \quad \text{and} \quad \sum_{k=0}^{\infty} b_k (x - x_0)^k$$

be the Taylor series for f and g, converging to f and g on $\{x : |x - x_0| < R\} \subseteq U$. If $|x - x_0| < R$, then the above series converge absolutely, so the product series

$$\sum_{n=0}^{\infty} \left[\left(\sum_{k=0}^{n} a_k b_{n-k} \right) (x - x_0)^k \right]$$

converges (absolutely) to $f(x)g(x)$ by IV.2.11.8.. 👍

V.16.3.11. It can be readily checked by repeated applications of the Product Rule that if f and g are infinitely differentiable at x_0, then so is fg, and that the Taylor series for fg around x_0 is the product of the Taylor series for f and g around x_0 (i.e. the Taylor series for fg around x_0 is the standard revision of the unordered power series obtained by multiplying the two Taylor series termwise). This observation is not needed for the proof of V.16.3.10., but it and V.16.3.10. give additional motivation for why the definition (IV.2.11.5.) of the product of two infinite series is the "right" one.

Compositions of Analytic Functions

V.16.3.12. THEOREM. Let f be analytic on an open set U, g analytic on an open set V, and $f(U) \subseteq V$. Then $g \circ f$ is analytic on U.

PROOF: We may proceed formally as follows, ignoring convergence questions for the moment. Let $x_0 \in U$, and $y_0 = f(x_0) \in V$. Let

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k \quad \text{and} \quad g(y) = \sum_{k=0}^{\infty} b_k (y - y_0)^k$$

be the Taylor series for f around x_0 and for g around y_0. Then we can expand

$$g(f(x)) = \sum_{k=0}^{\infty} b_k (f(x) - y_0)^k = \sum_{k=0}^{\infty} b_k \left(\sum_{j=1}^{\infty} a_j (x - x_0)^j \right)^k .$$

If we expand the k'th power series

$$\left(\sum_{j=1}^{\infty} a_j (x - x_0)^j \right)^k$$

into a product series

$$\sum_{n=0}^{\infty} c_{kn} (x - x_0)^n$$

for each k (cf. IV.2.11.11.), we have a drastic revision

$$\sum_{k=0}^{\infty} b_k \left(\sum_{n=1}^{\infty} c_{kn} (x - x_0)^n \right) \tag{V.1}$$

of an unordered power series

$$\sum_{(n,k)} b_k c_{kn} (x - x_0)^n \ . \tag{V.2}$$

The drastic revision V.1 converges to $g(f(x))$ if x is sufficiently close to x_0; so if the unordered series V.2 converges for some $x \neq x_0$, we can conclude from V.16.3.8. that its sum (which is $g \circ f$ by IV.2.10.16.) is analytic on a neighborhood of x_0. Since $x_0 \in U$ is arbitrary, the result follows.

The two power series

$$\sum_{k=1}^{\infty} |a_k| x^k \quad \text{and} \quad \sum_{k=0}^{\infty} |b_k| y^k$$

have a positive radius of convergence, hence define analytic functions F and G near 0. Thus, if t is a sufficiently small positive number, we can write

$$G(F(t)) = \sum_{k=0}^{\infty} |b_k| F(t)^k = \sum_{k=0}^{\infty} |b_k| \left(\sum_{j=1}^{\infty} |a_j| t^j \right)^k \tag{V.3}$$

where all the series converge (absolutely). Expanding the k'th power series

$$\left(\sum_{j=1}^{\infty} |a_j| t^j \right)^k = \sum_{n=0}^{\infty} d_{kn} t^n$$

for each k (note that $d_{kn} \geq 0$ for all k and n) and converting V.3 to an unordered series

$$\sum_{(n,k)} |b_k| d_{kn} t^n$$

we obtain a convergent nonnegative unordered series which converges to $G(F(t))$.

If we compare the series

$$\left(\sum_{j=1}^{\infty} a_j (x - x_0)^j \right)^k = \sum_{n=0}^{\infty} c_{kn} (x - x_0)^n$$

$$\left(\sum_{j=1}^{\infty} |a_j| t^j \right)^k = \sum_{n=0}^{\infty} d_{kn} t^n$$

we obtain easily from the triangle inequality that $|c_{kn}| \leq d_{kn}$ for all k and n. Thus the unordered series

$$\sum_{(n,k)} b_k c_{kn} t^n$$

converges absolutely by comparison, which was what needed to be shown to complete the proof (taking $x = x_0 + t$). 👍

V.16.3.13. Remarks. Here are some remarks about the theorem and its proof.

(i) In most of the proof we work with $\sum_{k=1}^{\infty} a_k (x - x_0)^k$, not $\sum_{k=0}^{\infty} a_k (x - x_0)^k$. This is because we necessarily have $a_0 = f(x_0) = y_0$ and we are working with $f(x) - y_0$.

(ii) The k'th power sum

$$\left(\sum_{j=1}^{\infty} a_j (x - x_0)^j \right)^k$$

when converted into the product series

$$\sum_{n=0}^{\infty} c_{kn}(x-x_0)^n$$

has summation starting at k, i.e. $c_{kn}=0$ for $n<k$. Thus the standard revision of the unordered series

$$\sum_{(n,k)} b_k c_{kn}(x-x_0)^n$$

representing $g\circ f$ is an ordinary revision, not a drastic revision:

$$\sum_{n=0}^{\infty}\left[\left(\sum_{k=0}^{n} b_k c_{kn}\right)(x-x_0)^n\right] \tag{V.4}$$

which must be the Taylor series for $g\circ f$ around x_0.

(iii) If f is $\mathcal{C}^\infty$ on U, g is $\mathcal{C}^\infty$ on V, and $f(U)\subseteq V$, then $g\circ f$ is $\mathcal{C}^\infty$ on U by the Chain Rule. If $x_0\in U$ and the Taylor series for $g\circ f$ around x_0 is calculated using Faà di Bruno's formula (V.3.7.8.), the same power series as in V.4 is obtained, as can be verified by unpleasant direct calculation. In fact, the derivation of this Taylor series in the proof of V.16.3.12. is a good way to derive (but not prove!) Faà di Bruno's formula.

Quotients of Analytic Functions

The next result is an immediate corollary of V.16.3.12., since $g(x)=\frac{1}{x}$ is analytic on $\{x : x\neq 0\}$:

V.16.3.14. Corollary. Let f be analytic on an open set U, with $f(x)\neq 0$ for all $x\in U$. Then $\frac{1}{f}$ is analytic on U.

Combining this with V.16.3.10., we obtain:

V.16.3.15. Theorem. Let f and g be analytic on an open set U, with $g(x)\neq 0$ for all $x\in U$. Then $\frac{f}{g}$ is analytic on U.

Bernstein's Theorem

We prove a theorem obtained by S. Bernstein in 1912, which is at least interesting if not too useful. (Bernstein only stated (ii), but the more precise statements of (i) and (iii) follow easily from his proof.)

V.16.3.16. Theorem. Let $I=(a,b)$ be an open interval in $\mathbb{R}$ (where $a\in\mathbb{R}\cup\{-\infty\}$ and $b\in\mathbb{R}\cup\{+\infty\}$), and let f be a C^∞ function on I. If $f^{(k)}(x)\geq 0$ for all $k\geq 0$ and all $x\in I$, then:

(i) If $x_0\in I$ and $R=b-x_0$ ($R=+\infty$ if $b=+\infty$), then the Taylor series for f around x_0 converges to $f(x)$ for all $x\in I$ with $|x-x_0|<R$. (In particular, the radius of convergence of this Taylor series is at least R.)

(ii) f is analytic on I.

(iii) If $\ell = b - a$ ($\ell = +\infty$ if $b = +\infty$ or $a = -\infty$), then f extends uniquely to an analytic function on $(a - \ell, b)$ (all of $\mathbb{R}$ if $b = +\infty$).

Note, however, that the extension in (iii) does not necessarily have $f^{(k)}(x) \geq 0$ for all k and x: consider $f(x) = x^3$ on $(0, \infty)$.

PROOF: (i): Fix $x_0 \in I$, and fix $x > x_0$. Let $c \in I$ with $c > x$. Let p_n be the n'th Taylor polynomial for f around x_0. We have $p_n(x) \geq 0$ since all the terms in the Taylor series are nonnegative. By the Integral Form of the Remainder (V.10.4.2.), we have

$$R_n(x) = f(x) - p_n(x) = \frac{1}{n!} \int_{x_0}^{x} f^{(n+1)}(t)(x - t)^n \, dt \ .$$

Using two changes of variable ($u = t - x_0$ and $s = \frac{u}{x - x_0}$), we obtain

$$R_n(x) = \frac{1}{n!} \int_0^{x - x_0} f^{(n+1)}(u + x_0)(x - x_0 - u)^n \, du = \frac{(x - x_0)^{n+1}}{n!} \int_0^1 f^{(n+1)}((x - x_0)s + x_0)(1 - s)^n \, ds$$

and, since $f^{(n+1)}$ is nonnegative and nondecreasing ($f^{(n+2)}$ is also nonnegative), we have

$$0 \leq R_n(x) \leq \left(\frac{x - x_0}{c - x_0} \right)^{n+1} \frac{(c - x_0)^{n+1}}{n!} \int_0^1 f^{(n+1)}((c - x_0)s + x_0)(1 - s)^n \, ds = \left(\frac{x - x_0}{c - x_0} \right)^{n+1} R_n(c) \ .$$

We have $R_n(c) = f(c) - p_n(c) \leq f(c)$ since $p_n(c) \geq 0$, so we have

$$0 \leq R_n(x) \leq \left(\frac{x - x_0}{c - x_0} \right)^{n+1} f(c)$$

and thus $R_n(x) \to 0$ as $n \to \infty$, i.e. the Taylor series for f around x_0 converges to $f(x)$.

Now suppose $x \in I$, $x < x_0$, $|x - x_0| < R$. Set $y = x_0 + |x - x_0|$, i.e. $y > x_0$ and $|y - x_0| = |x - x_0|$. By comparison of integrals we have $|R_n(x)| \leq R_n(y)$ since $f^{(n+1)}$ is nondecreasing, so $R_n(x) \to 0$ as $n \to \infty$.

(ii): Immediate from (i).

(iii): If $x_0 > a$ and $R = b - x_0$, then by (i) the Taylor series for f around x_0 defines an analytic function on $(x_0 - R, x_0 + R)$ which agrees with f on I. Let $x_0 \to a+$. (All the extensions defined this way must coincide on the intersections of their domains by V.16.2.1., which also gives the uniqueness of the extension.) 👍

Formal Manipulation of Power Series

V.16.3.17. The results of this section can be used to justify the formal manipulation of power series. For example, suppose we are given analytic functions f and h taking the same value at 0 and we want to find an analytic function g satisfying $g(0) = 0$ and $h = f \circ g$ (we stick to analytic functions around 0 for notational simplicity; more generally, we may take power series centered at other points than 0). Temporarily ignoring questions of convergence, we can proceed formally: write

$$f(x) = \sum_{k=0}^{\infty} a_k x^k$$

$$h(x) = \sum_{k=0}^{\infty} c_k x^k$$

with $a_0 = c_0$. The unknown function

$$g(x) = \sum_{k=1}^{\infty} b_k x^k$$

can be "found" by manipulation of the power series:

$$f(g(x)) = a_0 + a_1(b_1 x + b_2 x^2 + \cdots) + a_2(b_1 x + b_2 x^2 + \cdots)^2 + \cdots$$

We can formally multiply out the last expression to get

$$(b_1 x + b_2 x^2 + \cdots)^2 = b_1^2 x^2 + 2b_1 b_2 x^3 + \cdots$$

and similarly for higher powers. Collecting together like powers of x, we get

$$f(g(x)) = a_0 + a_1 b_1 x + (a_1 b_2 + a_2 b_1^2) x^2 + \cdots$$

and for this to be equal to $h(x)$, we get the equations

$$c_0 = a_0$$

$$c_1 = a_1 b_1$$

$$c_2 = a_1 b_2 + a_2 b_1^2$$

etc., which must be solved for the b_k. It is easy to see exactly when this can be done recursively, and what the (necessarily unique) solution is when it exists.

To justify the formal manipulations, the theorems of this section can be used. In fact, using (), we can show:

V.16.3.18. THEOREM. If the system is solvable for the b_k, and if the power series

$$\sum_{k=1}^{\infty} b_k x^k$$

has a positive radius of convergence, then the function g defined by the series satisfies $h = f \circ g$ in a neighborhood of 0. Conversely, if g is an analytic function in a neighborhood of 0 satisfying $g(0) = 0$ and $h = f \circ g$ in a neighborhood of 0, and

$$g(x) = \sum_{k=1}^{\infty} b_k x^k$$

is its Taylor series around 0, then the b_k satisfy the equations. In particular, g is the unique analytic function with these properties.

V.16.3.19. It is a more delicate matter to find or estimate *how large* a neighborhood of 0 the formula is valid in. In fact, it is valid in the largest open disk around 0 mapped by g into the open disk of convergence of the Taylor series for f around 0.

V.16.4. The Analytic Morse Lemma

V.16.4.1. If f is analytic on an interval I around x_0, and $r_{n-1} = f - p_{n-1}$ is the remainder function, then

$$\lim_{x\to x_0} \frac{r_{n-1}(x)}{(x-x_0)^n} = \frac{f^{(n)}(x_0)}{n!}$$

so if we set

$$g_n(x) = \frac{r_{n-1}(x)}{(x-x_0)^n}\ (x \neq x_0), \quad g_n(x_0) = \frac{f^{(n)}(x_0)}{n!}$$

for $x \in I$, then g_n is continuous on I and analytic on $I \setminus \{x_0\}$, and

$$f(x) = p_{n-1}(x) + (x-x_0)^n g_n(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n-1)}(x_0)}{(n-1)!}(x-x_0)^{n-1} + (x-x_0)^n g_n(x)$$

for $x \in I$.

V.16.4.2. Proposition. If f is analytic on I, then for each $n \in \mathbb{N}$ g_n is analytic on I.

Proof: Since $r_{n-1} = f - p_{n-1}$, hence also g_n, is analytic on $I \setminus \{x_0\}$, analyticity near x_0 is the only issue. The Taylor series for f around x_0 converges u.c. to f on an open interval $I_0 \subseteq I$, i.e.

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n-1)}(x_0)}{(n-1)!}(x-x_0)^{n-1} + \sum_{k=n}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

for $x \in I_0$, so if we set

$$g_n(x) = \sum_{k=n}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^{k-n}$$

for $x \in I_0$ (the power series converges u.c. to an analytic function on I_0), then

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n-1)}(x_0)}{(n-1)!}(x-x_0)^{n-1} + (x-x_0)^n g_n(x)$$

for $x \in I_0$, and g_n is analytic on I_0. 👍

The Analytic One-Dimensional Morse Lemma

V.16.4.3. If f is analytic in an interval around x_0, $f(x_0) = c$, and $f'(x_0) = \cdots = f^{(m-1)}(x_0) = 0$ but $f^{(m)}(x_0) \neq 0$, then $f(x)$ is approximately $p_m(x) = c + d(x-x_0)^m$ for x near x_0, where d is a constant ($d = \frac{f^{(m)}(x_0)}{m!}$). But we can say more: up to an analytic change of coordinates, f is *exactly* of this form, and we can even take $d = \pm 1$.

V.16.4.4. Theorem. [Analytic Generalized One-Dimensional Morse Lemma] Let $m \in \mathbb{N}$, and let f be analytic on an interval I around x_0, with $f(x_0) = c$ and $f'(x_0) = \cdots = f^{(m-1)}(x_0) = 0$ but $f^{(m)}(x_0) \neq 0$. Then there is an analytic order-preserving diffeomorphism ϕ from an open interval J around 0 onto an open interval $\tilde{I}$ around x_0, with $\phi(0) = x_0$, such that

$$[f \circ \phi](t) = c \pm t^m$$

for all $t \in J$, where the sign is the sign of $f^{(m)}(x_0)$.

Proof: By V.16.4.2., for $x \in I$ f can be written $f(x) = c + (x - x_0)^m g_m(x)$ with $g_m(x_0) = d \neq 0$. Thus there is an open interval $K \subseteq I$ around x_0 on which g_m is nonzero. For $x \in \tilde{I}$, set

$$\psi(x) = (x - x_0)h(x)$$

where

$$h(x) = |g_m(x)|^{1/m} .$$

Then h, hence ψ, is an analytic function on K since $\rho(x) = |x|^{1/m}$ is analytic on $\mathbb{R} \setminus \{0\}$, $\psi(x_0) = 0$, and

$$\psi'(x) = h(x) + (x - x_0)h'(x) .$$

Since h' is bounded near x_0 and $h(x_0) > 0$, there is an open interval $\tilde{I}$ around x_0 on which $\psi' > 0$. Thus ψ maps $\tilde{I}$ diffeomorphically onto an open interval J containing 0. Let $\phi : J \to \tilde{I}$ be the inverse function. Then ϕ is also analytic by the Inverse Function Theorem (). If $t = \psi(x) \in J$, then

$$[f \circ \phi](t) = f(x) = c + (x - x_0)^m g_m(x) = c \pm [(x - x_0)h(x)]^m = c \pm \psi(x)^m = c \pm t^m .$$

The classical one-dimensional Morse Lemma (due to Marston Morse) is the case $m = 2$:

V.16.4.5. Corollary. [Analytic One-Dimensional Morse Lemma] Let f be analytic on an interval I around x_0, with $f(x_0) = c$ and $f'(x_0) = 0$ but $f''(x_0) \neq 0$. Then there is an analytic order-preserving diffeomorphism ϕ from an open interval J around 0 onto an open interval $\tilde{I}$ around x_0, with $\phi(0) = x_0$, such that

$$[f \circ \phi](t) = c \pm t^2$$

for all $t \in J$, where the sign is the sign of $f''(x_0)$.

V.16.4.6. If we allow ϕ to be orientation-reversing, if m is odd we can always take the + sign in the conclusion. But if m is even (in particular, if $m = 2$), the sign does not depend on the orientation of ϕ.

V.17. Smooth Non-Analytic Functions

A striking and important phenomenon is that a smooth (C^∞) function on an interval I need not be analytic on I; in fact, it can be "completely nonanalytic" on I. In this section we describe various ways in which a smooth function can be nonanalytic.

This is a topic in which the fairly extensive work over a rather long period of time does not seem to be widely known in the mathematical community (e.g. the author only rather recently became aware of it). It is remarkable how often the basic results and examples (with slight variations) have been rediscovered, reproved, and republished. Only variations of Cauchy's original example, and Borel's Theorem, have apparently become standard knowledge among mathematicians. For an extensive survey of the work done in this area, see http://mathforum.org/kb/message.jspa?messageID=387148 .

V.17.1. Analyticity at a Point

The definition of an analytic function (V.16.1.1.) only defines analyticity on an open interval in $\mathbb{R}$ (or an open set in $\mathbb{C}$). It is useful to make the following definition of "analyticity at a point."

V.17.1.1. Definition. Let I be an interval in $\mathbb{R}$, and $x_0 \in I$. A real-valued C^∞ function f on I is *analytic at* x_0 if the Taylor series of f around x_0 converges to f in some relatively open subinterval of I containing x_0 (an open interval if x_0 is an interior point of I). A point where f is not analytic is called a *singular point* of f.

The next result, an immediate corollary of V.16.1.6., is fundamental:

V.17.1.2. Proposition. Let $f \in C^\infty(I)$. The set

$$A_f = \{x_0 \in I : f \text{ is analytic at } x_0\}$$

is relatively open in I.

Thus the set of singular points of f is a relatively closed subset of I. (It is not yet clear that this set can be nonempty.)

Types of Singular Points

V.17.1.3. If f is C^∞ on I and $x_0 \in I$, there are two potential ways f can fail to be analytic at x_0:

(i) The Taylor series of f around x_0 has radius of convergence zero. Such a point is sometimes called a *divergence point* of f. In [Zah47], such points are called *points of type (P)*, or *(P)-points* (for Pringsheim).

(ii) The Taylor series of f around x_0 has positive radius of convergence, but does not represent the function on any interval around x_0. Such a point is sometimes called a *false convergence point* of f. The name from [Zah47] is *point of type (C)*, or *(C)-point* (for Cauchy).

V.17.1.4. Thus, if f is C^∞ on I, I decomposes as the disjoint union of the set A_f of points where f is analytic, the set P_f of the (P)-points of f, and the set C_f of the (C)-points of f. A_f is (relatively) open by V.17.1.2., so $P_f \cup C_f$ is (relatively) closed.

V.17.2. (C)-Points and Smooth Partitions

V.17.2.1. The first example of a (C)-point, the best-known example in this subject, was given by Cauchy in 1823.

V.17.2.2. Example. Let

$$f(x) = \begin{cases} e^{-1/x^2} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases} .$$

Then f is clearly analytic on $\mathbb{R} \setminus \{0\}$. On $\mathbb{R} \setminus \{0\}$, it can easily be shown by induction on k () that $f^{(k)}(x)$ is a finite sum of terms of the form $c\frac{e^{-1/x^2}}{x^m}$ for $c \in \mathbb{R}$, $m \in \mathbb{N} \cup \{0\}$. Thus, since $f(0) = 0$, we have by induction that $f^{(k)}(0)$ exists and is 0 for all k (cf. V.10.7.8.).

So f is infinitely differentiable also at 0, thus C^∞ on $\mathbb{R}$. The Taylor series at 0 (Maclaurin series) for f has all terms zero, and thus has infinite radius of convergence. But it does not converge to $f(x)$ for any $x \neq 0$ since $f(x) > 0$ for $x \neq 0$. Thus 0 is a (C)-point of f and f is not analytic at 0. See Figure (). See XII.5.2.17.(ii) for an "explanation" of why the function is not analytic at 0.

V.17.2.3. This example is usually presented with an exponent of $-\frac{1}{x^2}$, but this is just somewhat a matter of taste. Identical arguments show that each of the following functions has the same properties:

$$f_2(x) = \begin{cases} e^{-1/|x|} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$$

$$f_3(x) = \begin{cases} e^{-1/x} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases}$$

$$f_4(x) = \begin{cases} e^{-1/x^2} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases}$$

i.e. each of these is C^∞ on $\mathbb{R}$, analytic on $\mathbb{R} \setminus \{0\}$, and has a (C)-point at 0 where its Taylor series is identically 0. Any of these can be called "Cauchy's example." The last two have the interesting property that they are identically zero on the negative real axis. Note that all these examples take values in $[0, 1)$, and are in particular bounded.

Smooth Bump Functions

We can combine translates of these functions to make smooth "bump functions" which are identically zero outside a specified interval:

V.17.2.4. Example. Let $a, b \in \mathbb{R}$, $a < b$. Define

$$g_{a,b}(x) = f_3(x-a)f_3(b-x) = \begin{cases} e^{-\frac{b-a}{(x-a)(b-x)}} & \text{if} \quad a < x < b \\ 0 & \text{if} \quad x \leq a \text{ or } x \geq b \end{cases} .$$

Then g is C^∞ on $\mathbb{R}$, analytic on $\mathbb{R} \setminus \{a, b\}$, has (C)-points at a and b where its Taylor series is identically zero, and is positive on (a, b) and identically zero on $\mathbb{R} \setminus (a, b)$. See Figure ().

Smooth Heaviside Functions

V.17.2.5. The function

$$\phi(x) = \int_a^x g_{a,b}(t)\,dt$$

is C^∞ on $\mathbb{R}$, analytic on $\mathbb{R} \setminus \{a, b\}$, identically 0 for $x \leq a$, takes a constant positive value for $x \geq b$, and is strictly increasing on $[a, b]$. A suitable scalar multiple $h_{a,b}$ has the same properties and is identically 1 for $x \geq b$. This can be regarded as a smooth version of the Heaviside function $\chi_{[b,\infty)}$. See Figure ().

Smooth Mesa Functions

V.17.2.6. Let $a < b < c < d$, and let f be the function

$$f(x) = h_{a,b}(x)(1 - h_{c,d}(x)) \ .$$

Then f is C^∞ on $\mathbb{R}$, analytic on $\mathbb{R} \setminus \{a, b, c, d\}$, identically 0 on $(-\infty, a]$ and on $[d, +\infty)$, identically 1 on $[b, c]$, strictly increasing on $[a, b]$, and strictly decreasing on $[c, d]$. See Figure ().

Smooth Partitions of Unity

Recall the definition of partition of unity (). We will show that on $\mathbb{R}$, smooth partitions of unity exist:

V.17.2.7. THEOREM. Let $\mathcal{U}$ be an open cover of $\mathbb{R}$. Then there is a countable partition of unity $\{f_n\}$ subordinate to $\mathcal{U}$ consisting of C^∞ functions of compact support.

PROOF: This is an easy consequence of (). There is a collection $\mathcal{V} = \{(a_n, b_n) : n \in \mathbb{Z}\}$ of bounded open intervals refining $\mathcal{U}$ such that $a_n < b_{n-1} < a_{n+1} < b_n$ for all n. Take f_n to be the smooth mesa function

$$f_n = h_{a_n,b_{n-1}}(1 - h_{a_{n+1},b_n})$$

defined in V.17.2.6. and V.17.2.5.. 👍

Smooth Functions With Prescribed Zero Sets

V.17.2.8. If f is a continuous function, then its *zero set*

$$Z_f = \{x : f(x) = 0\}$$

is always closed. We now show that every closed subset of $\mathbb{R}$ is the zero set of a smooth function. This is in stark contrast to the situation for analytic functions, where the zero set is either discrete or all of $\mathbb{R}$ (). For a simple example of a $\mathcal{C}^\infty$ function with a nonisolated zero, see Exercise V.17.5.1..

V.17.2.9. THEOREM. Let E be a closed subset of $\mathbb{R}$. Then there is a nonnegative bounded $\mathcal{C}^\infty$ function θ on $\mathbb{R}$ for which $Z_\theta = E$ and θ is analytic on $\mathbb{R} \setminus E$.

PROOF: There is a simple and appealing argument that almost works to prove this result, which can be adapted to give a real proof. The open set $U = \mathbb{R} \setminus E$ is a countable disjoint union of open intervals I_n. For each n, we can take a smooth bump function f_n which is nonzero exactly on I_n (if I_n is unbounded, take a translation of the f_3 of V.17.2.3. or its reflection instead of a bump function). Define θ to be f_n on I_n and 0 on E. Alternately, define $\theta = \sum_{n=1}^{\infty} f_n$; the series converges for each x since at most one term is nonzero.

However, there is a problem with this approach. Since the set of endpoints of the I_n is not closed and discrete in general (e.g. consider the case where E is the Cantor set), the sum of the series will not even be continuous in general unless the heights of the bumps are controlled.

We can do this as follows. The complement of E consists of at most a countable number of bounded open intervals $I_n = (a_n, b_n)$, plus possibly one or two unbounded open intervals we will deal with separately at the end of the argument. We may assume there are infinitely many I_n since otherwise a finite sum of bump functions works for f. Let $g = g_{0,1}$ be the smooth bump function of V.17.2.4., normalized so that $\max(g(x)) = 1$, and for each $k \geq 0$ let $M_k = \max |g^{(k)}|$. Let g_n be g translated to I_n, i.e.

$$g_n(x) = g\left(\frac{x - a_n}{b_n - a_n}\right)$$

for $x \in \mathbb{R}$. Set $f_n = c_n g_n$, where

$$c_n = \frac{\min(1, (b_n - a_n)^n)}{2^n \left(\max_{0 \leq k \leq n} M_k\right)} .$$

Then f_n is $\mathcal{C}^\infty$ on $\mathbb{R}$, and $|f_n^{(k)}(x)| \leq 2^{-n}$ for $0 \leq k \leq n$ and all $x \in \mathbb{R}$. Thus

$$\sum_{n=1}^{\infty} f_n^{(k)}$$

converges uniformly on $\mathbb{R}$ for each k. If $\theta = \sum_{n=1}^{\infty} f_n$, then by V.15.1.6. (the simpler version suffices), θ is $\mathcal{C}^\infty$ on $\mathbb{R}$ and $\theta^{(k)} = \sum_{n=1}^{\infty} f_n^{(k)}$ for all k. The function θ takes values in $[0, 1]$, and is clearly zero precisely on E and any unbounded intervals in E^c. There will be such an unbounded interval if and only if E has a minimum element a and/or a maximal element b. If there is such a b, adding to f a translated version of Cauchy's example f_3 makes θ positive on $(b, +\infty)$. A similar correction can be made on $(-\infty, a)$ if a exists. Since, for each $x \in E^c$, θ is equal to f_n for some n (or a translated Cauchy function) on a neighborhood of x, θ is analytic at x. 👍

The Smooth Urysohn Lemma

With somewhat more work along the same lines, we can get a smooth version of Urysohn's Lemma () for $\mathbb{R}$:

V.17.2.10. THEOREM. [SMOOTH URYSOHN LEMMA FOR $\mathbb{R}$] Let A be a closed subset of $\mathbb{R}$ and U an open set in $\mathbb{R}$ containing A. Then there is a C^∞ function ϕ from $\mathbb{R}$ to $[0, 1]$ such that $\phi \equiv 1$ on A and $\phi \equiv 0$ on $\mathbb{R} \setminus U$. We may choose ϕ so that $\phi^{-1}(\{1\}) = A$ and $\phi^{-1}(\{0\}) = \mathbb{R} \setminus U$.

Alternatively, if E and F are disjoint closed subsets of $\mathbb{R}$, there is a smooth function $\phi : \mathbb{R} \to [0,1]$ such that $\phi^{-1}(\{1\}) = E$ and $\phi^{-1}(\{0\}) = F$ (set $A = E$ and $U = \mathbb{R} \setminus F$).

PROOF: We prove the alternative form. Let E and F be disjoint closed subsets of $\mathbb{R}$. Then the open set $V = \mathbb{R} \setminus (E \cup F)$ is a countable disjoint union of open intervals, at most two of which are unbounded. Let $\{(a_i, b_i) : i \in I\}$ be the set of bounded intervals in this union such that $a_i \in E$ and $b_i \in F$, and let $\{(c_j, d_j) : j \in J\}$ be the set of bounded intervals in the collection with $c_j \in F$ and $d_j \in E$. Then I and J are countable. Between any two a_i there must be at least one of the intervals (c_j, d_j), and similarly for the other combinations.

Let $L_E = \{a_i : i \in I\}$, $L_F = \{c_j : j \in J\}$, $R_E = \{d_j : j \in J\}$, $R_F = \{b_i : i \in I\}$. L_E is countable; we claim it is closed and discrete in $\mathbb{R}$. If not, there is a sequence (a_{i_n}) in L_E of distinct points converging to some $x \in \mathbb{R}$, and passing to a subsequence we may assume that the sequence is strictly monotone (). Since each $a_{i_n} \in E$, $x \in E$. If (a_{i_n}) is strictly monotone, we also have $b_{i_n} \to x$ since b_{i_n} is between $a_{i_{n-1}}$ and $a_{i_{n+1}}$, so $x \in F$, contradicting that E and F are disjoint. Thus L_E is closed and discrete. By similar arguments, L_F, R_E, R_F are all closed and discrete in $\mathbb{R}$. Thus $L_E \cup L_F \cup R_E \cup R_F$ is also closed and discrete, i.e. there are only finitely many points in any bounded interval.

Thus the endpoints of the intervals can be reindexed $\{(a_k, b_k), (c_k, d_k) : k \in \mathbb{Z}\}$, where

$$a_k < b_k < c_k < d_k < a_{k+1}$$

for all k. If V contains an unbounded interval on either side, the indexing on that side will be finite. Define ϕ separately on each of the subintervals of $\mathbb{R}$ so defined. On $[b_k, a_{k+1}]$, let ϕ be the smooth Heaviside function h_{c_k,d_k} (V.17.2.5.), and on $[d_{k-1}, c_k]$ let ϕ be $1 - h_{a_k,b_k}$. Note that these definitions coincide on the overlaps of the intervals. If there is an unbounded interval to $+\infty$ in V, let ϕ be h_{c_k,d_k} on $[b_k, +\infty)$ if (c_k, d_k) is the last bounded interval, and $1 - h_{a_k,b_k}$ on $[d_{k-1}, +\infty)$ if (a_k, b_k) is the last bounded interval; define ϕ analogously if V has an unbounded interval to $-\infty$. Since the endpoints of the intervals form a closed discrete set, the ϕ so defined is a C^∞ function from $\mathbb{R}$ to $[0,1]$ which is identically 1 on E and 0 on F.

This ϕ, however, does not necessarily have the property that $\phi^{-1}(\{1\}) = E$ and $\phi^{-1}(\{0\}) = F$. We can obtain this property by a trick. Let θ be a C^∞ function from $\mathbb{R}$ to $[0,1)$ such that $\theta = 0$ precisely on $E \cup F$ (V.17.2.9.). Then $\psi_1 = \phi(1 - \theta)$ is a C^∞ function from $\mathbb{R}$ to $[0,1]$ which is 0 on F and which is 1 precisely on E. Similarly, $\psi_2 = 1 - (1 - \phi)(1 - \theta)$ is a C^∞ function from $\mathbb{R}$ to $[0,1]$ which is 1 on E and 0 precisely on F. Then $\psi = \frac{1}{2}(\psi_1 + \psi_2)$ is a C^∞ function from $\mathbb{R}$ to $[0,1]$ which is 1 precisely on E and 0 precisely on F. 👍

Multivariable Versions

V.17.2.11. There are versions of all these examples and results for real-valued functions on $\mathbb{R}^n$ (VIII.4.2.).

V.17.3. (P)-Points and Borel's Theorem

The first example of a (P)-point was apparently obtained by CELLÈRIER in the early 1860's, only published posthumously in 1890 (Exercise ()). The first published example was by DU BOIS-REYMOND in 1876 (Exercise ()).

Some specific examples of (P)-points will be given in (). In this subsection we will be primarily concerned with a striking theorem of E. BOREL from 1895, which provides (P)-points in profusion. The theorem can be stated in two equivalent ways.

V.17.3.1. THEOREM. Let $(a_k)_{k\geq 0}$ be any sequence of real numbers. Then there is a C^∞ function on $\mathbb{R}$ whose Taylor series around 0 (Maclaurin series) is

$$\sum_{k=0}^{\infty} a_k x^k \ .$$

V.17.3.2. COROLLARY. Let $(b_k)_{k\geq 0}$ be any sequence of real numbers. Then there is a C^∞ function f on $\mathbb{R}$ such that $f^{(k)}(0) = b_k$ for all k.

The two statements are equivalent by setting $b_k = k!a_k$.

We first give a simple proof from [Mir56] (the same proof works in the multidimensional case ()). Then we will give another argument giving an example with additional analyticity properties.

PROOF: Fix a smooth mesa function g such that g is supported on $(-1, 1)$, and for each n let M_n be a bound on $g^{(n)}$. Note that $g^{(n)}(0) = 0$ for all $n > 0$ since g is constant on an interval around 0. For $k \geq 0$ define

$$f_k(x) = a_k x^k g(c_k x)$$

where (a_k) is the given sequence and (c_k) is a sequence of natural numbers to be chosen later. Then f_k is C^∞ on $\mathbb{R}$, and by the higher product rule () and the chain rule for the g factors,

$$f_k^{(n)}(x) = \sum_{m=0}^{n} \binom{n}{m} k(k-1)\cdots(k-n+m+1) a_k c_k^m x^{k-n+m} g^{(m)}(c_k x)$$

for all n (the product $k(k-1)\cdots(k-n+m+1)$ is 1 if $m = n$). In particular, we have

$$f_k^{(k)}(0) = \sum_{m=0}^{k} \binom{k}{m} k(k-1)\cdots(m+1) a_k c_k^m 0^m g^{(m)}(0) = k!a_k = b_k$$

and $f_k^{(n)}(0) = 0$ if $n \neq k$ (for $n > k$, note that $f_k^{(n)}$ is identically 0 in a neighborhood of 0). Since $f_k(x) = 0$ for all x with $|x| \geq \frac{1}{c_k}$, we have

$$|f_k^{(n)}(x)| \leq \sum_{m=0}^{n} \binom{n}{m} k(k-1)\cdots(k-n+m+1) a_k c_k^m c_k^{-k+n-m} M_m$$

$$\leq \frac{1}{c_k} \sum_{m=0}^{n} \binom{n}{m} k(k-1)\cdots(k-n+m+1) a_k M_m$$

for $0 \leq n < k$ and all x, and thus by choosing c_k sufficiently large we obtain that

$$|f_k^{(n)}(x)| \leq 2^{-k}$$

for $0 \le n < k$ and all x. Thus the series

$$\sum_{k=0}^{\infty} f_k^{(n)}$$

converges uniformly on $\mathbb{R}$ for every $n \ge 0$ by the M-test. So the function

$$f = \sum_{k=0}^{\infty} f_k$$

is C^∞ on $\mathbb{R}$ and

$$f^{(n)} = \sum_{k=0}^{\infty} f_k^{(n)}$$

for every n by (). In particular, $f^{(n)}(0) = b_n$ for all n. 👍

V.17.3.3. To obtain an example of a smooth function with a (P)-point, apply the theorem with a sequence $a_k \to +\infty$ rapidly, e.g. $a_k = k!$.

V.17.3.4. Note that the example constructed in the above proof is analytic except on a countable set; in fact, if $x_0 \neq 0$ and U is an open neighborhood of x_0 bounded away from 0, then all but finitely many of the terms in the series are identically zero on U. Thus this theorem is strictly a single-point result.

We can in fact improve this result to make f analytic on all of $\mathbb{R} \setminus \{0\}$. This result is due to Besicovitch [Bes24].

V.17.3.5. Theorem. Let (b_k) be an arbitrary sequence of real numbers. Then there is a real-valued function f, C^∞ on $[0, +\infty)$ and analytic on $(0, +\infty)$, with $f^{(k)}(0) = b_k$ for all k (these are one-sided derivatives).

Proof: [KP02b, Lemma 2.2.4] Define

$$f_0(x) = \frac{1}{\sqrt{c_0 + x}}$$

and for each $k \ge 1$, define

$$f_k(x) = \int_0^x \int_0^{t_{k-1}} \cdots \int_0^{t_1} \frac{1}{\sqrt{c_k + t}}\, dt\, dt_1 \cdots dt_{k-1}$$

where the c_k are positive real numbers which will be chosen later, i.e. f_k is the unique function such that $f_k^{(k)}(x) = \frac{1}{\sqrt{c_k + x}}$ and $f_k^{(n)}(0) = 0$ for $0 \le n < k$. For each k, f_k is C^∞ on $(-c_k, +\infty)$.

Fix $N > 1$. If $0 \le x \le N$, we have, for $0 \le n < k$,

$$0 \le f_k^{(n)}(x) \le \int_0^N \int_0^{t_{k-n-1}} \cdots \int_0^{t_1} \frac{1}{\sqrt{t}}\, dt\, dt_1 \cdots dt_{k-n-1} = \frac{N^{k-n-1/2}}{\frac{1}{2}\frac{3}{2}\cdots\frac{2(k-n)-1}{2}} < \frac{2N^{k-n}}{(k-n-1)!}$$

and so the series

$$\sum_{k=0}^{\infty} \sigma_k f_k^{(n)}$$

converges uniformly on $[0, N]$ for every $n \geq 0$ and any choice of $\sigma_k = -1, 0, 1$ by the M-test. Thus, for any such choice of σ_k, we have that

$$f = \sum_{k=0}^{\infty} \sigma_k f_k$$

is C^∞ on $[0, \infty)$ and for each n,

$$f^{(n)} = \sum_{k=0}^{\infty} \sigma_k f_k^{(n)}$$

by (). We also have, for $\epsilon > 0$,

$$|f^{(n)}(x)| \leq \sum_{k=0}^{n} \frac{2}{(c_k + \epsilon)^{k+1/2}} + \sum_{k=n+1}^{\infty} \frac{2N^{k-n}}{(k-n-1)!} \leq 2(n+1)\left(\frac{1}{\epsilon}\right)^{n+1} + 2e^N$$

for $\epsilon \leq x \leq N$, so by () f is analytic on $(0, \infty)$.

The above results hold for any choice of c_k and σ_k. We need only show that these can be chosen so that $f^{(k)}(0) = b_k$ for all k. Since $f_k^{(n)}(0) = 0$ if $0 \leq n < k$, we must solve the equations

$$b_0 = \sigma_0 \frac{1}{c_0^{1/2}}$$

$$b_1 = -\frac{1}{2}\sigma_0 \frac{1}{c_0^{3/2}} + \sigma_1 \frac{1}{c_1^{1/2}}$$

$$b_2 = \frac{1}{2} \cdot \frac{3}{2}\sigma_0 \frac{1}{c_0^{5/2}} - \frac{1}{2}\sigma_1 \frac{1}{c_1^{3/2}} + \sigma_2 \frac{1}{c_2^{1/2}}$$

$$\cdots$$

We may solve these successively. If $b_0 = 0$, set $\sigma_0 = 0$ and $c_0 = 1$; otherwise, set $\sigma_0 = \frac{b_0}{|b_0|}$ and $c_0 = |b_0|^{-2}$. Then set σ_1 equal to $1, -1, 0$ depending whether

$$b_1 + \frac{1}{2}\sigma_0 \frac{1}{c_0^{3/2}}$$

is positive, negative, or 0, and then c_1 is determined from the previous choices. This process can clearly be continued indefinitely. 👍

As a corollary, we get a two-sided version which improves Borel's result:

V.17.3.6. Corollary. Let (b_k) be an arbitrary sequence of real numbers. Then there is a real-valued function f, C^∞ on $\mathbb{R}$ and analytic on $\mathbb{R} \setminus \{0\}$, with $f^{(k)}(0) = b_k$ for all k.

Proof: Let f be as in the Theorem for the sequence (b_k), and g as in the theorem for the sequence $((-1)^k b_k)$. Extend f to $\mathbb{R}$ by setting $f(x) = g(-x)$ for $x < 0$. The extended f is C^∞ at 0. 👍

We now obtain the main result of [Bes24], that the derivatives of a C^∞ function can be independently specified at two points with the function analytic in between:

V.17.3.7. Theorem. Let (a_k) and (b_k) be arbitrary sequences of real numbers. Then there is a real-valued function f, C^∞ on $\mathbb{R}$ and analytic on $\mathbb{R} \setminus \{0, 1\}$, such that $f^{(k)}(0) = a_k$ and $f^{(k)}(1) = b_k$ for all k.

Proof: Let $h(x) = h_{0,1}$ be the smooth Heaviside function from V.17.2.5.. Let g_0 and g_1 be functions as in V.17.3.6. for the sequences (a_k) and (b_k) respectively, and set

$$f_0(x) = g_0(x)(1 - h(x))$$
$$f_1(x) = g_1(x - 1)h(x)$$

and $f = f_0 + f_1$. Then f_0, f_1, f are C^∞ on $\mathbb{R}$ and analytic on $\mathbb{R} \setminus \{0, 1\}$. We have that $f_0^{(k)}(0) = a_k$ and $f_0^{(k)}(1) = 0$ for all k, and $f_1^{(k)}(0) = 0$ and $f_1^{(k)}(1) = b_k$ for all k. Thus f is the desired function. 👍

V.17.4. Everywhere Nonanalytic Functions

There are smooth functions on an interval I which are not analytic anywhere. By V.17.1.2., these are precisely the functions f for which the set $C_f \cup P_f = I$. Since $C_f \cup P_f$ is closed in I, to find examples it suffices to find f's for which C_f or P_f is dense in I. (It turns out that for any such example, P_f must be dense in I (V.17.4.10.), but this is often not easy to explicitly verify.)

V.17.4.1. There are two standard methods for constructing smooth functions which are not analytic anywhere, both involving infinite series. The first is to take a smooth function g which is not analytic at one, or a few, points, such as Cauchy's example, and form a sum of the form

$$f(x) = \sum_{n=1}^{\infty} c_n g(x - r_n)$$

where the r_n form a dense sequence (e.g. an enumeration of $\mathbb{Q}$) and $c_n \to 0$ rapidly, or some variation of this involving translation and scaling of g. This method is called the method of *condensation of singularities*.

The other method is to consider an infinite sum of the form

$$\sum_{n=1}^{\infty} c_n \cos(\lambda_n x)$$

(sin can be used instead of cos) where $c_n \to 0$ rapidly, $\lambda_n \to \infty$ rapidly, and there are appropriate relations between the c_n and λ_n. The first example of a function with a dense set of (P)-points was a function of this second kind due to Cellèrier in the 1860's (Exercise ()). We give a nice example of this sort in V.17.4.3.. (Caution: there is an example of this kind on Wikipedia which is not correct.) ✋

The Fabius Function

V.17.4.2. EXAMPLE. Perhaps the nicest example of a C^∞ function which is not analytic anywhere is the *Fabius function* [Fab66], which has a dense set of (C)-points. Although there is no closed form formula or even simple infinite series of functions which defines this function, there are at least two nice characterizations.

FABIUS defined the function probabilistically. Let (ξ_n) be a sequence of independent nonnegative uniformly distributed random variables on $[0,1]$ with range $[0,1]$. Then the Fabius function $F(t)$ is the distribution function for the random variable

$$\sum_{n=1}^{\infty} 2^{-n}\xi_n \ .$$

This can be described as the process of choosing a number x_1 randomly in $\left[0,\frac{1}{2}\right]$, a number $x_2 \in \left[0,\frac{1}{4}\right]$, ..., $x_n \in [0,2^{-n}]$, ..., then adding them up; $F(t)$ is the probability that $\sum_{n=1}^{\infty} x_n \le t$.

See e.g. http://www.math.osu.edu/~edgar/selfdiff/ for the graph of F.

This function can be shown to have the following properties, which give an alternate characterization:

(i) $0 \le F(t) \le 1$ for $0 \le t \le 1$; $F(0) = 0$, $F(1) = 1$, and F is strictly increasing on $[0,1]$.

(ii) $F(1-t) = 1 - F(t)$ for $0 \le t \le 1$.

(iii) F is differentiable on $[0,1]$ and $F'(t) = 2F(2t)$ for $0 \le t \le \frac{1}{2}$.

It then follows from (ii) that $F'(t) = F'(1-t) = 2F(2-2t)$ for $\frac{1}{2} \le t \le 1$.

We can then extend F to $[1,2]$ by setting $F(t) = F(2-t)$ for $1 \le t \le 2$. Extension beyond $[0,2]$ can be done by iterating, with or without reflection across the x-axis. It is customary to extend F to $[2,4]$ by setting $F(t) = -F(t-2)$ for $2 \le t \le 4$. We then successively extend F to $[0,+\infty)$ by setting $F(t) = -F(t-2^n)$ for $2^n \le t \le 2^{n+1}$. We have that $F(t) = \frac{1}{2}F'\left(\frac{t}{2}\right)$ for $0 \le t \le 2$; in particular, the extended F is differentiable at 0, 1, and 2 and $F'(0) = F'(1) = F'(2) = 0$. So F is differentiable everywhere and $F'(t) = 2F(2t)$ for all $t \in [0,+\infty)$. It follows that F is C^∞ on $[0,+\infty)$ and $F^{(n)}(t) = 2^{\sigma(n)}F(2^n t)$ for all n and t, where

$$\sigma(n) = \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \ .$$

F can be extended to be C^∞ on all of $\mathbb{R}$ by setting $F(-t) = F(t)$ or $F(-t) = -F(t)$ (there seems to be no convention). Thus the graph of $F^{(n)}$ is similar to the graph of F, stretched vertically by a factor of $2^{\sigma(n)}$ and compressed horizontally by a factor of 2^n. See Figure () for the graph of F'.

However, F is not analytic anywhere. For if $t_0 = \frac{k}{2^m}$ is a dyadic rational number, it follows that $2^m t_0 \in \mathbb{Z}$, so $F^{(n)}(t_0) = 0$ for $n > m$. Thus the Taylor series for $F^{(n)}$ around t_0 has all terms zero except possibly the constant term. This series therefore has infinite radius of convergence, but represents $F^{(n)}$ only for $t = t_0$ since $F^{(n)}$ is not constant on any interval by (i). So by working backwards, the Taylor series for F around t_0 has infinite radius of convergence (it is a polynomial), but represents F only at t_0. So every dyadic rational is a (C)-point for F, i.e. F has a dense set of (C)-points..

Smooth Functions With Only (P)-Points

The following example of a smooth function with only (P)-points is from http://mathoverflow.net/questions/43462/existence-of-a-smooth-function-with-nowhere-converging-taylor-series-at-every-poi/81465#81465. This is a specific instance of a general class of examples from [SZ55]. The first example of this kind was given by H. Cartan in 1940 [Car40].

V.17.4.3. Example. Let

$$f_n = (n!)^{-3n+2}\cos((n!)^3 x)$$

which has the following properties:

(i) $|f_n^{(k)}(x)| \le \frac{1}{n!}$ for all x if $0 \le k \le n-1$.

(ii) $[f_n^{(n)}(x)]^2 + [f_n^{(n+1)}(x)]^2 \ge (n!)^4$ for all x; hence $\max(|f_n^n(x)|, |f_n^{(n+1)}(x)|) \ge \frac{1}{2}(n!)^2$ for all x (a better constant can be obtained, but this is sufficient for the argument).

We will inductively choose a sequence (n_m) so that, if $f = \sum_{m=1}^{\infty} f_{n_m}$, f is C^∞ and the Taylor series of f around any x_0 has radius of convergence zero. We need to insure that at least some of the derivatives of f increase rapidly in size at every x.

First note that however the n_m are chosen (strictly increasing), the resulting f will be C^∞ and its derivatives will be given by successive term-by-term differentiation. For the k'th term-by-term derived series is dominated by the series

$$\sum_{n=1}^{\infty}(n!)^{-3n+2+3k}$$

which obviously converges for each k, so the k'th derived series converges uniformly on $\mathbb{R}$ by the M-test, and we can apply (). Also note that if $r > k$, we have, for all x,

$$\sum_{n=r}^{\infty}|f_n^{(k)}(x)| \le \sum_{n=r}^{\infty}\frac{1}{n!}$$

which is a tail sum for the Maclaurin series for e^1, so by Taylor's Theorem we have, for some $c < 1$,

$$\sum_{n=r}^{\infty}|f_n^{(k)}(x)| \le \frac{e^c}{r!} < \frac{e}{r!}\ .$$

Let $n_1 = 1$. Suppose $n_1, \ldots, n_r$ have been chosen. Let $g(x) = \sum_{m=1}^{r} f_{n_m}(x)$. Then the Maclaurin series for g converges to g on all of $\mathbb{R}$, and g is periodic, so there is a constant C such that $|g^{(k)}(x)| \le C(k+1)!$ for all k and all $x \in \mathbb{R}$ (). Choose $n \ge n_r + 2$ large enough that $C(n+2)! < \frac{1}{4}(n!)^2$, and set $n_{r+1} = n$. This completes the inductive construction of the n_m.

Fix r and let $k = n_r$. Set $g = \sum_{m=1}^{r-1} f_{n_m}$ and $h = \sum_{m=r+1}^{\infty} f_{n_m}$. Then $f = f_k + g + h$. We have

$$\max(|f^{(k)}(x)|, |f^{(k+1)}(x)|) \ge \max(|f_k^{(k)}(x)|, |f_k^{(k+1)}(x)|) - \max(|g^{(k)}(x)|, |g^{(k+1)}(x)|) - \max(|h^{(k)}(x)|, |h^{(k+1)}(x)|)$$

$$\geq \frac{1}{2}(k!)^2 - C(k+2)! - \frac{1}{k!} \geq \frac{1}{5}(k!)^2$$

for any x, and so it follows that, for any x_0,

$$\limsup_{k\to\infty} \left| \frac{f^{(k)}(x_0)}{k!} \right|^{1/k} \geq \lim_{k\to\infty} \left[\frac{(k!)^2}{5(k+1)!} \right]^{1/k} = \lim_{k\to\infty} \frac{(k!)^{1/k}}{5^{1/k}(k+1)^{1/k}} = +\infty$$

(Stirling's formula () is helpful in evaluating this limit.) Thus the Taylor series for f around x_0 has radius of convergence 0.

Most Smooth Functions are Nowhere Analytic

We can use the Baire Category Theorem to show that "most" C^∞ functions on an interval have only (P)-points and are thus nowhere analytic. This result was first proved by D. MORGENSTERN [Mor54a]. The result has been reproved and republished several times, sometimes in the weaker form of Corollary V.17.4.5..

V.17.4.4. THEOREM. Let I be an interval in $\mathbb{R}$. The set of all smooth functions on I for which every point of I is a (P)-point is a residual subset of $C^\infty(I)$ with its usual topology. Thus the set of smooth functions which have a Taylor series at at least one point which has a positive radius of convergence is a meager set in $C^\infty(I)$.

PROOF: We will prove the result for I closed and bounded; the proof can be adapted to the general case.

Let $\mathcal{P}$ be the set of all g in $C^\infty(I)$ such that $P_g = I$. First note that $\mathcal{P}$ is dense in $C^\infty(I)$: since the polynomials are dense in $C^\infty(I)$ (), the functions $f+\phi$, where f is the function of Example V.17.4.3. and ϕ is a polynomial, are dense in $C^\infty(I)$. But addition of a polynomial to f does not change the radius of convergence of any Taylor series of f since only finitely many terms are changed; in particular, $P_{f+\phi} = P_f = I$, so $f+\phi \in \mathcal{P}$ for any polynomial ϕ.

So we need only show that $\mathcal{P}$ is a G_δ, i.e. that $\mathcal{P}^c$ is an F_σ. Let $E_{m,n}$ be the set of all $f \in C^\infty(I)$ such that there is an $x \in I$ with

$$|f^{(k)}(x)| \leq k! \cdot m \cdot n^k$$

for all $k \in \mathbb{N}$. We show that $E_{m,n}$ is closed for each m and n and $\cup_{m,n} E_{m,n} = \mathcal{P}^c$.

To show that $E_{m,n}$ is closed, suppose $f_n \in E_{m,n}$ and $f_n \to f$ in $C^\infty(I)$. For each n there is an x_n such that $|f_n^{(k)}(x_n)| \leq k! \cdot m \cdot n^k$. Passing to a subsequence, we may assume $x_n \to x$ for some $x \in I$ since I is compact (recall that the topology of $C^\infty(I)$ is metrizable). Then $f_n^{(k)}(x_n) \to f^{(k)}(x)$ for all k, since $f_n^{(k)} \to f^{(k)}$ uniformly on I for each k (). Thus

$$|f^{(k)}(x)| = \lim_{n\to\infty} |f_n^{(k)}(x_n)| \leq k! \cdot m \cdot n^k$$

for all k, i.e. $f \in E_{m,n}$.

Now suppose $f \in E_{m,n}$. Then there is an $x_0 \in I$ such that

$$\left| \frac{f^{(k)}(x_0)}{k!} \right|^{1/k} \leq m^{1/k} n$$

for all k, so we have

$$\limsup_{k\to\infty}\left|\frac{f^{(k)}(x_0)}{k!}\right|^{1/k} \leq n$$

so the radius of convergence of the Taylor series for f around x_0 is at least $\frac{1}{n}$, and $f \notin \mathcal{P}$. Conversely, suppose $f \notin \mathcal{P}$, i.e. there is an $x_0 \in I$ where the Taylor series has positive radius of convergence. Thus there is an $n \in \mathbb{N}$ such that

$$\limsup_{k\to\infty}\left|\frac{f^{(k)}(x_0)}{k!}\right|^{1/k} < n$$

so there is a k_0 such that

$$\left|\frac{f^{(k)}(x_0)}{k!}\right|^{1/k} < n$$

for all $k > k_0$, i.e. $|f^{(k)}(x_0)| < k! \cdot n^k$ for $k \geq k_0$.

Let $m \in \mathbb{N}$, $m \geq \max(|f(x_0)|, |f'(x_0)|, \ldots, |f^{(k_0)}(x_0)|)$. Then $|f^{(k)}(x_0)| < k! \cdot m \cdot n^k$ for all k, i.e. $f \in E_{m,n}$. 👍

V.17.4.5. Corollary. Let I be an interval in $\mathbb{R}$. The set of all smooth functions on I which are nowhere analytic is a residual subset of $C^\infty(I)$ with its usual topology. Thus the set of smooth functions which are analytic at at least one point is a meager set in $C^\infty(I)$.

Can a Smooth Function Have Only (C)-Points?

V.17.4.6. We can similarly apply the Baire Category Theorem to the question of whether a nowhere analytic smooth function f on an interval I can have a Taylor series around each $x_0 \in I$ with a positive radius of convergence, i.e. whether we can have $C_f = I$. For $m, n \in \mathbb{N}$, define

$$E_{m,n}(f) = \{x_0 \in I : |f^{(k)}(x_0)| \leq k! \cdot m \cdot n^k \text{ for all } k \geq 0\} .$$

V.17.4.7. Proposition. Let $f \in C^\infty(I)$. Then each $E_{m,n}(f)$ is relatively closed in I, and $\cup_{m,n} E_{m,n}(f) = I \setminus P_f$. Thus C_f is an F_σ and P_f is a G_δ.

Proof: The proof of the first statement is a simplified form of the argument in the proof of V.17.4.4., and is left to the reader. For the second statement, note that each of the following subsets of $\mathbb{R}$ is both an F_σ and a G_δ:

I, an interval ().

A_f, a relatively open subset of I.

$C_f \cup P_f = I \setminus A_f$, a relatively closed subset of I.

$E_{m,n}(f)$, a relatively closed subset of I.

So by the first statement $I \setminus P_f$ is an F_σ, and thus $C_f = (C_f \cup P_f) \cap (I \setminus P_f)$ is an F_σ, and $P_f = I \setminus (I \setminus P_f)$ is a G_δ. 👍

Each $E_{m,n}(f)$ is contained in $A_f \cup C_f$. The key fact is that interior points of $E_{m,n}(f)$ are in A_f:

V.17.4.8. PROPOSITION. If x_0 is an interior point of $E_{m,n}(f)$ for some (m,n), then f is analytic in a neighborhood of x_0.

PROOF: Let J be an open interval around x_0 contained in $E_{m,n}(x_0)$, and let $x \in J$ with $|x - x_0| = r < \frac{1}{n}$. If p_k is the k'th Taylor polynomial of f around x_0, we have, by Taylor's Theorem,

$$f(x) - p_k(x) = \frac{f^{(k+1)}(c)}{(k+1)!}(x - x_0)^{k+1}$$

for some $c \in J$, and so we have

$$|f(x) - p_k(x)| \leq \frac{(k+1)! \cdot m \cdot n^{k+1}}{(k+1)!}|x - x_0|^{k+1} = m(nr)^{k+1} \to 0$$

and so the Taylor series for f around x_0 converges to f on $J \cap \left(x_0 - \frac{1}{n}, x_0 + \frac{1}{n}\right)$. 👍

V.17.4.9. THEOREM. Let $f \in C^\infty(I)$. Then the set C_f is a meager subset of I, and in particular has empty interior.

PROOF: We have that $C_f = \cup_{m,n}[C_f \cap E_{m,n}(f)]$, and by V.17.4.8. $C_f \cap E_{m,n}(f)$ is nowhere dense (note that $C_f \cap E_{m,n}(f) = A_f^c \cap E_{m,n}(f)$ is relatively closed in I). 👍

V.17.4.10. COROLLARY. Let $f \in C^\infty(I)$.

(i) If $P_f = \emptyset$, i.e. if the Taylor series for f around x_0 has positive radius of convergence for every $x_0 \in I$, then f is analytic on a dense open subset of I.

(ii) If f is not analytic at any point of I, then P_f is dense in I.

Note that a function satisfying (i) is not necessarily analytic everywhere: CAUCHY'S example is a counterexample. But there is a companion result to (i), first stated by PRINGSHEIM in 1893 with a faulty proof and proved by R. BOAS in 1935 in the same paper [BJ35] where V.17.4.9. was first proved:

V.17.4.11. THEOREM. Let $f \in C^\infty(I)$. If there is an $r > 0$ such the Taylor series for f around x_0 has radius of convergence at least r, for every $x_0 \in I$, then f is analytic at every point of I (i.e. analytic on I if I is open).

V.17.4.12. We give only a partial argument; for the full proof see (). Since P_f is empty, it follows from V.17.4.10. that C_f is a nowhere dense relatively closed subset of I. We only show that C_f has no isolated points. If x_0 is an isolated point of C_f, i.e. f is analytic on a deleted open interval J around x_0, and $r > 0$ is as in the statement, take $x_1 \in J$ with $x_0 < x_1 < x_0 + r$. The Taylor series for f around x_1 defines an analytic function g on $(x_1 - r, x_1 + r)$, and g agrees with f on (x_0, x_1) by V.16.2.2., since they agree on an interval around x_1. By continuity of all derivatives of f and g, the Taylor series for g around x_0 agrees with the Taylor series for f around x_0, so the Taylor series for f around x_0 converges to f on $[x_0, x_1)$ by V.16.2.7. (applied to g). A similar argument shows that the Taylor series for f around x_0 converges to f on an interval $(x_2, x_0]$ for some $x_2 < x_0$. Thus f is analytic at x_0, contradicting that $x_0 \in C_f$.

This argument does not rule out that C_f could look like a Cantor set. Showing that this is impossible requires a more delicate argument involving application of the Baire Category Theorem to the set C_f, which is a locally compact metrizable space.

Zahorski's Theorem

In 1947 Z. Zahorski [Zah47] gave a complete characterization of which subsets of I occur as C_f and P_f for a smooth function f on an interval I. We have shown that these sets have the following properties:

C_f is a meager F_σ.

P_f is a G_δ disjoint from C_f.

$C_f \cup P_f$ is relatively closed in I.

Zahorski's Theorem asserts that these conditions are sufficient as well as necessary:

V.17.4.13. Theorem. [Zahorski's Theorem] Let I be an interval in $\mathbb{R}$, and let C and P be subsets of I such that

C is a meager F_σ.

P is a G_δ disjoint from C.

$C \cup P$ is relatively closed in I.

Then there is an $f \in C^\infty(I)$ such that $C_f = C$ and $P_f = P$.

See [SZ55] for a proof which is somewhat simpler than Zahorski's, but still complicated.

V.17.4.14. Thus, although C_f is always a meager set, it can be large in a measure-theoretic sense. For example, there is a smooth function f on $\mathbb{R}$ for which P_f is the set of Liouville numbers and C_f the complement. This f is not analytic anywhere, and C_f is a set whose complement has Lebesgue measure zero, and in fact Hausdorff dimension zero (III.8.2.37.).

V.17.4.15. There is a multidimensional version of Zahorski's Theorem [**?**].

V.17.5. Exercises

V.17.5.1. Let

$$f(x) = \begin{cases} e^{-1/x^2} \sin \frac{1}{x} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases} .$$

Show that f is $\mathcal{C}^\infty$ on all of $\mathbb{R}$, and has a nonisolated zero at 0.

V.17.5.2. [?, p. 69] Let

$$f(x) = \sum_{k=1}^{\infty} \frac{\cos(k!x)}{(k!)^k} .$$

(a) Show that this series and the series of term-by-term n'th derivatives converge uniformly on $\mathbb{R}$ for every n. In particular, $f(x)$ is defined for all $x \in \mathbb{R}$, and f is periodic with period 2π.

(b) Use V.15.1.6. inductively to conclude that f is $\mathcal{C}^\infty$ on $\mathbb{R}$, and that n-fold term-by-term differentiation is justified for every n.

(c) Show that 0 is a (P)-point of f.

(d) Show that for any $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, $f\left(x + \frac{\pi p}{q}\right) - f(x)$ is analytic on $\mathbb{R}$ [the infinite series differ in only finitely many terms], and that $\frac{\pi p}{q}$ is a (P)-point of f. Thus f has a dense set of (P)-points. Conclude from V.17.1.2. that f is not analytic anywhere.

(e) Is every $x \in \mathbb{R}$ a (P)-point of f?

Chapter VI

The Topology of $\mathbb{R}$, $\mathbb{R}^n$, and Metric Spaces

In this chapter we discuss the basic notions of open and closed sets, closure of a set, interior and boundary of a set, and some of the elementary results about them, beginning with the real numbers, then moving on to $\mathbb{R}^n$ and general metric spaces. This chapter should be considered a parallel thread to Chapter () on sequences and series; there is much interplay between the two chapters.

The general theory of topology and more detailed results along the lines of this chapter are the subject of Chapter ().

VI.1. The Topology of $\mathbb{R}$

We begin with the real numbers. Recall the definition of an interval in $\mathbb{R}$ (III.1.9.12.). The open intervals and the closed intervals are most important for this section. The open intervals are the intervals of the form (a, b), where $a \in \mathbb{R} \cup \{-\infty\}$, $b \in \mathbb{R} \cup \{+\infty\}$, and $a < b$. The closed intervals are of the form $[a, b]$ with $a, b \in \mathbb{R}$, $a < b$, or $[a, +\infty)$, $(-\infty, b]$, or all of $\mathbb{R}$. We generalize both open intervals and closed intervals into open sets and closed sets respectively.

VI.1.1. Open and Closed Sets in $\mathbb{R}$

VI.1.1.1. Definition. (i) A subset U of $\mathbb{R}$ is an *open set* if, whenever $x \in U$, there is an open interval I with $x \in I$ and $I \subseteq U$.

(ii) A subset E of $\mathbb{R}$ is a *closed set* if $E^c = \mathbb{R} \setminus E$ is an open set.

A slight rephrasing of the definition gives:

VI.1.1.2. Proposition. A subset of $\mathbb{R}$ is an open set if and only if it is a union of open intervals.

VI.1.1.3. An open set is, roughly, a set which locally looks like an open interval. Any open interval is an open set. The empty set vacuously satisfies the definition of open set, hence is also an open set (it is an "empty union" of open intervals). The complement of a closed interval is either one open interval, the disjoint union of two open intervals, or the empty set, hence open; thus a closed interval is a closed set, so the terminology is consistent. The complement of the empty set is $\mathbb{R}$, which is open, so $\emptyset$ is closed as well as open. There are many nonempty closed sets which look nothing like closed intervals: for example, any finite set is closed since its complement is a union of open intervals; similarly $\mathbb{Z}$ is a closed set. There are much more complicated examples; see e.g. (), ().

A half-open interval of the form $[a, b)$ for $a < b$ is not an open set since it does not contain an open interval around a; similarly, its complement is not open, so $[a, b)$ is not a closed set either. $\mathbb{Q}$ is also neither open nor closed, since every open interval contains both rational and irrational numbers. In fact, "most" subsets of $\mathbb{R}$ are neither open nor closed.

The only subsets of $\mathbb{R}$ which are both open and closed are $\emptyset$ and $\mathbb{R}$ ().

Here is a useful rephrasing of the definition of open set:

VI.1.1.4. PROPOSITION. Let U be a subset of $\mathbb{R}$. Then U is an open set if and only if, for every $x \in U$, there is an $\epsilon > 0$ such that the open interval $(x - \epsilon, x + \epsilon) = \{y \in \mathbb{R} : |y - x| < \epsilon\}$ is contained in U. (The ϵ can depend on x.)

PROOF: Any such set is obviously open, since $(x - \epsilon, x + \epsilon)$ is an open interval. Conversely, if U is open and $x \in U$, and I is an open interval with $x \in I$ and $I \subseteq U$, then there is a smaller open interval J centered at x contained in I (and hence in U); J will be of the form $(x - \epsilon, x + \epsilon)$ for some $\epsilon > 0$. 👍

Unions and Intersections

VI.1.1.5. PROPOSITION. (i) An arbitrary union of open sets is open. A *finite* intersection of open sets is open.

(ii) An arbitrary intersection of closed sets is closed. A *finite* union of closed sets is closed.

PROOF: (i): It is obvious from VI.1.1.2. (or just straight from the definition) that a union of any collection of open sets is open. For a finite intersection, we use VI.1.1.4.. Let $U_1, \ldots, U_n$ be a finite collection of open sets, and $U = U_1 \cap \cdots \cap U_n$. Let $x \in U$. Then $x \in U_j$ for each j, so for each j there is an $\epsilon_j > 0$ such that $(x - \epsilon_j, x + \epsilon_j) \subseteq U_j$. Set $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_n\}$. Then $\epsilon > 0$, and $(x - \epsilon, x + \epsilon) \subseteq U$. Since this can be done for any $x \in U$, U is open.

(ii) follows from (i) by taking complements and using de Morgan's Laws (). 👍

VI.1.1.6. The intersection of an infinite number of open sets is not open in general. For example, $\cap_{n=1}^{\infty} \left(x - \frac{1}{n}, x + \frac{1}{n}\right) = \{x\}$, which is not open (in fact it is closed). Similarly, an infinite union of closed sets is not closed in general. In fact, every subset of $\mathbb{R}$ is a union of singleton sets, each of which is closed.

VI.1.1.7. COROLLARY. Let U be an open set and E a closed set. Then $U \setminus E$ is open and $E \setminus U$ is closed.

PROOF: $U \setminus E = U \cap E^c$ and $E \setminus U = E \cap U^c$. 👍

Neighborhoods and Deleted Neighborhoods

VI.1.1.8. Let x be a real number. An open interval containing x, or more generally an open set containing x, is often called a *neighborhood*, or *open neighborhood*, of x. In topology, a neighborhood of x is generally thought of as a set of points "close to x"; this is literally true if the neighborhood is something like $(x-\epsilon, x+\epsilon)$ for a small positive ϵ, but may be somewhat of a stretch for general neighborhoods. Nonetheless, the notion of neighborhood is a useful one, as we will see. An open set is an open neighborhood of each of its points.

Sometimes we use the term "neighborhood" to mean a not necessarily open set: a neighborhood of x is a set containing an open neighborhood of x. For example, it is sometimes useful to consider closed neighborhoods (e.g. ()).

VI.1.1.9. It is also often useful to consider deleted neighborhoods of a point. If $x \in \mathbb{R}$, a *deleted neighborhood* of x is a subset of $\mathbb{R}$ of the form $U \setminus \{x\}$, where U is a neighborhood of x. Deleted neighborhoods are usually taken to be open, but not necessarily.

The most commonly used deleted neighborhoods of an $x \in \mathbb{R}$ are sets of the form

$$(x-\epsilon, x) \cup (x, x+\epsilon) = \{y \in \mathbb{R} : 0 < |y-x| < \epsilon\}$$

for some (typically "small") $\epsilon > 0$. Every deleted neighborhood of x contains a deleted neighborhood of this form.

VI.1.2. Limit Points and Closures

There is an alternate way to view closed sets, via the notions of limit points and closures, which we now describe.

VI.1.2.1. DEFINITION. Let A be a subset of $\mathbb{R}$, and $x \in \mathbb{R}$. Then x is a *limit point* of A if every open neighborhood of x contains at least one point of A.

The word "open" may be deleted from the definition without effect.

VI.1.2.2. Any point of A is a limit point of A: if $x \in A$, then any neighborhood of x contains at least one point of A, namely x itself. But the importance of the definition is that there may be real numbers which are not in A but are limit points of A. For example, $0 \notin (0,1)$, but 0 is a limit point of $(0,1)$ since every neighborhood of 0 contains small positive numbers. For a more dramatic example, every real number is a limit point of $\mathbb{Q}$, since every nonempty open set in $\mathbb{R}$ contains rational numbers.

Some authors define "limit point" to mean what is more commonly called an *accumulation point* of A. The number x is an accumulation point of A if every neighborhood of x contains a point of A other than x, i.e. every deleted neighborhood of x contains a point of A. Not every element of A is necessarily an accumulation point of A; those which are not are called *isolated points* of A. We will not need to make much use of the notion of accumulation point; limit points are more important for most purposes.

As with open sets, the notion of limit point can be phrased in terms of ϵ's:

VI.1.2.3. Proposition. Let $A \subseteq \mathbb{R}$, $x \in \mathbb{R}$. Then x is a limit point of A if and only if, for every $\epsilon > 0$, there is a $y \in A$ with $|x - y| < \epsilon$.

Intuitively, x is a limit point of A if and only if there are points of A arbitrarily close to x.

VI.1.2.4. Definition. Let A be a subset of $\mathbb{R}$. The *closure* of A is the set $\bar{A}$ of all limit points of A.

VI.1.2.5. Since every element of A is a limit point of A, we have $A \subseteq \bar{A}$. If $A \subseteq B$, then it follows immediately from the definition of limit point that every limit point of A is a limit point of B, so $\bar{A} \subseteq \bar{B}$.

The term "closure" is justified by the next result.

VI.1.2.6. Theorem. Let A be a subset of $\mathbb{R}$. Then

(i) $\bar{\bar{A}} = \bar{A}$, i.e. every limit point of $\bar{A}$ is in $\bar{A}$.

(ii) $\bar{A}$ is a closed set.

(iii) $\bar{A}$ is the smallest closed set containing A, i.e. if E is a closed set containing A, then $\bar{A} \subseteq E$.

Proof: (i): Let x be a limit point of $\bar{A}$, and U an open neighborhood of x. Then there is a $y \in \bar{A}$ with $y \in U$. Then y is a limit point of A, and U is a neighborhood of y, so there is a $z \in A$ with $z \in U$. This is true for every open neighborhood U of x, so x is a limit point of A.

(ii): Let $x \in (\bar{A})^c$. Then x is not a limit point of A, so there is an open neighborhood U of x containing no points of A. Then U also cannot contain any limit points of A, for if $y \in U$ were a limit point of A, then the neighborhood U of y would have to contain a point of A. Thus $U \subseteq (\bar{A})^c$. Since such a U exists for every $x \in (\bar{A})^c$, $(\bar{A})^c$ is open.

(iii): There is a smallest closed set F containing A, namely the intersection of all closed sets containing A (cf. VI.1.1.5.; there is at least one closed set containing A, $\mathbb{R}$ itself). Since $\bar{A}$ is a closed set containing A by (ii), we have $F \subseteq \bar{A}$. To show the opposite inclusion, let $x \notin F$. Since F is closed, F^c is an open neighborhood of x containing no point of A. Thus x cannot be a limit point of A. 👍

VI.1.3. Compactness and Continuity

VI.1.3.1. A collection $\mathcal{U} = \{U_i : i \in \Omega\}$ of open sets in $\mathbb{R}$ (or in a general topological space) is an *open cover* of a set A if $A \subseteq \cup_i U_i$. A *subcover* of $\mathcal{U}$ for A is a subset of $\mathcal{U}$ which is also a cover of A. A subset F of $\mathbb{R}$ is *compact* if every open cover of F has a finite subcover.

VI.1.3.2. LEMMA. A closed bounded interval $[a,b]$ in $\mathbb{R}$ is compact.

PROOF: Let $\mathcal{U} = \{U_i : i \in \Omega\}$ be an open cover of $[a,b]$. Let S be the set of all $c \in [a,b]$ such that there is a finite subcover of $\mathcal{U}$ for $[a,c]$. S is nonempty since $a \in S$ ($a \in U$ for some $U \in \mathcal{U}$, so $\{U\}$ is a finite subcover of $[a,a] = \{a\}$). S is bounded since $S \subseteq [a,b]$.
Let $d = \sup(S)$. Then $d \in U_0$ for some $U_0 \in \mathcal{U}$. There is an $\epsilon > 0$ such that $(d-\epsilon, d+\epsilon) \subseteq U_0$ since U_0 is open. Since d is the least upper bound of S, there is a $c \in S$ with $c > d - \epsilon$. Let $\{U_1, \ldots, U_n\}$ be a finite subcover of $\mathcal{U}$ for $[a,c]$. Then $\{U_0, U_1, \ldots, U_n\}$ is a finite subcover of $[a,d]$, so $d \in S$. Suppose $d < b$. Then if $d < x < \min(d+\epsilon, b)$, then $\{U_0, U_1, \ldots, U_n\}$ is a finite subcover of $[a,x]$, contradicting that $d = \sup(S)$. Thus $d = b$ and $b \in S$. 👍

VI.1.3.3. COROLLARY. Every closed bounded subset of $\mathbb{R}$ is compact.

PROOF: If F is closed and contained in $[a,b]$, and $\mathcal{U}$ is an open cover of F, then $\mathcal{U} \cup \{F^c\}$ is an open cover of $[a,b]$, and this open cover has a finite subcover $\mathcal{V}$ for $[a,b]$. $\mathcal{V} \setminus \{F^c\} \subseteq \mathcal{U}$ is then a finite subcover of $\mathcal{U}$ for F. 👍

Virtually the same proof shows that any closed subset of a compact set is compact. The converse is also true:

VI.1.3.4. PROPOSITION. A compact subset of $\mathbb{R}$ is closed and bounded.

PROOF: Let E be a compact subset of $\mathbb{R}$. If $U_m = (-m, m)$, then $\{U_m : m \in \mathbb{N}\}$ is an open cover of E, so there is a finite subcover $\{U_{m_1}, \ldots, U_{m_n}\}$. If $M = \max\{m_1, \ldots, m_n\}$, then $E \subseteq (-M, M)$. To see that E is closed, let $c \in E^c$, and for each $\epsilon > 0$ let $U_\epsilon = \{x \in \mathbb{R} : |x - c| > \epsilon\}$. Then $\{U_\epsilon : \epsilon > 0\}$ is an open cover of E, so there is a finite subcover $\{U_{\epsilon_1}, \ldots, U_{\epsilon_n}\}$. If $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_n\}$, then $(c-\epsilon, c+\epsilon) \subseteq E^c$, so E^c is open. 👍

Combining these two results, we obtain one of the most important theorems of analysis:

VI.1.3.5. THEOREM. [HEINE-BOREL THEOREM FOR $\mathbb{R}$] A subset of $\mathbb{R}$ is compact if and only if it is closed and bounded.

The Heine-Borel Theorem gives a generalization of the Nested Interval Theorem:

VI.1.3.6. COROLLARY. Let (F_m) be a decreasing sequence of nonempty closed bounded subsets of $\mathbb{R}$. Then $\cap_m F_m$ is not empty.

PROOF: Suppose $F_1 \subseteq [a,b]$. If $\cap F_m = \emptyset$, then $\{F_m^c : m \in \mathbb{N}\}$ is an open cover of $[a,b]$, which has a finite subcover $\{F_{m_1}, \ldots, F_{m_n}\}$. If $M = \max\{m_1, \ldots, m_n\}$, then $F_M = F_{m_1} \cap \cdots \cap F_{m_n} = \emptyset$, a contradiction. 👍

From this, we get the important Bolzano-Weierstrass Theorem (see IV.1.7.1. for a different proof):

THEOREM. [BOLZANO-WEIERSTRASS THEOREM] Every bounded sequence in $\mathbb{R}$ has a convergent subsequence.

PROOF: Let (x_n) be a bounded sequence in $[a,b] \subseteq \mathbb{R}$. It suffices to show that (x_n) has a cluster point (). For each $m \in \mathbb{N}$, let F_m be the closure of $\{x_n : n \geq m\}$. Then F_m is a nonempty closed subset of $\mathbb{R}$, and $F_{m+1} \subseteq F_m$ and $F_m \subseteq [a,b]$ for all m. Thus $\cap_m F_m$ is nonempty, and any $c \in \cap_m F_m$ is a cluster point of (x_n). 👍

We then get some basic theorems about continuous functions as consequences:

VI.1.3.7. THEOREM. Let E be a compact subset of $\mathbb{R}$, and let $f : E \to \mathbb{R}$ be continuous. Then $f(E)$ is compact.

PROOF: We first show that $f(E)$ is bounded. Suppose for each n there is an $x_n \in E$ with $|f(x_n)| \geq n$. Since E is bounded, (x_n) has a subsequence (x_{k_n}) converging to some $y \in \mathbb{R}$, and $y \in E$ since E is closed. Because f is continuous, the sequence $(f(x_{k_n}))$ converges to $f(y)$ by the Sequential Criterion. But this is impossible since $(f(x_{k_n}))$ is unbounded by construction, a contradiction.

Now suppose that c is in the closure of $f(E)$, and choose $x_n \in E$ such that $f(x_n) \to c$. Again by the Bolzano-Weierstrass Theorem, (x_n) has a subsequence (x_{k_n}) converging to some $y \in E$. But then $(f(x_{k_n}))$ converges to $f(y)$ by the Sequential Criterion; since $(f(x_{k_n}))$ converges to c, we have $c = f(y) \in f(E)$ and $f(E)$ is closed. 👍

See () for a slicker proof.

VI.1.3.8. COROLLARY. [MAX-MIN THEOREM] Let E be a closed bounded subset of $\mathbb{R}$, and $f : E \to \mathbb{R}$ a continuous function. Then f is bounded on E and attains its maximum and minimum on E, i.e. there are $c, d \in E$ such that $f(c) \leq f(x) \leq f(d)$ for all $x \in E$.

For the proof, just note that $f(E)$ is compact and that a compact subset of $\mathbb{R}$ has a maximum and a minimum element.

See () for a more complete discussion of compactness.

The next theorem is a variation of the Tietze Extension Theorem of topology (). It is certainly not true that a continuous function on a subset E of $\mathbb{R}$ can be extended to a continuous function on all of $\mathbb{R}$; there are many counterexamples, such as $f(x) = \frac{1}{x}$ on $(0, \infty)$. But if E is closed, there is always an extension:

VI.1.3.9. THEOREM. [TIETZE EXTENSION THEOREM FOR $\mathbb{R}$] Let E be a closed subset of $\mathbb{R}$, and $f : E \to \mathbb{R}$ a continuous function. Then f can be extended to a continuous function on all of $\mathbb{R}$, i.e. there is a continuous function $g : \mathbb{R} \to \mathbb{R}$ which agrees with f on E.

PROOF: The definition of g is very simple. Let $U = E^c$; then U is open, so it consists of a union of countably many disjoint open intervals $I_n = (a_n, b_n)$ and possibly an unbounded interval of the form $(-\infty, a)$ (if E has a smallest element a) and/or an unbounded interval (b, ∞) (if E has a largest element b.) Define g to be

equal to f on E, linear on $[a_n, b_n]$ for each n, and if necessary constant on $(-\infty, a]$ and/or $[b, \infty)$. If $x \in I_n$, the formula for $g(x)$ is

$$g(x) = f(a_n) + \frac{f(b_n) - f(a_n)}{b_n - a_n}(x - a_n)\,.$$

It is geometrically clear that g is continuous everywhere, but takes a little work to write down a careful proof. Suppose $c \in \mathbb{R}$; we will show g is continuous at c. If $c \notin E$, this is obvious since g is linear in a neighborhood of c. Suppose $c \in E$. We will show that $\lim_{x \uparrow c} f(x) = f(c)$. Let $\epsilon > 0$ be given. If $c = b_n$ for some n, let $\delta = \min(b_n - a_n, \frac{b_n - a_n}{|f(b_n) - f(a_n)|}\epsilon)$ (if $f(a_n) = f(b_n)$, set $\delta = b_n - a_n$.) Then if $c - \delta < x < c$, g is linear on $[x, c]$, so $|f(x) - f(c)| < \epsilon$.

Now suppose that c is not one of the b_n. For $\epsilon > 0$, choose $\delta_1 > 0$ such that $c - \delta_1 < x < c$ and $x \in E$ imply $|f(x) - f(c)| < \epsilon$ (using the fact that f is continuous on E.) Since c is not one of the b_n, there is a $d \in E$ with $c - \delta_1 < d < c$. Set $\delta = c - d$. If $c - \delta < x < c$, we will show that $|f(x) - f(c)| < \epsilon$. This is obvious if $x \in E$ since $\delta < \delta_1$; and if $x \in U$, then there is an n such that $d \le a_n < x < b_n < c$, so $f(x)$ is between $f(a_n)$ and $f(b_n)$, each of which is in $(f(c) - \epsilon, f(c) + \epsilon)$, so $|f(x) - f(c)| < \epsilon$. This proves that g is continuous from the left; the proof that g is continuous from the right is almost identical. 👍

VI.1.3.10. None of the arguments in this section require any form of the Axiom of Choice.

VI.2. Metric Spaces

Much of what was done in the last section works almost verbatim in the more general setting of metric spaces. Besides Euclidean spaces, there are many other metric spaces which are important in analysis, as well as in other parts of mathematics, so it is quite worthwhile working in this generality.

VI.2.1. Definitions

A *metric space* is a set X with a specified numerical distance between any two points, which behaves reasonably like the usual distance between numbers or points in Euclidean space. The distance may be regarded as a function $\rho : X \times X \to \mathbb{R}$, with $\rho(x, y)$ denoting the distance between x and y. Formally:

VI.2.1.1. DEFINITION. A *metric* on a set X is a function $\rho : X \times X \to \mathbb{R}$ satisfying:

(i) $\rho(x, y) \geq 0$ for all $x, y \in X$.

(ii) $\rho(x, y) = 0$ if and only if $x = y$ (*definiteness*).

(iii) $\rho(x, y) = \rho(y, x)$ for all $x, y \in X$ (*symmetry*).

(iv) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ for all $x, y, z \in X$ (*triangle inequality*).

A pair (X, ρ), where X is a set and ρ is a metric on X, is called a *metric space*.
The *diameter* of a metric space (X, ρ) is $diam(X, \rho) = \sup\{\rho(x, y) : x, y \in X\}$. The metric ρ is *bounded* if $diam(X, \rho)$ is finite.

VI.2.1.2. It is customary to call a set X with a given metric a "space", and to call elements of X "points", even though there may be no geometric picture or interpretation of X as an actual space in the usual sense. The geometric terminology helps to motivate some aspects of the theory of metric spaces.

VI.2.1.3. EXAMPLES.

(i) On $\mathbb{R}$, the function $\rho(x, y) = |x - y|$ is a metric. This is called the *standard metric* on $\mathbb{R}$, and we will often use this metric on $\mathbb{R}$ without specific mention.

(ii) On $\mathbb{R}^n$, the Euclidean distance function

$$\rho_2(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

for $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ is a metric, the *standard metric* on $\mathbb{R}^n$.

(iii) There are other important metrics on $\mathbb{R}^n$. Two of them are:

$$\rho_1(x, y) = |x_1 - y_1| + \cdots + |x_n - y_n|$$

$$\rho_\infty(x, y) = \max(|x_1 - y_1|, \ldots, |x_n - y_n|)$$

for $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$.

(iv) There are important metrics on function spaces analogous to those of (ii) and (iii). For example, let $C[a,b]$ be the set of real-valued continuous functions on the interval $[a,b]$, and for $f,g \in C[a,b]$ define

$$\rho_2(f,g) = \left[\int_a^b (f(t)-g(t))^2\,dt\right]^{1/2}$$

$$\rho_1(f,g) = \int_a^b |f(t)-g(t)|\,dt$$

$$\rho_\infty(f,g) = \max_{a\le t\le b} |f(t)-g(t)|$$

(It is not obvious that ρ_2 satisfies the triangle inequality. See Exercise () or () for a proof.)

(v) Let X be a sphere (recall that a *sphere* in mathematics is a two-dimensional surface, not including the points "inside"). If $x,y \in X$, let $\rho(x,y)$ be the arc length of the shortest great circle path between x and y. (The triangle inequality is not entirely obvious here, but can be proved.) This example has obvious practical applications: it is the appropriate "distance" to consider on the surface of the earth.

(vi) Let X be any set, $\rho(x,y) = \begin{cases} 1 & \text{if} \quad x \neq y \\ 0 & \text{if} \quad x = y \end{cases}$. Then ρ is a metric on X, called the *discrete metric*.

(vii) Let X be the set of all n-tuples of 0's and 1's, and define $\rho(x,y)$ to be the number of coordinates in which x and y differ. This is a metric, called the *Hamming metric*, which is important in coding theory.

In these examples, only the metrics in (v)–(vii) are bounded.

VI.2.1.4. A "distance function" ρ on a set X satisfying (i), (iii), and (iv) is called a *pseudometric*. Pseudometrics play an important role in more advanced parts of analysis.

VI.2.2. Balls, Limit Points, and Closures

VI.2.2.1. DEFINITION. Let (X,ρ) be a metric space, and $x_0 \in X$. If $r > 0$, the *(open) ball of radius r around x_0* (or *centered at x_0*) is

$$B_r^\rho(x_0) = \{x \in X : \rho(x_0,x) < r\}$$

(the ρ is often omitted if the metric is understood and there is no possibility of confusion.) The *closed ball of radius r around x_0* is

$$\bar{B}_r^\rho(x_0) = \{x \in X : \rho(x_0,x) \le r\}.$$

VI.2.2.2. The term "ball" is a standard term even though $B_r(x_0)$ need not look anything like a geometric ball. (In fact, there is usually no geometric interpretation of distance in a general metric space; see Exercise VI.2.8.6., for example, for interpretations of ρ_1 and ρ_∞ on $\mathbb{R}^2$.)

The term "ball" without the qualification "closed" will always mean "open ball".

VI.2.2.3. EXAMPLES.

(i) In $\mathbb{R}$ with the usual metric, the open ball of radius r around x_0 is the open interval of length $2r$ centered at x_0; $\bar{B}_r(x_0)$ is the closed interval of length $2r$ centered at x_0.

(ii) In $(\mathbb{R}^2, \rho_2)$, $B_r(x_0)$ is an open disk centered at x_0. In $(\mathbb{R}^3, \rho_2)$, $B_r(x_0)$ is the usual geometric open ball of radius r centered at x_0. The closed balls have a similar interpretation.

(iii) In $(\mathbb{R}^2, \rho_1)$, balls (open or closed) are square diamonds; in $(\mathbb{R}^2, \rho_\infty)$ they are squares with horizontal and vertical sides (see Exercise ().)

(iv) There is a geometric interpretation of balls in $C[a,b]$ for the ρ_∞ metric. Fix $f_0 \in C[a,b]$, and $r > 0$. Draw the graphs of the functions $f(t) - r$ and $f(t) + r$ (Figure ().) Then $B_r^{\rho_\infty}(f_0)$ consists of all functions in $C[a,b]$ whose graph lies within the strip strictly between the top and bottom curves. $\bar{B}_r^{\rho_\infty}(f_0)$ consists of the functions whose graphs lie in the closed strip including the two boundary functions.

(v) If ρ is the discrete metric on a set X, and $x_0 \in X$, then $B_r^\rho(x_0) = \{x_0\}$ if $r \leq 1$, and $B_r^\rho(x_0) = X$ if $r > 1$. $\bar{B}_r^\rho(x_0) = \{x_0\}$ if $r < 1$ and $\bar{B}_r^\rho(x_0) = X$ if $r \geq 1$. Thus $B_1^\rho(x_0) = \{x_0\}$ but $\bar{B}_1^\rho(x_0) = X$, so $\bar{B}_r(x_0)$ can be much larger than the "closure" of $B_r(x_0)$ in general (this will be made precise in ().)

VI.2.2.4. DEFINITION. Let (X, ρ) be a metric space, and $A \subseteq X$. A point $x \in X$ is a *limit point* (or *point of closure*) of A (with respect to ρ) if every ball centered at x contains at least one point of A, i.e. if, for every $\epsilon > 0$, there is a point $y \in A$ with $\rho(x,y) < \epsilon$. The *closure* of A (with respect to ρ), denoted $\bar{A}$, is the set of all limit points of A. The set A is *closed* if $A = \bar{A}$.

When there is any possibility of confusion about which metric is being used, we will use the terms *ρ-limit point* and *ρ-closed set*.

VI.2.2.5. Note that a limit point of A is not an element of A in general: for example, 0 and 1 are limit points of $(0,1)$ in $\mathbb{R}$. But a point of A is always a limit point of A, so $A \subseteq \bar{A}$ for any A. If $A \subseteq B$, then $\bar{A} \subseteq \bar{B}$. We also have that $\bar{\emptyset} = \emptyset$ and $\bar{X} = X$, i.e. $\emptyset$ and X are closed sets in X.

VI.2.2.6. PROPOSITION. Let (X, ρ) be a metric space, and $A \subseteq X$. Then every limit point of $\bar{A}$ is a limit point of A, i.e. $\bar{\bar{A}} = \bar{A}$, so $\bar{A}$ is closed.

PROOF: Suppose x is a limit point of $\bar{A}$, and $\epsilon > 0$. We must show that there is a $z \in A$ with $\rho(x,z) < \epsilon$. But x is a limit point of $\bar{A}$, so there is a $y \in \bar{A}$ with $\rho(x,y) < \epsilon/2$; and there is a $z \in A$ with $\rho(y,z) < \epsilon/2$. By the triangle inequality, $\rho(x,z) < \epsilon$. 👍

The complement of a ρ-closed set is called an *open set* (with respect to ρ, or *ρ-open set*). The next proposition gives two other useful characterizations of open sets:

VI.2.2.7. PROPOSITION. Let (X, ρ) be a metric space, and $U \subseteq X$. Then the following are equivalent:

(i) U is an open set.

(ii) U is a union of open balls.

(iii) For every $x \in U$, there is an $\epsilon > 0$ such that $B_\epsilon(x) \subseteq U$.

PROOF: (i) $\Rightarrow$ (iii): If $x \in U$, then x is not a limit point of U^c since U^c is closed. Thus there is an $\epsilon > 0$ such that $B_\epsilon(x)$ contains no point of U^c, i.e. $B_\epsilon(x) \subseteq U$.
(iii) $\Rightarrow$ (i): If $x \in U$, then there is an $\epsilon > 0$ such that $B_\epsilon(x) \subseteq U$, hence $B_\epsilon(x)$ contains no point of U^c and x is not a limit point of U^c. Thus U^c is closed.
(ii) $\Rightarrow$ (iii): Suppose $x \in U$. Then $x \in B_r(y) \subseteq U$ for some r and y, since U is a union of open balls. Set $\epsilon = r - \rho(x, y)$. Then by the triangle inequality, $B_\epsilon(x) \subseteq B_r(y) \subseteq U$.
(iii) $\Rightarrow$ (ii) is obvious. 👍

In particular, every open ball is an open set.

VI.2.2.8. COROLLARY. An arbitrary union of open sets is open. A finite intersection of open sets is open.
An arbitrary intersection of closed sets is closed. A finite union of closed sets is closed.

PROOF: It is obvious from either (ii) or (iii) that an arbitrary union of open sets is open. To prove that a finite intersection of open sets is open, use (iii): if $U_1, \ldots, U_n$ are open and $x \in U_1 \cap \cdots \cap U_n$, choose $\epsilon_1, \ldots, \epsilon_n > 0$ with $B_{\epsilon_k}(x) \subseteq U_k$ for each k. Set $\epsilon = \min(\epsilon_1, \ldots, \epsilon_n)$; then $\epsilon > 0$ and $B_\epsilon(x) \subseteq U_1 \cap \cdots \cap U_n$.
The statements about unions and intersections of closed sets are obtained by taking complements. 👍

VI.2.2.9. Note that an intersection of infinitely many open sets is not open in general: for example, $\cap_{n=1}^{\infty} B_{1/n}(x) = \{x\}$, which is rarely an open set ($\{x\}$ is actually a closed set for any x.) Similarly, an arbitrary union of closed sets is not closed in general: in fact, every subset of X is a union of singleton subsets, which are closed.

VI.2.2.10. It is important to note that not every subset of a metric space is either open or closed; in fact, this is far from true in most metric spaces. ✋

> "Most sets, like doors, are neither open nor closed, but ajar."
>
> *C. Pugh*[1]

In fact, the "most" can be precisely quantified: in $\mathbb{R}$ (and in most other metric spaces encountered in analysis), there are $2^{2^{\aleph_0}}$ subsets in all, but only $2^{\aleph_0}$ open sets and $2^{\aleph_0}$ closed sets (), so there are $2^{2^{\aleph_0}}$ subsets which are "ajar." Although "most" of these subsets are rather bizarre and cannot be even explicitly described, many frequently encountered and easily described sets are ajar, such as half-open intervals of the form $(a, b]$ or $[a, b)$, or the rational numbers $\mathbb{Q}$.

Interior and boundary points

[1] [Pug15, p. 60].

VI.2.3. Limits and Continuity

Convergence of sequences and limits and continuity of functions work in any metric space almost exactly as in $\mathbb{R}$. The statements and proofs are essentially identical to the case of $\mathbb{R}$; all that is needed is to "redecorate" using the metric distance in place of the absolute value. We will do a few examples of redecoration in proofs, and leave others as exercises.

Limits of Sequences

VI.2.3.1. DEFINITION. Let (X, ρ) be a metric space, $x \in X$, and (x_n) a sequence in X. Then (x_n) *converges to* x (with respect to ρ), or x is the *limit* of the sequence (x_n), written $x_n \to x$ or $x = \lim_{n\to\infty} x_n$, if for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that $\rho(x_n, x) < \epsilon$ for all $n \geq n_0$.

Where it is necessary to explicitly specify the metric ρ, or where there is any chance of confusion, we will say that (x_n) *ρ-converges* to x, or x is the *ρ-limit* of (x_n), written $x_n \to_\rho x$ or $x = \rho\text{-}\lim_{n\to\infty} x_n$.

As in $\mathbb{R}$, limits of sequences are unique:

VI.2.3.2. PROPOSITION. Let (X, ρ) be a metric space, $x, y \in X$, and (x_n) a sequence in X. If $x_n \to x$ and $x_n \to y$, then $x = y$.

PROOF: If $x \neq y$, set $\epsilon = \rho(x, y)/3 > 0$, and choose $n_1, n_2 \in \mathbb{N}$ with $\rho(x_n, x) < \epsilon$ for all $n \geq n_1$ and $\rho(x_n, y) < \epsilon$ for all $n \geq n_2$. Then, if $n \geq \max(n_1, n_2)$,

$$\rho(x, y) \leq \rho(x, x_n) + \rho(x_n, y) < 2\epsilon = 2\rho(x, y)/3 < \rho(x, y)$$

which is a contradiction. 👍

VI.2.3.3. PROPOSITION. Let (X, ρ) be a metric space, $x \in X$, and (x_n) a sequence in X. If $x_n \to x$, then every subsequence of (x_n) also converges to x.

There is also a notion of a cluster point of a sequence in a metric space:

VI.2.3.4. DEFINITION. Let (X, ρ) be a metric space, $x \in X$, and (x_n) a sequence in X. Then x is a *cluster point* of the sequence (x_n) if for every $\epsilon > 0$ and $n_0 \in \mathbb{N}$, $\rho(x_n, x) < \epsilon$ for some $n \geq n_0$. (Equivalently, $\rho(x_n, x) < \epsilon$ for infinitely many n.) When ρ needs to be specified, we say x is a *ρ-cluster point* of (x_n).

If $x_n \to x$, then x is a cluster point of (x_n), but the converse is generally not true. A sequence can have many cluster points.

VI.2.3.5. The term "limit point" is sometimes used in place of "cluster point", but this can lead to confusion, so we will use the term "limit point" only in reference to a *set*, and "cluster point" in reference to a *sequence*. If x is a cluster point of a sequence (x_n), then x is a limit point of the set $\{x_n : n \in \mathbb{N}\}$, but the converse is not necessarily true: each term x_k is a limit point of $\{x_n : n \in \mathbb{N}\}$, but x_k is usually not a cluster point of (x_n). In fact, we have:

VI.2.3.6. PROPOSITION. Let (x_n) be a sequence in a metric space (X, ρ), and let $x \in X$. Then x is a cluster point of (x_n) if and only if x is a limit point of

$$A_k = \{x_n : n \geq k\}$$

for every $k \in \mathbb{N}$.

This proposition is an immediate consequence of the definitions, and the proof is left as an exercise.

Cluster points can also be characterized using subsequences:

VI.2.3.7. PROPOSITION. Let (x_n) be a sequence in a metric space (X, ρ), and let $x \in X$. Then x is a cluster point of (x_n) if and only if there is a subsequence (x_{k_n}) which converges to x.

PROOF: Suppose there is a subsequence of (x_n) converging to x. Then, for any $\epsilon > 0$, infinitely many terms of the subsequence lie within ϵ of x, so x is a cluster point of (x_n). Conversely, if x is a cluster point of (x_n), then set $k_1 = 1$ and inductively let k_{n+1} be an integer $\geq k_n + 1$ for which $\rho(x_{k_{n+1}}, x) < 1/(n+1)$. Then $x_{k_n} \to x$. 👍

A related result about subsequences has a somewhat complicated statement, but is frequently useful:

VI.2.3.8. PROPOSITION. Let (X, ρ) be a metric space, $x \in X$, and (x_n) a sequence in X. Then $x_n \to x$ if and only if x is a cluster point of every subsequence of (x_n).

PROOF: If $x_n \to x$, then every subsequence (x_{k_n}) of (x_n) also converges to x (), hence x is a cluster point of (x_{k_n}). Conversely, suppose (x_n) does *not* converge to x. Then there is an $\epsilon > 0$ such that, for every $n_0 \in \mathbb{N}$, there is an $n \geq n_0$ with $\rho(x_n, x) \geq \epsilon$. Let $k_1 = 1$, and inductively choose $k_{n+1} \geq k_n + 1$ with $\rho(x_{k_{n+1}}, x) \geq \epsilon$. Then x is not a cluster point of the subsequence (x_{k_n}). 👍

Limits of sequences and functions, continuity, sequential criterion, dependence on metric; cluster points; subsequences; closure in terms of sequences; uniform continuity; mean and mean-square convergence

Dominant and Equivalent Metrics

It is sometimes important to compare two or more metrics on the same set. The next definition is reminiscent of the definition of continuity, and is in fact very closely related ().

VI.2.3.9. DEFINITION. Let ρ and σ be metrics on the same set X. Then ρ *dominates* σ if, for every $x_0 \in X$ and $\epsilon > 0$, there is a $\delta > 0$ such that $\sigma(x, x_0) < \epsilon$ whenever $x \in X$ and $\rho(x, x_0) < \delta$. The metric ρ *uniformly dominates* σ if, for every $\epsilon > 0$, there is a $\delta > 0$ such that $\sigma(x, y) < \epsilon$ whenever $x, y \in X$ and $\rho(x, y) < \delta$.

The metrics ρ and σ are *equivalent* if each dominates the other, and *uniformly equivalent* if each uniformly dominates the other.

The metrics ρ and σ are *incomparable* if neither one dominates the other.

VI.2.3.10. Roughly speaking, "ρ dominates σ" means that $\sigma(x,y)$ is small whenever $\rho(x,y)$ is small. The most common way a metric ρ uniformly dominates another metric σ is if there is a $C > 0$ with $\sigma(x,y) \leq C\rho(x,y)$ for all $x, y \in X$.

VI.2.3.11. PROPOSITION. Let ρ and σ be metrics on a set X. Then the following are equivalent:

(i) ρ dominates σ.

(ii) The identity function from (X,ρ) to (X,σ) is continuous.

(iii) Whenever (x_n) is a sequence in X converging to a limit x with respect to ρ (i.e. $\lim_{n\to\infty} \rho(x_n, x) = 0$), (x_n) also converges to x with respect to σ (i.e. $\lim_{n\to\infty} \sigma(x_n, x) = 0$).

(iv) Every set which is σ-closed is also ρ-closed.

(v) Every set which is σ-open is also ρ-open.

Paraphrasing (iii)–(v), if ρ dominates σ, it is *easier* for a set to be ρ-open (or ρ-closed) than σ-open (or σ-closed), but *harder* for a sequence to converge with respect to ρ than with respect to σ.
PROOF: (i) $\iff$ (ii) is a straightforward comparison of definitions, and is left to the reader. (ii) $\iff$ (iii) is the Sequential Criterion for continuity.
(iii) $\Rightarrow$ (iv): Suppose $A \subseteq X$ is σ-closed, and x is a ρ-limit point of A. We must show that $x \in A$. But by the Sequential Criterion for limit points, there is a sequence (x_n) in A with $x_n \to x$ with respect to ρ. By (iii), $x_n \to x$ with respect to σ, so $x \in A$ because A is σ-closed.
(iv) $\iff$ (v) by taking complements. (v) $\Rightarrow$ (ii) by **??**. 👍

Similarly, we obtain:

VI.2.3.12. PROPOSITION. Let ρ and σ be metrics on a set X. Then ρ uniformly dominates σ if and only if the identity function from (X,ρ) to (X,σ) is uniformly continuous.

VI.2.3.13. EXAMPLES.

(i) On $\mathbb{R}^n$, the metrics ρ_2, ρ_1, and ρ_∞ are uniformly equivalent since, for any $x, y \in \mathbb{R}^n$,

$$\rho_\infty(x,y) \leq \rho_2(x,y) \leq \rho_1(x,y) \leq n\rho_\infty(x,y) \leq n\rho_2(x,y)$$

(the inequality $\rho_2(x,y) \leq \rho_1(x,y)$ is the only one which is not obvious; see Exercise ().)

(ii) On $C[a,b]$, ρ_∞ uniformly dominates ρ_1 and ρ_2 since $\rho_1(f,g) \leq (b-a)\rho_\infty(f,g)$ and $\rho_2(f,g) \leq (b-a)^{1/2}\rho_\infty(f,g)$ for any f, g. Actually, $\rho_1(f,g) \leq (b-a)^{1/2}\rho_2(f,g)$ for any $f, g \in C[a,b]$ (Exercise ()), so ρ_2 also uniformly dominates ρ_1.

However, none of the three are equivalent. For example, let f_n be the function whose graph is in Figure ().

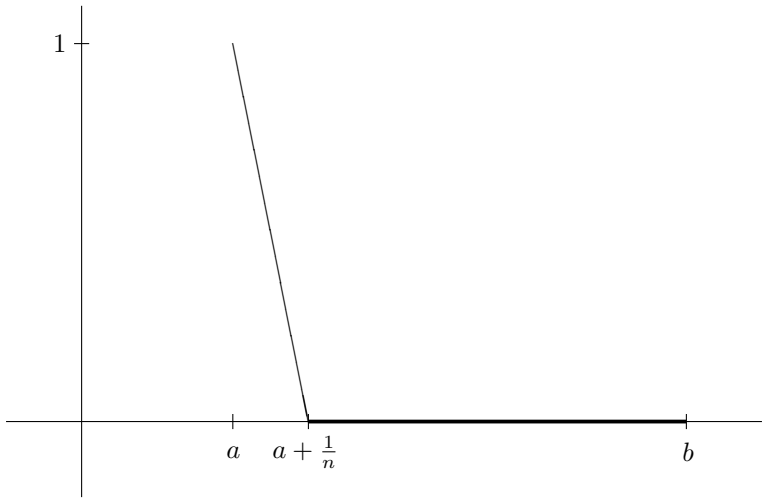


Figure VI.1: The graph of f_n

Then

$$\lim_{n\to\infty} \rho_1(f_n, 0) = \lim_{n\to\infty} \rho_2(f_n, 0) = 0,$$

but $\rho_\infty(f_n, 0) = 1$ for all n. Apply the criterion of VI.2.3.11.(iii) to conclude that ρ_1 and ρ_2 do not dominate ρ_∞. See Exercise () for a proof that ρ_1 does not dominate ρ_2.

(iii) Let ρ be a metric on X, and for $x, y \in X$, set $\sigma(x, y) = \min(\rho(x, y), 1)$. Then σ is a bounded metric on X which is uniformly equivalent to ρ [it is obvious that ρ uniformly dominates σ; to prove that σ uniformly dominates ρ, for any ϵ choose $\delta = \min(\epsilon, 1)$.]

(iv) Let ρ be the usual metric on $\mathbb{R}$, $\rho(x, y) = |x - y|$, and let $\sigma(x, y) = |\arctan x - \arctan y|$. Then σ is a metric on $\mathbb{R}$, and ρ uniformly dominates σ (in fact, $\sigma(x, y) \leq \rho(x, y)$ for all $x, y \in \mathbb{R}$.) Conversely, σ dominates ρ by continuity of the tangent function, but not uniformly since $\sigma(n, n + 1) \to 0$ as $n \to \infty$; so ρ and σ are equivalent but not uniformly equivalent.

(v) The discrete metric on a set X uniformly dominates any other metric on X [for any $\epsilon > 0$ we can choose $\delta = 1$.]

(vi) If ρ is a metric on X, then $c\rho$ is a metric which is uniformly equivalent to ρ, for any $c > 0$.

(vii) If ρ and σ are metrics on a set X, then $\rho + \sigma$ (defined by $[\rho + \sigma](x, y) = \rho(x, y) + \sigma(x, y)$) and $\max(\rho, \sigma)$ are uniformly equivalent metrics which uniformly dominate both ρ and σ. If (ρ_n) is a sequence of metrics on X, set $\sigma_n = \min(\rho_n, 1)$ for all n (cf. (iii)), and define ρ by

$$\rho(x, y) = \sum_{n=1}^{\infty} 2^{-n} \sigma_n(x, y).$$

Then ρ is a metric which uniformly dominates all the ρ_n's.

VI.2.4. Continuity Without the Countable AC

The proof () that the Sequential Criterion for continuity is equivalent to the $(\epsilon-\delta)$-definition uses the Countable AC. We can show without Choice that for functions between metric spaces, the Sequential Criterion for continuity is equivalent to the $(\epsilon-\delta)$-definition if the source space is separable.

VI.2.4.1. DEFINITION. (i) Let X and Y be topological spaces, $f: X \to Y$ a function, and $x_0 \in X$. Then f is *sequentially continuous at* x_0 if, whenever $x_n \to x_0$ in X, $f(x_n) \to f(x_0)$ in Y.

(ii) Let (X,ρ) and (Y,σ) be metric spaces, $f: X \to Y$ a function, and $x_0 \in X$. Then f is *$(\epsilon-\delta)$-continuous at* x_0 if for every $\epsilon > 0$ there is a $\delta > 0$ such that $\sigma(f(x), f(x_0)) < \epsilon$ whenever $x \in X$ and $\rho(x, x_0) < \delta$.

(iii) f is *sequentially continuous* [resp. *$(\epsilon-\delta)$-continuous*] if it is sequentially continuous [resp. $(\epsilon-\delta)$-continuous] at x_0 for each $x_0 \in X$.

A function between metric spaces is $(\epsilon-\delta)$-continuous if and only if it is continuous in the topological sense (); a function which is $(\epsilon-\delta)$-continuous at x_0 is sequentially continuous at x_0 (no Choice is needed for this).

VI.2.4.2. THEOREM. Let (X,ρ) and (Y,σ) be metric spaces, with (X,ρ) separable. Let $f: X \to Y$ be a function. Then f is $(\epsilon-\delta)$-continuous if and only if it is sequentially continuous.

PROOF: One direction has already been observed. Conversely, suppose f is sequentially continuous. Fix a countable dense set D in X and arrange it into a sequence $S = \{s_1, s_2, \dots\}$. If f is not $(\epsilon-\delta)$-continuous at x_0, there is an $\epsilon > 0$ such that, for any n, there is a $p \in X$ with $\rho(p, x_0) < \frac{1}{n}$ and $\sigma(f(p), f(x_0)) > \epsilon$. Fix n. There is a sequence (p_r) in D with $p_r \to p$ (such a sequence can be generated algorithmically if desired). Then by assumption $f(p_r) \to f(p)$, so sufficiently large r we have $\rho(p_r, x_0) < \frac{1}{n}$ and $\sigma(f(p_r), f(x_0)) > \epsilon$, i.e. there is a $q \in D$ with $\rho(q, x_0) < \frac{1}{n}$ and $\sigma(f(q), f(x_0)) > \epsilon$.

For each n let z_n be the first s_k for which $\rho(s_k, x_0) < \frac{1}{n}$ and $\sigma(f(s_k), f(x_0)) > \epsilon$. Then $z_n \to x_0$ but $f(z_n) \not\to f(x_0)$, contradicting that f is sequentially continuous. 👍

VI.2.4.3. This result does not extend to pointwise continuity, or to limits (). In the absence of separability, the result can fail without the Countable AC, even if X and Y are compact. Let (X,ρ) and $(X^\dagger,\bar{\rho})$ be as in XIII.15.12.8., and define $f: X^\dagger \to [0,1]$ by setting f identically 0 on X and $f(\infty) = 1$. Then f is sequentially continuous, but not $(\epsilon-\delta)$-continuous at ∞.

VI.2.4.4. The result applies to functions from $\mathbb{R}^n$ to any metric space, or to functions defined on an open or closed subset of $\mathbb{R}^n$ (cf. XIII.15.9.11.). But one cannot prove without the Countable AC that arbitrary subsets of $\mathbb{R}$ are separable, or even that an infinite subset of $\mathbb{R}$ contains a sequence of distinct points ().

VI.2.5. Extended Metrics

It turns out to be quite harmless to allow a metric to also take the value ∞ $(+\infty)$, i.e. to allow points to be an infinite distance apart.

VI.2.5.1. DEFINITION. An *extended metric* on a set X is a function $\rho : X \times X \to [0, \infty]$ satisfying:

(i) $\rho(x, y) \geq 0$ for all $x, y \in X$.

(ii) $\rho(x, y) = 0$ if and only if $x = y$ (*definiteness*).

(iii) $\rho(x, y) = \rho(y, x)$ for all $x, y \in X$ (*symmetry*).

(iv) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ for all $x, y, z \in X$ (*triangle inequality*).

A pair (X, ρ), where X is a set and ρ is a metric on X, is called an *extended metric space*.

In the definition, we take the usual order on $[0, \infty]$, i.e. $x < \infty$ for all $x \in \mathbb{R}$, and $x + \infty = \infty$ for any $x \in [0, \infty]$. Note that any metric is an extended metric. We may also consider *extended pseudometrics* which have all the properties except (ii).

VI.2.5.2. All the results of this section hold for extended metrics with no modifications. If ρ is an extended metric on X, define $\tilde{\rho}$ by $\tilde{\rho}(x, y) = \min(\rho(x, y), 1)$; then $\tilde{\rho}$ is a metric on X which is equivalent to ρ, so if desired any extended metric can be converted to a metric without changing the topology.

VI.2.5.3. DEFINITION. Let (X, ρ) be an extended metric space, and $x \in X$. The *galaxy* of x in X is

$$\{y \in X : \rho(x, y) < \infty\} .$$

The proof of the next simple proposition is left as an exercise for the reader.

VI.2.5.4. PROPOSITION. Let (X, ρ) be an extended metric space. Then

(i) Any two distinct galaxies in X are disjoint. Thus X partitions into a disjoint union of galaxies. Any points in distinct galaxies are an infinite distance apart.

(ii) Each galaxy is a clopen set in X.

(iii) As a topological space, X is the separated union () of its galaxies.

(iv) The restriction of ρ to any galaxy is a metric.

A metric space is just an extended metric space with only one galaxy.

Maximal Metrics

The next simple but important observation is most naturally phrased using extended metrics:

VI.2.5.5. PROPOSITION. Let X be a set, $\{\rho_i : i \in I\}$ a collection of extended pseudometrics on X. For $x, y \in X$, set

$$\bar{\rho}_I(x,y) = \sup_{i\in I} \rho_i(x,y) \ .$$

Then $\bar{\rho}_I$ is an extended pseudometric on X.

PROOF: The only thing that needs to be checked is the triangle inequality. If $x, y, z \in X$, then, by (),

$$\bar{\rho}_I(x,z) = \sup_i \rho_i(x,z) \leq \sup_i[\rho_i(x,y) + \rho_i(y,z)] \leq \sup_i \rho_i(x,y) + \sup_i \rho_i(y,z) = \bar{\rho}_I(x,y) + \bar{\rho}_I(y,z) \ .$$



VI.2.5.6. If the ρ_i are separating (in particular, if at least one ρ_i is an extended metric), then $\bar{\rho}_I$ is an extended metric. The metric $\bar{\rho}_I$ dominates each ρ_i, i.e. the $\bar{\rho}_I$ topology is stronger than all the ρ_i topologies, but it might not be the weakest topology stronger than all the ρ_i topologies. For a simple example, let σ be a metric on X, and for each n let $\rho_n = n\sigma$. Then all the ρ_n are equivalent to σ, but $\bar{\rho}_{\mathbb{N}}$ is the discrete extended metric. (For an actual metric version, set $\rho_n(x,y) = \min(1, n\sigma(x,y))$ for $x, y \in X$.) If I is finite, then the $\bar{\rho}_I$ topology is indeed the weakest topology stronger than all the ρ_i topologies.



VI.2.5.7. A particularly important example (cf. XIII.8.4.2.) comes from a bound function. Let X be a set, and $\beta : X \times X \to [0, +\infty]$ a function. Let I be the set of all pseudometrics σ on X such that $\sigma(x,y) \leq \beta(x,y)$ for all $x, y \in X$ (I is nonempty since the zero pseudometric is in I; one could also allow extended pseudometrics without affecting the construction). Set $\rho_\beta = \bar{\rho}_I$. The extended pseudometric ρ_β is called the *maximal pseudometric* of β; it is the largest pseudometric ρ satisfying $\rho(x,y) \leq \beta(x,y)$ for all $x, y \in X$.

VI.2.6. Cauchy Sequences

VI.2.7. Normed Vector Spaces

Many of the important metric spaces in analysis are vector spaces in which the metric comes from a norm.

VI.2.7.1. DEFINITION. Let X be a vector space over $\mathbb{R}$ or $\mathbb{C}$. A *norm* on X is a function $x \mapsto \|x\|$ from X to $\mathbb{R}$ satisfying

(i) $\|x\| \geq 0$ for all $x \in X$.

(ii) $\|x\| = 0$ if and only if $x = 0$ (*definiteness*).

(iii) $\|\alpha x\| = |\alpha|\|x\|$ for all $x \in X$ and all scalars α (*homogeneity*).

(iv) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in X$ (*triangle inequality*).

VI.2.7.2. A norm on a vector space X defines a metric on X by $\rho(x,y) = \|x - y\|$. The metrics in VI.2.1.3.(i)–(iv) arise this way:

VI.2.7.3. EXAMPLES.

(i) The usual metric on $\mathbb{R}$ comes from the norm $\|x\| = |x|$.

(ii) The metric ρ_2 on $\mathbb{R}^n$ comes from $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$.

(iii) The metrics ρ_1 and ρ_∞ on $\mathbb{R}^n$ also come from norms:

$$\|x\|_1 = |x_1| + \cdots + |x_n|$$

$$\|x\|_\infty = \max(|x_1|, \ldots, |x_n|)$$

(iv) The metrics ρ_2, ρ_1, ρ_∞ on $C([a,b])$ come from norms

$$\|f\|_2 = \left[\int_a^b (f(t))^2\, dt\right]^{1/2}$$

$$\|f\|_1 = \int_a^b |f(t)|\, dt$$

$$\|f\|_\infty = \max_{a \le t \le b} |f(t)|$$

(See Exercise () or () for a proof that $\|\cdot\|_2$ satisfies the triangle inequality.)

(v) On $\mathbb{C}^n$, the same formulas as in (ii) and (iii) define norms, except that the 2-norm must be written

$$\|x\|_2 = \sqrt{|x_1|^2 + \cdots + |x_n|^2}\ .$$

VI.2.7.4. The metric ρ induced by a norm is *translation-invariant*, i.e. $\rho(x+z, y+z) = \rho(x,y)$ for all $x, y, z \in X$, and *homogeneous*, i.e. $\rho(\alpha x, \alpha y) = |\alpha|\rho(x,y)$ for all $x, y \in X$ and all scalars α. A consequence of translation-invariance is that $B_r(z) = z + B_r(0)$ for any $r > 0$, $z \in X$, and $B_{\alpha r}(0) = \alpha B_r(0)$ for any $\alpha, r > 0$. Thus all balls in a normed vector space "look the same" up to scaling and translation.

Another consequence of homogeneity is:

VI.2.7.5. PROPOSITION. Let X be a vector space, $\|\cdot\|$ and $\|\cdot\|'$ norms on X, ρ and σ the associated metrics. Then the following are equivalent:

(i) ρ dominates σ.

(ii) ρ uniformly dominates σ.

(iii) There is a $C > 0$ such that $\|x\|' \le C\|x\|$ for all $x \in X$.

This is actually a special case of a fundamental principle about operators between normed vector spaces ().

PROOF: (iii) ⇒ (ii) ⇒ (i) are trivial, so we need only prove (i) ⇒ (iii). Choose $\delta > 0$ so that $\|x\| = \rho(x, 0) < \delta$ implies $\|x\|' = \sigma(x, 0) < 1$, for any $x \in X$, and set $C = 1/\delta$. Then $\|x\| < 1$ implies that $\|x\|' < C$. If $x \in X$ and $\|x\|' > C\|x\|$, choose $\alpha \in \mathbb{R}$ with

$$\frac{C}{\|x\|'} < \alpha < \frac{1}{\|x\|}$$

and then $\|\alpha x\| < 1$ but $\|\alpha x\|' > C$, a contradiction; so there can be no such x, proving (iii). 👍

[Move to later section]

THEOREM. Let X and Y be vector spaces (over the same field, $\mathbb{R}$ or $\mathbb{C}$), with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ respectively, and let $T : X \to Y$ be a linear operator. Then the following are equivalent:

(i) T is continuous at 0.

(ii) T is continuous.

(iii) T is uniformly continuous.

(iv) There is a constant $K \geq 0$ such that $\|Tx\|_Y \leq K\|x\|_X$ for all $x \in X$.

If T satisfies (iv), T is said to be *bounded*; the smallest such K is called the *operator norm* of T, denoted $\|T\|_{op}$ or usually just $\|T\|$.

Seminorms

Banach Spaces

VI.2.8. Exercises

VI.2.8.6. Here are physical interpretations of ρ_1 and ρ_∞ on $\mathbb{R}^2$.
(a) The metric ρ_1 is often called the *city metric* or *driving metric*. Imagine a city in which all streets run north-south or east-west. Then the ρ_1-distance between two points located on streets is the shortest driving distance between the points (ignoring one-way streets, etc.)
(b) There is a similar, though perhaps less satisfactory, interpretation of ρ_∞. In this city, there are buses on each street. The fare structure is unusual: the cost of riding a bus is strictly proportional to the distance traveled, say \$1 per mile. However, a rider also receives a transfer good for a free ride on a bus in the perpendicular direction, of length not exceeding the length of the paid ride. The ρ_∞ distance between two points on streets is then the minimum cost of a bus ride between the two points.

VI.2.8.7. Show that ρ_1, ρ_2, and ρ_∞ are equivalent metrics on $\mathbb{R}^n$.

VI.2.8.8. (a) If $x_1, \ldots, x_n$ are nonnegative real numbers, show that

$$\sum_{k=1}^{n} x_k^2 \leq \left[\sum_{k=1}^{n} x_k\right]^2.$$

[Hint: Multiply out the right side.] Give a careful proof by induction on n.

(b) If x_1, x_2 are real numbers, show that $(x_1 + x_2)^2 \le 2(x_1^2 + x_2^2)$. [Hint: $(x_1 - x_2)^2 \ge 0$.]
(c) If $x_1, \ldots, x_n$ are arbitrary real numbers, and $n = 2^m$, show by induction on m that

$$\left[\sum_{k=1}^{n} x_k\right]^2 \le n \sum_{k=1}^{n} x_k^2.$$

(d) Let f be Riemann integrable on $[a, b]$ (e.g. $f \in C[a, b]$.) Show that

$$\left[\int_a^b |f(t)|\, dt\right]^2 \le (b - a) \int_a^b [f(t)]^2\, dt\ .$$

[Hint: approximate the integrals by Riemann sums, dividing the interval into 2^m subintervals of equal length, and apply (c).]
(e) Show that the inequality in (c) is also true if n is not a power of 2. [Given $x_1, \ldots, x_n$, let f be the function on $[0, n]$ with $f(t) = x_k$ if $k - 1 \le x < k$. Apply (d).]

The inequalities in (b)–(e) are special cases of Jensen's Inequality. See also VI.2.8.11. for an alternate proof of (c) for general n.

VI.2.8.9. (a) Show that on $C[a, b]$, the ∞-norm dominates the 2-norm, and the 2-norm dominates the 1-norm.
(b) Let g_n be the function whose graph is given in Figure VI.2, and $h_n = \sqrt{g_n}$.

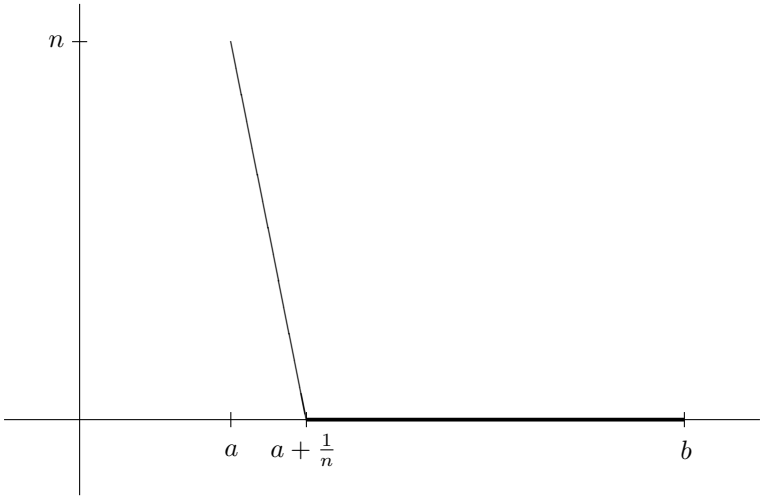


Figure VI.2: The graph of g_n

Then $\rho_1(h_n, 0) < n^{-1/2}$, so $\lim_{n\to\infty} \rho_1(h_n, 0) \to 0$. However, for any n,

$$\rho_2(h_n, 0) = \left[\int_a^b g_n(t)\, dt\right]^{1/2} = \frac{1}{\sqrt{2}}$$

so $\rho_2(h_n, 0) \not\to 0$.

VI.2.8.10. Give a direct proof (not involving taking complements) that the collection of closed sets in a metric space is closed under arbitrary intersections and under finite unions.

The graph of the Cantor function is given in Figure VI.4.

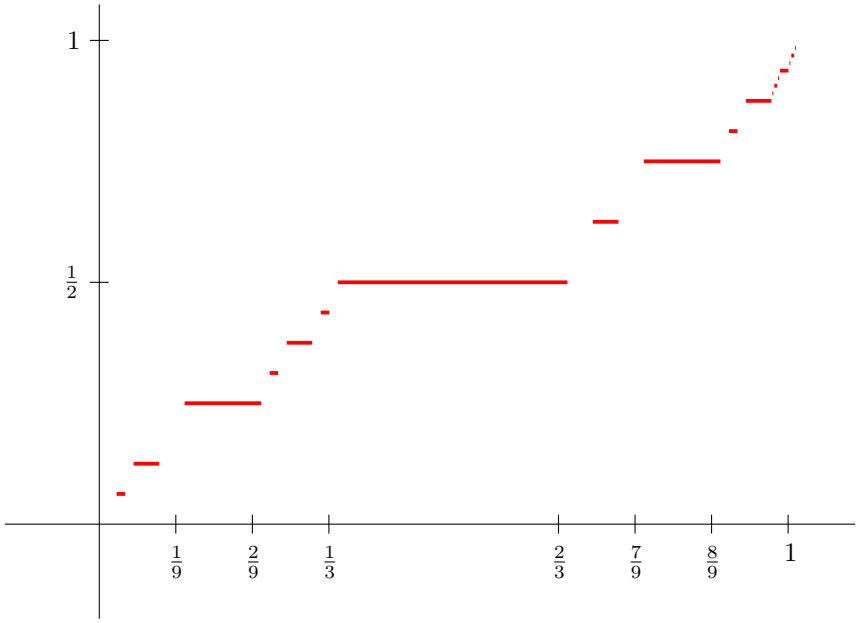


Figure VI.3: The Cantor function

[Move to later section]

VI.2.8.11. Let $x_1, \ldots, x_n, y_1, \ldots, y_n$ be real numbers. Show that

$$(x_1 + \cdots + x_n)(y_1 + \cdots + y_n) \leq n\sqrt{x_1^2 + \cdots + x_n^2}\sqrt{y_1^2 + \cdots + y_n^2} \ .$$

[Hint: let $\mathbf{x} = (x_1, \ldots, x_n)$, and for $1 \leq k \leq n$ let

$$\mathbf{y}^{(k)} = (y_k, y_{k+1}, \ldots, y_n, y_1, \ldots, y_{k-1}) \ .$$

Then $(x_1 + \cdots + x_n)(y_1 + \cdots + y_n) = \mathbf{x} \cdot \mathbf{y}^{(1)} + \cdots + \mathbf{x} \cdot \mathbf{y}^{(n)}$. Apply the CBS inequality.] In particular, $(x_1 + \cdots + x_n)^2 \leq n(x_1^2 + \cdots + x_n^2)$.
The n in this inequality is the best possible constant in general: if all the x_i and y_j are equal, we have equality.

VI.2.8.12. Adapt the proof of Theorem VI.2.4.2. to the case where X is separable metric and Y is T_3. Can the separation assumption on Y be further weakened?

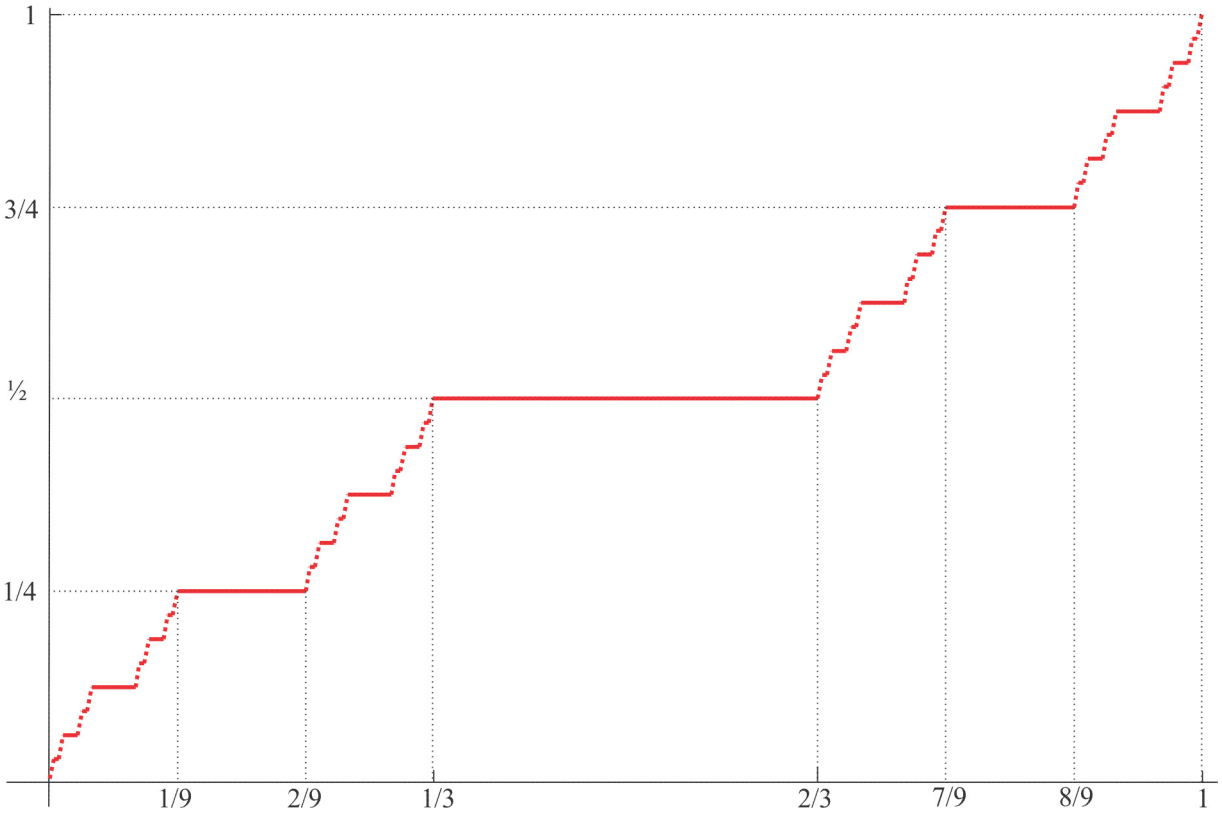


Figure VI.4: The Cantor Function

VI.2.9. How Many Separable Metric Spaces Are There?

In this subsection, we answer the following two questions:

VI.2.9.1. Question. How many separable metric spaces are there?

VI.2.9.2. Question. How many separable complete metric spaces are there?

VI.2.9.3. These questions are meaningless unless we specify a notion of equivalence. After all, there are as many one-point metric spaces as there are sets! (For a similar reason, it is meaningless to ask how many metric spaces there are, even up to isometry: any set with the discrete metric is a (complete) metric space.)

The two obvious notions of equivalence are homeomorphism and isometry. Isometry is a much finer equivalence: each homeomorphism class of metric spaces (with more than one point) includes many isometry classes. Thus we will obtain, and answer, four questions by inserting the phrase "up to homeomorphism" or "up to isometry" in each. The answers, however, will turn out not to depend on which equivalence we take: the answer to VI.2.9.1. in both cases is $2^{2^{\aleph_0}}$, and to VI.2.9.2. in both cases is $2^{\aleph_0}$.

VI.2.9.4. Proposition. There are exactly $2^{\aleph_0}$ functions from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{R}$.

VI.2.9.5. Corollary. There are at most $2^{\aleph_0}$ metrics on $\mathbb{N}$. Hence up to isometry there are at most $2^{\aleph_0}$ countable metric spaces.

In fact, it is easily seen that there are at least $2^{\aleph_0}$ metrics on $\mathbb{N}$ (the distance between 1 and 2 can be any positive number), hence exactly $2^{\aleph_0}$. (One should also consider finite metric spaces: there are clearly exactly $2^{\aleph_0}$ metrics on a finite set with more than one element.)

VI.2.9.6. Corollary. Up to isometry, there are at most $2^{\aleph_0}$ separable complete metric spaces.

Proof: Every separable complete metric space is isometric to the completion of a countable metric space, and the completion of a metric space is unique up to isometry. 👍

VI.2.9.7. There are $2^{\aleph_0}$ topologically distinct compact subsets of $\mathbb{R}^2$ (XIII.19.5.1.; there are even $2^{\aleph_0}$ topologically distinct compact subsets of $\mathbb{R}$ – see XIII.19.5.4.), each of which is a separable complete metric space. Thus up to homeomorphism, there are at least $2^{\aleph_0}$ compact metric spaces, each of which is separable and complete (). Thus we obtain the answer to Question VI.2.9.2.:

VI.2.9.8. THEOREM. (i) Up to homeomorphism, there are exactly $2^{\aleph_0}$ compact metric spaces.
(ii) Up to isometry, there are exactly $2^{\aleph_0}$ compact metric spaces.
(iii) Up to homeomorphism, there are exactly $2^{\aleph_0}$ separable complete metric spaces.
(iv) Up to isometry, there are exactly $2^{\aleph_0}$ separable complete metric spaces.

Recall () that a *Polish space* is a topological space which is homeomorphic to a separable complete metric space. Thus we can rephrase (iii):

VI.2.9.9. COROLLARY. Up to homeomorphism, there are exactly $2^{\aleph_0}$ Polish spaces.

We can now answer Question VI.2.9.1.:

VI.2.9.10. THEOREM. (i) Up to homeomorphism, there are exactly $2^{2^{\aleph_0}}$ separable metric spaces.
(ii) Up to isometry, there are exactly $2^{2^{\aleph_0}}$ separable metric spaces.

PROOF: Every separable metric space is isometric to a subset of a separable complete metric space. Every separable complete metric space has cardinality $\leq 2^{\aleph_0}$ (XIII.7.10.6.(c)) (in fact, exactly $2^{\aleph_0}$ if uncountable; cf. XVI.3.1.10.). Thus a separable complete metric space has at most $2^{2^{\aleph_0}}$ subsets. Since there are $2^{\aleph_0}$ separable complete metric spaces up to isometry (VI.2.9.8.), there are up to isometry at most $2^{\aleph_0} \cdot 2^{2^{\aleph_0}} = 2^{2^{\aleph_0}}$ subsets of complete metric spaces. Hence there are at most $2^{2^{\aleph_0}}$ separable metric spaces up to isometry.

Conversely, there are $2^{2^{\aleph_0}}$ topologically distinct subsets of $\mathbb{R}$ (XIII.19.5.45.), each of which is a separable metric space. Thus there are at least $2^{2^{\aleph_0}}$ separable metric spaces up to homeomorphism. 👍

VI.2.10. The Banach Fixed-Point Theorem

There are many fixed-point theorems in topology, which assert that under certain conditions a function from a topological space to itself sends at least one, or exactly one, point to itself. Perhaps the simplest example is (); other examples are XIII.21.1.2., (), ... Here we discuss a fairly elementary but very useful example, the Banach fixed-point theorem, sometimes called the *Contraction Mapping Theorem*.

VI.2.10.1. DEFINITION. Let (X, ρ) be a metric space. A function $T : X \to X$ is a *strict contraction* if there is an α, $0 < \alpha < 1$, such that $\rho(T(x), T(y)) \leq \alpha\rho(x, y)$ for all $x, y \in X$.

A strict contraction is obviously a (uniformly) continuous function from X to X.

VI.2.10.2. **Caution:** For T to be a strict contraction, it is not sufficient that $\rho(T(x), T(y)) < \rho(x, y)$ whenever $x \neq y$; the distance must decrease by at least a constant factor less than 1. See VI.2.10.6.. ✋

VI.2.10.3. THEOREM. [BANACH FIXED-POINT THEOREM] Let (X, ρ) be a complete metric space, and $T : X \to X$ a strict contraction. Then T has a unique fixed point: there is a unique point $p \in X$ with $T(p) = p$. If x is any point of X, then the sequence $(T^n(x))$ converges to p.

PROOF: Fix $x \in X$. Inductively set $x_0 = x$ and $x_{n+1} = T(x_n)$ for all n. We claim (x_n) is a Cauchy sequence. By induction, we have $\rho(x_n, x_{n+1}) \leq \alpha^n \rho(x_0, x_1)$ for all n. Thus, if $n < m$, by the triangle inequality we have

$$\rho(x_n, x_m) \leq \rho(x_0, x_1) \sum_{k=n}^{m-1} \alpha^k \ .$$

The sum is a tail sum of the convergent geometric series $\sum_{k=1}^{\infty} \alpha^k$, hence is arbitrarily small if n is large enough. Thus (x_n) is a Cauchy sequence and therefore converges to a point $p \in X$. Then $T(x_n) \to T(p)$ since T is continuous; but $T(x_n) = x_{n+1}$, so $T(x_n) \to p$, i.e. $T(p) = p$ by uniqueness of limits in a metric space.
If q is any fixed point of T, then

$$\rho(p, q) = \rho(T(p), T(q)) \leq \alpha\rho(p, q)$$

which can only hold if $\rho(p, q) = 0$, i.e. $q = p$. Thus p is the only fixed point of T. 👍

VI.2.10.4. The proof of the theorem is constructive: the sequence of iterates of any starting point converge to the fixed point, i.e. the fixed point is a "universal attractor." Thus the fixed point can be found by an explicit iterative algorithm of successive approximation.

VI.2.10.5. If (X, ρ) is not complete, a strict contraction on X need not have a fixed point (VI.2.11.1.). However, *if* it does, the argument of the proof shows that the fixed point is unique (this part of the proof does not require completeness). And if p is a fixed point of T, then for any $x \in X$ we have $\rho(T^n(x), p) = \rho(T^n(x), T^n(p)) \leq \alpha^n \rho(x, p) \to 0$, so $T^n(x) \to p$. Alternatively, any strict contraction on (X, ρ) extends uniquely to a strict contraction $\bar{T}$ on the completion $(\bar{X}, \bar{\rho})$ of (X, ρ) by uniform continuity (); if T has a fixed point, it must be the unique fixed point of $\bar{T}$ and we have $T^n(x) \to p$ for any $x \in X$.

VI.2.10.6. EXAMPLE. To obtain a fixed point, it does not suffice to simply have $\rho(T(x),T(y)) < \rho(x,y)$ for $x \neq y$; this condition is not enough to insure that $(T^n(x))$ is a Cauchy sequence for any x. For example, let $T : [1,+\infty) \to [1,+\infty)$ be defined by $T(x) = x + \frac{1}{x}$. Then T is differentiable everywhere and $0 < T'(x) < 1$ for $1 < x < +\infty$, so by the MVT $|T(x)-T(y)| < |x-y|$ for $x, y \in [1,+\infty)$, $x \neq y$ (this also has an elementary proof not requiring calculus). But T has no fixed point.

Continuity of Fixed Points

If two strict contractions are uniformly close, their fixed points must be close:

VI.2.10.7. PROPOSITION. Let (X,ρ) be a metric space, and S, T strict contractions on X, i.e. there are $0 \leq \alpha, \beta < 1$ such that $\rho(S(x),S(y)) \leq \alpha\rho(x,y)$ and $\rho(T(x),T(y)) \leq \beta\rho(x,y)$ for all $x, y \in X$. Let p and q be (unique) fixed points of S and T respectively. If S and T are uniformly close within $\epsilon > 0$, i.e. $\rho(S(x),T(x)) < \epsilon$ for all $x \in X$, then $\rho(p,q) < \frac{\epsilon}{1-\gamma}$, where $\gamma = \min(\alpha,\beta)$.

PROOF: We have $\lim_{n\to\infty} S^n(q) = p$, and by the triangle inequality we have, for each n,

$$\rho(q,S^n(q)) = \rho(q,S(q)) + \rho(S(q),S^2(q)) + \cdots + \rho(S^{(n-1)}(q),S^n(q))$$

$$\leq (1 + \alpha + \cdots + \alpha^{n-1})\rho(q,S(q)) = \frac{1-\alpha^n}{1-\alpha}\rho(q,S(q)) < \frac{1}{1-\alpha}\rho(q,S(q))$$

and thus, letting $k \to \infty$,

$$\rho(q,p) \leq \frac{1}{1-\alpha}\rho(q,S(q)) = \frac{1}{1-\alpha}\rho(T(q),S(q)) < \frac{\epsilon}{1-\alpha}\ .$$

Similarly, we have $\rho(p,q) = \lim_{n\to\infty} \rho(p,T^n(p)) < \frac{\epsilon}{1-\beta}$. 👍

Strict Power Contractions and the Extended Banach Fixed-Point Theorem

There is a substantial extension of the Banach Fixed-Point Theorem.

VI.2.10.8. DEFINITION. Let (X,ρ) be a metric space. A function $T : X \to X$ is a *strict power contraction* if T^n is a strict contraction for some $n \in \mathbb{N}$.

VI.2.10.9. EXAMPLE. A strict power contraction is not necessarily a strict contraction. It is not necessarily even continuous! Let $X = \mathbb{R}$ with its usual metric, and define $T : \mathbb{R} \to \mathbb{R}$ by

$$T(x) = \begin{cases} x/2 & \text{if} \quad x \in \mathbb{Q} \\ -x/2 & \text{if} \quad x \notin \mathbb{Q} \end{cases}\ .$$

Then T is not continuous, but T^2 is a strict contraction, so T is a strict power contraction.
Note that T has a unique fixed point 0, and is continuous at 0 (and only at 0).

VI.2.10.10. THEOREM. [EXTENDED BANACH FIXED-POINT THEOREM] Let (X, ρ) be a complete metric space, and $T : X \to X$ a strict power contraction. Then T has a unique fixed point p. If x is any point of X, then the sequence $(T^n(x))$ converges to p.

PROOF: Suppose T^n is a strict contraction. Then T^n has a unique fixed point p by Theorem VI.2.10.3.. We have

$$T^n(T(p)) = T^{n+1}(p) = T(T^n(p)) = T(p)$$

so $T(p)$ is also a fixed point of T^n; thus $T(p) = p$ by uniqueness, and p is a fixed point of T. If q is any fixed point of T, then q is also a fixed point of T^n, hence $q = p$ by uniqueness, so T has a unique fixed point.
Fix $x \in X$. Then $\lim_{k\to\infty} T^{kn}(x) = p$. Also, for each j, $T^{kn+j}(x) = T^{kn}(T^j(x)) \to p$ as $k \to \infty$. It follows that $\lim_{m\to\infty} T^m(x) = p$. 👍

VI.2.10.11. If (X, ρ) is not complete, a strict power contraction on X need not have a fixed point (VI.2.11.1.). However, *if* it does, the fixed point is unique (VI.2.10.5.), and the proof shows that $T^n(x) \to p$ for any $x \in X$ (this part of the proof does not require completeness). The alternate argument of VI.2.10.5. does not work, however, since in general there is no reasonable way to extend a power contraction to the completion.

VI.2.11. Exercises

VI.2.11.1. Let $X = \mathbb{Q} \cap [1, 2]$ with the usual metric, and define $T : X \to X$ by

$$T(x) = \frac{1}{2}\left(x + \frac{2}{x}\right) .$$

(a) Show that T is a strict contraction on X.

(b) Show that T has no fixed point in X.

(c) Identify $\bar{X}$ with $[1, 2]$ with the usual metric, and $\bar{T}$ with the function from $\bar{X}$ to $\bar{X}$ defined by the same formula as T. Find the fixed point of $\bar{T}$.

VI.2.11.2. A strict power contraction does not have to be continuous (VI.2.10.9.). If it has a fixed point (e.g. if the metric space is complete), does it have to be continuous at the fixed point? Give a proof or counterexample.

VI.2.11.3. Discuss the difficulties in trying to extend VI.2.10.7. to power contractions. Is there any reasonable result?

VI.3. Euclidean Space

VI.3.1. Metric Axiomatization of Euclidean Space

There are several approaches one can take to defining n-dimensional Euclidean space. We can just define it to *be* $\mathbb{R}^n$ (with the standard metric ρ_2). We will eventually take this approach, but it is worth noting that Euclidean space can be defined in a more abstract manner and shown to be effectively the same as $\mathbb{R}^n$.

There are two such approaches possible, characterizing $\mathbb{R}^n$ directly as a metric space, and giving a set of purely geometric axioms. Since this is a book about analysis, we will take the metric space approach, which may be rather unsatisfying to geometers. The metric approach is, however, quite elegant (although some of the proofs are rather involved), and not as well known as it should be. Our exposition is adapted from [Blu70] (cf. also [Blu75]), where the history of various metric approaches is discussed. For the purely geometric approach, see [Gre93], [Har00], and/or [Moi90].

VI.3.1.1. One easy metric approach is to define n-dimensional Euclidean space to be a metric space isometric to $\mathbb{R}^n$. Since the isometries of $\mathbb{R}^n$ can be nicely described (XX.2.2.4.), and preserve the affine structure, an isometry from an abstract metric space E to $\mathbb{R}^n$ is almost unique (up to a choice of origin and an ordered orthonormal basis) and can be used to transfer the structure of $\mathbb{R}^n$ to E. We will regard a choice of an isometry of E with $\mathbb{R}^n$ to be a *choice of a coordinate system* in E. By choosing a coordinate system, we can identify E with $\mathbb{R}^n$ and all analysis we do on E can be described using the chosen coordinates; it will all turn out to be independent of the coordinate system chosen in an appropriate sense ().

VI.3.1.2. We can, however, characterize n-dimensional Euclidean space abstractly as a metric space by three axioms (VI.3.1.43.), or four simpler axioms plus completeness (VI.3.1.44.). These axioms will presuppose having $\mathbb{R}$ available by some previous construction; this is undesirable from a geometric point of view, but necessary from the metric standpoint (indeed, the definition of a metric space itself depends on already having $\mathbb{R}$, or at least $\mathbb{R}_+$). $\mathbb{R}$ is always assumed to have its usual metric.

We will give the three axioms for a metric space (E, ρ), and discuss some variations and equivalent versions of them.

The Line Axiom

VI.3.1.3. Definition. A subset L of a metric space (E, ρ) is a *line* if L is isometric to $\mathbb{R}$. A subset of E isometric to a closed bounded interval $[\alpha, \beta]$ in $\mathbb{R}$ is called a *line segment* in E, with the points corresponding to α and β the *endpoints*. A subset of E isometric to $[0, \infty)$ is called a *ray* in E, beginning at (or with endpoint) the point corresponding to 0.

VI.3.1.4. **Line Axiom:** For any two points $a, b \in E$ there is a line in E containing a and b.

Such a subset is called a *line through a and b*. Note that no uniqueness is required as part of the axiom (uniqueness will follow from the second axiom; cf. VI.3.1.27.). A metric space in which each pair of distinct points lies on a unique line will be said to satisfy the *Unique Line Axiom*. Similarly:

VI.3.1.5. **Line Segment Axiom:** For any two (distinct) points $a, b \in E$ there is a line segment in E with endpoints a and b.

A metric on a set which satisfies the Line Segment Axiom is called a *strictly intrinsic metric*. See () for an explanation of this terminology from metric geometry. The Line Axiom is more important than the Line Segment Axiom in this section, but the Line Segment Axiom is central in metric geometry ().

VI.3.1.6. EXAMPLES. (i) $(\mathbb{R}^2, \rho_\infty)$ () satisfies the Line Axiom (as does any normed real vector space viewed as a metric space). But lines through two given points are not unique in general: for example, the graph of any Lipschitz function from $\mathbb{R}$ to $\mathbb{R}$ () with Lipschitz constant ≤ 1 (contractive function), regarded as a subset of $\mathbb{R}^2$ in the usual way, is a line in this metric space.

(ii) Any convex set in $\mathbb{R}^n$ (or any normed vector space) satisfies the Line Segment Axiom. Again the line segment need not be unique.

(iii) A circle with the arc length metric satisfies the Line Segment Axiom, as does a 2-sphere with the great circle arc length metric. The line segment is not unique for diametrically opposed points. They do not satisfy the Line Axiom.

If the metric space (E, ρ) is assumed to be complete, an apparently much weaker assumption implies the Line Axiom. We first make a definition:

VI.3.1.7. DEFINITION. Let $a, c \in E$. A point $b \in E$ is *between* a and c if $b \neq a$, $b \neq c$, and $\rho(a, c) = \rho(a, b) + \rho(b, c)$.

In $\mathbb{R}$, b is between a and c in this sense if and only if b is between a and c with respect to the usual ordering on $\mathbb{R}$. More generally, if L is a line in a metric space, then the various isometries of L with $\mathbb{R}$ give two orderings on L, and if $a, b, c \in L$, then b is between a and c in the sense of VI.3.1.7. if and only if b is between a and c with respect to either (both) of these orderings. In particular, if a, b, and c are distinct points on a line in a metric space, exactly one of the points is between the other two.

Note, however, that the set of points between two given points in a metric space need not look anything like a line segment (cf. Exercise VI.3.3.1.); indeed, this set can easily be empty or finite.

There is a limited sense in which betweenness is transferable:

VI.3.1.8. PROPOSITION. Let a, b, c, d be points in a metric space (E, ρ). If b is between a and d and c is between b and d, then c is between a and d and b is between a and c.

PROOF: We have

$$\rho(a, d) = \rho(a, b) + \rho(b, d) = \rho(a, b) + \rho(b, c) + \rho(c, d)$$

and since $\rho(a, d) \leq \rho(a, c) + \rho(c, d)$ by the triangle inequality, we have

$$\rho(a, c) \geq \rho(a, b) + \rho(b, c)$$

by subtraction; the opposite inequality holds by the triangle inequality, so

$$\rho(a, c) = \rho(a, b) + \rho(b, c)$$

and c is between a and b. Then we have

$$\rho(a, d) = \rho(a, b) + \rho(b, c) + \rho(c, d) = \rho(a, c) + \rho(c, d)$$

so c is between a and d. 👍

VI.3.1.9. Care must be exercised in combining betweenness relations, however; overlapping relations cannot be combined in general (betweenness is not "transitive"). For example, if $(E,\rho) = (\mathbb{R}^2, \rho_\infty)$ and $a = (0,0)$, $b = (1,1)$, $c = (2,0)$, and $d = (1,-1)$, then b is between a and c and c is between b and d, but neither b nor c is between a and d.

There is a useful consequence of VI.3.1.8.:

VI.3.1.10. Corollary. Let (E,ρ) be a metric space, $a,b,c \in E$, S_1 a line segment with endpoints a and b, and S_2 a line segment with endpoints b and c. Then $S = S_1 \cup S_2$ is a line segment with endpoints a and c if and only if b is between a and c.

Proof: If S is a line segment, then b is clearly between a and c. Conversely, if b is between a and c, to show S is a line segment it suffices to show that if $x \in S_1$, $y \in S_2$, $x \neq b$, $y \neq b$, then $\rho(x,y) = \rho(x,b) + \rho(b,y)$, i.e. that b is between x and y. Apply VI.3.1.8. to a,b,y,c to conclude that b is between a and y, and then apply VI.3.1.8. again to y,b,x,a to conclude that b is between y and x. 👍

VI.3.1.11. **Between Axiom:** If $a,c \in E$, $a \neq c$, then there is a $b \in E$ which is between a and c.

VI.3.1.12. **Extension Axiom:** If $a,b \in E$, $a \neq b$, then there is a $c \in E$ such that b is between a and c.

In metric geometry (e.g. in [Blu70]), a metric space satisfying the Between Axiom is often called *metrically convex*, and a metric space satisfying the Extension Axiom is called *externally convex*. We have not used these terms since they do not seem appropriate in general: for example, the subset $[0,1) \cup \{2\}$ of $\mathbb{R}$ satisfies the Between Axiom, and $\mathbb{Z}$ satisfies the Extension Axiom, but neither resembles a convex set in the usual sense (). However, the two axioms together, along with completeness, are quite powerful:

VI.3.1.13. Theorem. Let (E,ρ) be a complete metric space satisfying the Between Axiom and the Extension Axiom. Then (E,ρ) satisfies the Line Axiom.

Completeness is essential for this result: the subsets $\mathbb{Q}$ and $(0,1)$ of $\mathbb{R}$ each satisfy both the Between Axiom and the Extension Axiom, but not the Line Axiom. Note also that even for complete metric spaces, neither axiom implies the other: consider $[0,1]$ and $\mathbb{Z}$.

The proof of VI.3.1.13. consists of three lemmas.

VI.3.1.14. Lemma. Let (E,ρ) be a complete metric space satisfying the Between Axiom. Then (E,ρ) satisfies the Line Segment Axiom.

Proof: We give a simple proof using Zorn's Lemma. A somewhat more complicated proof can be given which requires only the Axiom of Dependent Choice; cf. [Blu70, Theorem 14.1] and Exercise VI.3.3.8..

Let $\mathcal{X}$ be the set of all subsets of E containing a and b and isometric to a subset of $[0,\lambda]$, where $\lambda = \rho(a,b)$, ordered by inclusion. (Note that $\{a,b\} \in \mathcal{X}$, so $\mathcal{X}$ is nonempty.) For every set $X \in \mathcal{X}$, there is a unique

isometry ϕ_X from X to $[0,\lambda]$ with $\phi_X(a)=0$ and $\phi_X(b)=\lambda$. If $\{X_i\}$ is a chain in $\mathcal{X}$, and $X=\cup_i X_i$, define $\phi_X : X \to [0,1]$ by $\phi_X(x)=\phi_{X_i}(x)$ for $x\in X_i$; this is well defined by uniqueness of the ϕ_{X_i}: if $X_i\subseteq X_j$, then $\phi_{X_j}|_{X_i}=\phi_{X_i}$. Then ϕ_X is an isometry since if $x,y\in X$, then $x,y\in X_i$ for some i.

If $X\in\mathcal{X}$, then since ϕ_X is an isometry it extends to an isometry from the closure $\bar{X}$ in E to $[0,\lambda]$, and hence $\bar{X}\in\mathcal{X}$.

Let X be a maximal element of $\mathcal{X}$. Then X is closed in E, and hence complete; thus $K=\phi_X(X)$ is complete, hence closed in $[0,\lambda]$. If $K\neq[0,\lambda]$, there is an interval $[\alpha,\beta]\subseteq[0,\lambda]$ with $\alpha,\beta\in K$ and $(\alpha,\beta)\cap K=\emptyset$. Let $c,d\in X$ with $\phi_X(c)=\alpha$, $\phi_X(d)=\beta$. There is a $p\in E$ between c and d. Since c and d are between a and b, by VI.3.1.8. we have p between a and b. Then $p\notin X$ since $\rho(a,p)\in(\alpha,\beta)$. If x is any element of X, then x is either between a and c or between d and b; it follows from VI.3.1.8. that if $x,y\in X$, then one of x,y,p is between the other two. Thus ϕ_X extends to an isometry of $X\cup\{p\}$ to $[0,\lambda]$, contradicting maximality of X. Thus $K=[0,\lambda]$ and X is a line segment with endpoints a and b. 👍

Now we turn to the procedure of extending line segments.

VI.3.1.15. Lemma. Let (E,ρ) be a complete metric space satisfying the Extension Axiom, and let $a,b\in E$. Then

$$\sup\{\rho(a,c) : c\in E, b \text{ is between } a \text{ and } c\} = +\infty\ .$$

Proof: Suppose $\lambda_1=\sup\{\rho(a,c) : c\in E, b \text{ is between } a \text{ and } c\}$ is finite. Let $c_1\in E$ with b between a and c_1 and $\rho(a,c_1)>\lambda_1-1$. If $c\in E$ and c_1 is between a and c, then b is between a and c by VI.3.1.8.. Thus

$$\lambda_2=\sup\{\rho(a,c) : c\in E, c_1 \text{ is between } a \text{ and } c\}$$

is finite and $\lambda_2\leq\lambda_1$. Inductively let

$$\lambda_{n+1}=\sup\{\rho(a,c) : c\in E, c_n \text{ is between } a \text{ and } c\}\leq\lambda_n$$

and let c_{n+1} be a point of E with c_n between a and c_{n+1} (hence by VI.3.1.8. b is between a and c_{n+1}) and $\rho(a,c_{n+1})>\lambda_{n+1}-\frac{1}{n+1}$.

Set $\lambda=\inf\lambda_n=\lim\lambda_n$. If $n<m$, we have c_n between a and c_m, hence

$$\lambda_n-\frac{1}{n}<\rho(a,c_n)<\rho(a,c_m)\leq\lambda_n$$

so $\rho(c_n,c_m)<\frac{1}{n}$ for $n<m$ and thus (c_n) is a Cauchy sequence in E. By completeness, (c_n) converges to some $c\in E$, and we have

$$\lambda_n-\frac{1}{n}<\rho(a,c)\leq\lambda_n$$

for all n; thus $\rho(a,c)=\lambda$ by the Squeeze Theorem (). By continuity, b and c_n are between a and c for all n.

Choose $d\in E$ with c between a and d. Then $\rho(a,d)>\lambda$, so $\rho(a,d)>\lambda_n$ for some n. But c_n is between a and d, so this contradicts the definition of λ_n. Thus the assumption that $\lambda_1<+\infty$ is false. 👍

VI.3.1.16. LEMMA. Let (E, ρ) be a complete metric space satisfying the Between and Extension Axioms. Then

(i) (E, ρ) satisfies the Line Axiom.

(ii) Every line segment in E can be extended to a line in E.

(iii) If $a, b \in E$, and $x_1, \ldots, x_n$ are points of E between a and b, with x_k between a and x_{k+1} for $1 \leq k \leq n-1$, then there is a line in E containing a, b, and all the x_k.

PROOF: For (i), let $a, b \in E$. By Lemma VI.3.1.14., there is a line segment S_0 with endpoints a and b. For (ii), let S_0 be a line segment in E, and let a and b be its endpoints. For (iii), there is a line segment with endpoints a and x_1, for each $k < n$ a line segment with endpoints x_k and x_{k+1}, and a line segment with endpoints x_n and b. By repeated use of VI.3.1.10., the union of these segments is a line segment S_0 with endpoints a and b containing all the x_k.

In each case, begin with the segment S_0. By Lemma VI.3.1.15. there is a $c_1 \in E$ with b between a and c_1 and $\rho(a, c_1) \geq 2\rho(a, b)$. By Lemma VI.3.1.14. there is a line segment S_1 in E with endpoints b and c_1. Since b is between a and c_1, $S_0 \cup S_1$ is a line segment with endpoints a and c_1 by VI.3.1.10..

Now let $c_2 \in E$ with a between c_1 and c_2 and $\rho(c_1, c_2) \geq 2\rho(a, c_1)$, and let S_2 be a line segment with endpoints a and c_2. Then $S_0 \cup S_1 \cup S_2$ is a line segment in E with endpoints c_1 and c_2. Now let $c_3 \in E$ with c_1 between c_2 and c_3 and $\rho(c_2, c_3) \geq 2\rho(c_1, c_2)$, and let S_3 be a line segment between c_1 and c_3. Continue inductively placing c_{n+1} at the c_{n-1} end of the already constructed segment with endpoints c_{n-1} and c_n. Then $\cup_{k=0}^{n+1} S_k$ is a line segment between c_n and c_{n+1} extending the previous segment, and the length of this segment is at least $2^{n+1}\rho(a, b)$. The union of the S_n is a line in E through a and b. 👍

This completes the proof of Theorem VI.3.1.13..

Here are variations of the Between and Extension Axioms which are sometimes useful:

VI.3.1.17. DEFINITION. If $a, b \in E$, then (a, b) has the *Interpolation Property* if for each $t \in [0, 1]$ there is a point $c \in E$ such that

$$\rho(a, c) = t\rho(a, b) \quad \text{and} \quad \rho(b, c) = (1 - t)\rho(a, b) \ .$$

If $a, b \in E$, and $t \in \mathbb{R}$, then (a, b) has the *Extrapolation Property* if for each $t \geq 1$ there is a point $c \in E$ such that

$$\rho(a, c) = t\rho(a, b) \quad \text{and} \quad \rho(b, c) = (t - 1)\rho(a, b) \ .$$

The pair (a, b) has the *Unique Interpolation Property* (resp. the *Unique Extrapolation Property*) if the c is unique.

The Interpolation Property is symmetric in a and b, but the Extrapolation Property is not.

VI.3.1.18. If there is a line segment in E with endpoints a and b, then (a, b) has the Interpolation Property. But even the Unique Interpolation Property for (a, b) does not imply that there is a line segment in E with endpoints a and b (Exercise VI.3.3.3.). Similarly, if there is a ray from a through b, then (a, b) has the Extrapolation Property; but even the Unique Extrapolation Property does not imply existence of a ray.

A metric space with the Interpolation Property satisfies the Between Axiom; the converse is false in general if the space is incomplete (e.g. $\mathbb{Q}$). Similarly, the Extrapolation Property implies the Extension Axiom, but not conversely (e.g. $\mathbb{Z}$).

Here is a result which, among other things, can substitute for completeness in yielding the Line Axiom:

VI.3.1.19. THEOREM. (i) If every pair (a, b) of points in E has the Unique Interpolation Property, then for every $a, b \in E$, $a \neq b$, there is a unique line segment in E with endpoints a and b.

(ii) If every pair (a, b) of points in E has the Unique Interpolation Property and the Unique Extrapolation Property, then through any two distinct points of E there is a unique line.

See Exercise VI.3.3.7. for the proof.

The Pythagorean Axiom

The second axiom for Euclidean space has several equivalent forms, at least in the presence of the Line Axiom.

VI.3.1.20. **Weak Euclidean Four-Point Axiom:** If $a, b, c, d \in E$, and b, c, d lie on a line in E, then $\{a, b, c, d\}$ is isometric to a set of four points in $\mathbb{R}^2$ (with the usual Euclidean metric).

The stronger *Euclidean Four-Point Axiom* is the same without the requirement that three of the points lie on a line, and $\mathbb{R}^2$ replaced by $\mathbb{R}^3$. These axioms trivially hold in $\mathbb{R}^n$ (four points in $\mathbb{R}^n$ generate an affine subspace of dimension ≤ 3).

One can similarly define the *Euclidean n-Point Axiom* for any n. Note that every metric space satisfies the Euclidean Three-Point Axiom, since if α, β, γ are any positive real numbers with each less than or equal to the sum of the other two, there is a (possibly degenerate) triangle in $\mathbb{R}^2$ with sides α, β, γ. It turns out that the Weak Euclidean Four-Point Axiom along with the Line Axiom implies the Euclidean n-Point Axiom for all n (VI.3.1.27., VI.3.1.35.).

VI.3.1.21. DEFINITION. Let (E, ρ) be a metric space and a, b distinct points of E. A point m in E is a *midpoint* for a and b if $\rho(a, m) = \rho(m, b) = \frac{1}{2}\rho(a, b)$. A midpoint m is a *linear midpoint* if there is a line in E containing a, b, and m.

A midpoint for a and b is obviously between a and b. Each line through a and b contains a unique midpoint of a and b; however, if lines are not unique, (linear) midpoints may not be unique either (Exercise VI.3.3.2.).

If (E, ρ) is complete and satisfies the Between and Extension Axioms, then any midpoint m for a and b automatically lies on a line through a and b (VI.3.1.16.(iii)) and is thus a linear midpoint. The Line Axiom is not enough to guarantee this, however (?)

VI.3.1.22. **Parallelogram Axiom:** If $a, b, c \in E$, and m is a linear midpoint of b and c, then

$$2\left[\rho(a,b)^2 + \rho(a,c)^2\right] = \rho(b,c)^2 + 4\rho(a,m)^2 \ .$$

This law expresses the familiar geometric fact that the sum of the squares of the lengths of the diagonals of a parallelogram equals the sum of the squares of the lengths of the sides. (The fourth vertex of the "parallelogram" would be a point d for which m is also a midpoint for a and d. In the presence of the Line Axiom, it will follow that there is a unique such d, and opposite sides of the resulting figure are automatically equal length.) This axiom holds in $\mathbb{R}^n$ (XX.11.2.6.).

VI.3.1.23. Definition. Let S be a subset of a metric space (E, ρ), and a a point of E. A *foot* of a on S is a point f of S such that $\rho(a,f) = \rho(a,S) = \min\{\rho(a,c) : c \in S\}$.

If $a \in S$, then a itself is a foot of a on S (and is the only foot of a on S). The concept of foot is only interesting if $\rho(a,S) > 0$.

VI.3.1.24. Proposition. Let (E, ρ) be a metric space, S a subset of E with the Bolzano-Weierstrass property () (e.g. a subset of E isometric to $\mathbb{R}^n$ for some n), and a a point of E. Then there exists a foot of a on S.

Proof: Let c be an element of S. Let

$$K = \{x \in S : \rho(x,c) \leq 2\rho(a,c)\} \ .$$

Then the triangle inequality implies that if e is a point of L which is not in K, then $\rho(a,e) > \rho(a,c)$. So we have $\rho(a,S) = \rho(a,K)$, and K is closed and bounded in E, hence compact, so the distance is minimized at some point of K (). 👍

VI.3.1.25. In particular, if L is a line in E, and a a point of E, then a has a foot on L. A foot of a point a on a line L need not be unique in general. For example, in $(\mathbb{R}^2, \rho_\infty)$, if L is the x-axis and $a = (0,1)$, then any point $(t,0)$ with $|t| \leq 1$ is a foot of a on L.

VI.3.1.26. **Pythagorean Axiom:** If L is a line in E, and a a point of E, and f is a foot of a on L, then for any $c \in L$ we have

$$\rho(a,c)^2 = \rho(a,f)^2 + \rho(f,c)^2 \ .$$

The Pythagorean Axiom holds in $\mathbb{R}^n$ (). It obviously implies that the foot of a point on a line is unique.

VI.3.1.27. Theorem. Let (E, ρ) be a metric space satisfying the Line Axiom. The following are equivalent:

(i) (E, ρ) satisfies the Weak Euclidean Four Point Axiom.

(ii) (E, ρ) satisfies the Parallelogram Axiom.

(iii) (E, ρ) satisfies the Pythagorean Axiom.

(iv) If L is a line in E and a a point of E not on L, then $L \cup \{a\}$ can be isometrically embedded in $\mathbb{R}^2$ (with the usual metric).

These conditions imply that there is a unique line through any two distinct points of E.

PROOF: (iv) $\Rightarrow$ (i) $\Rightarrow$ (ii) and (iv) $\Rightarrow$ (iii) are trivial. We will prove (iii) $\Rightarrow$ (ii) $\Rightarrow$ (iv). (In fact, the Line Axiom is only needed for (iv) $\Rightarrow$ (i) and for the last assertion.) 👍

VI.3.1.28. LEMMA. Let (E, ρ) be a metric space satisfying the Pythagorean Axiom. Then E satisfies the Parallelogram Axiom.

PROOF: Let $a, b, c \in E$ and L a line through b and c. Let m be the midpoint of b and c on L, and let f be the foot of a on L. Set $\alpha = \rho(a, c)$, $\beta = \rho(a, b)$, $\gamma = \rho(b, c)$. There are two cases to consider: (1) $f = b$, $f = c$, or f is between b and c, and (2) c is between b and f (the case where b is between f and c is symmetric and handled by interchanging b and c). See Figure (), although the reader is cautioned that since the points are in E and not in $\mathbb{R}^2$, the accuracy of the picture cannot (yet) be justified and it should be used only as a visual aid in following the proof (this is usually a good attitude to take with any picture or diagram).

Consider case (1). By interchanging b and c if necessary, we may assume that $f = c$, $f = m$, or f is between m and c. Then there is a t, $0 \leq t \leq \frac{1}{2}$, such that $\rho(f, c) = t\gamma$, $\rho(m, f) = \left(\frac{1}{2} - t\right)\gamma$, $\rho(b, f) = (1-t)\gamma$. Set $\eta = \rho(a, f)$ and $\zeta = \rho(a, m)$. By the Pythagorean Axiom we have

$$\alpha^2 = \eta^2 + t^2\gamma^2$$
$$\beta^2 = \eta^2 + (1-t)^2\gamma^2 = \eta^2 + \gamma^2 - 2t\gamma^2 + t^2\gamma^2$$
$$\zeta^2 = \eta^2 + \left(\frac{1}{2} - t\right)^2\gamma^2 = \eta^2 + \frac{1}{4}\gamma^2 - t\gamma^2 + t^2\gamma^2$$

so we have

$$2\alpha^2 + 2\beta^2 = 2\eta^2 + 2t^2\gamma^2 + 2\eta^2 + 2\gamma^2 - 4t\gamma^2 + 2t^2\gamma^2 = 4\eta^2 + 2\gamma^2 - 4t\gamma^2 + 4t^2\gamma^2$$
$$4\zeta^2 + \gamma^2 = 4\eta^2 + \gamma^2 - 4t\gamma^2 + 4t^2\gamma^2 + \gamma^2 = 2\alpha^2 + 2\beta^2$$

which is the Parallelogram formula.

Now consider case (2). There is a $t \geq 0$ such that $\rho(f, c) = t\gamma$, $\rho(f, m) = \left(\frac{1}{2} + t\right)\gamma$, $\rho(f, b) = (1+t)\gamma$. Again by the Pythagorean Axiom

$$\alpha^2 = \eta^2 + t^2\gamma^2$$
$$\beta^2 = \eta^2 + (1+t)^2\gamma^2 = \eta^2 + \gamma^2 + 2t\gamma^2 + t^2\gamma^2$$
$$\zeta^2 = \eta^2 + \left(\frac{1}{2} + t\right)^2\gamma^2 = \eta^2 + \frac{1}{4}\gamma^2 + t\gamma^2 + t^2\gamma^2$$

so we have

$$2\alpha^2 + 2\beta^2 = 2\eta^2 + 2t^2\gamma^2 + 2\eta^2 + 2\gamma^2 + 4t\gamma^2 + 2t^2\gamma^2 = 4\eta^2 + 2\gamma^2 + 4t\gamma^2 + 4t^2\gamma^2$$
$$4\zeta^2 + \gamma^2 = 4\eta^2 + \gamma^2 + 4t\gamma^2 + 4t^2\gamma^2 + \gamma^2 = 2\alpha^2 + 2\beta^2$$

which is the Parallelogram formula. 👍

VI.3.1.29. LEMMA. Let (E,ρ) be a metric space satisfying the Parallelogram Axiom. Let L be a line in E, b, c points on L, and a a point not on L. Set $\alpha = \rho(a,c)$, $\beta = \rho(a,b)$, $\gamma = \rho(b,c)$. Then for any $t \in \mathbb{R}$, if d_t is the unique point on L with $\rho(b,d_t) = |t|\gamma$ and $\rho(d,c) = |1-t|\gamma$, we have

$$\rho(a,d_t)^2 = t\alpha^2 + (1-t)\beta^2 - t(1-t)\gamma^2 \ .$$

PROOF: By continuity, it suffices to prove the formula if t is a dyadic rational number. The formula is trivially true if $t = 0$ or $t = 1$ (and the case $t = \frac{1}{2}$ is the Pythagorean Axiom, although we do not explicitly need this case).

We first show that if the formula holds for t, it also holds for $-t$. Set $\rho(a,d_t) = \eta$. By assumption,

$$\eta^2 = t\alpha^2 + (1-t)\beta^2 - t(1-t)\gamma^2 \ .$$

Since $\rho(d_t, d_{-t}) = 2|t|\gamma$ and b is the midpoint of d_t and d_{-t} on L, by the Parallelogram Axiom we have

$$2\rho(a,d_{-t})^2 + 2\eta^2 = 4\beta^2 + 4t^2\gamma^2$$

$$\rho(a,d_{-t})^2 = 2\beta^2 - \eta^2 + 2t^2\gamma^2 = [2-(1-t)]\beta^2 - t\alpha^2 + [2t^2 + t(1-t)]\gamma^2$$

$$= -t\alpha^2 + (1+t)\beta^2 + t(1+t)\gamma^2 = (-t)\alpha^2 + (1-(-t))\beta^2 - (-t)(1-(-t))\gamma^2$$

so the formula holds for $-t$.

Similarly, we show that if the formula holds for t, it holds for $2t$. Again let $\eta = \rho(a,d_t)$. Since $\rho(b,d_{2t}) = 2|t|\gamma$ and d_t is the midpoint of b and d_{2t} on L, by the Parallelogram Axiom we have

$$2\rho(a,d_{2t})^2 + 2\beta^2 = 4\eta^2 + 4t^2\gamma^2$$

$$\rho(a,d_{2t})^2 = 2\eta^2 - \beta^2 + 2t^2\gamma^2 = 2[t\alpha^2 + (1-t)\beta^2 - t(1-t)\gamma^2] - \beta^2 + 2t^2\gamma^2$$

$$= 2t\alpha^2 + [2(1-t)-1]\beta^2 + [2t^2 - 2t(1-t)]\gamma^2 = 2t\alpha^2 + (1-2t)\beta^2 - 2t(1-2t)\gamma^2$$

so the formula holds for $2t$.

Finally, suppose $s, t \in \mathbb{R}$ and the statement is true for s and t. We will show that the statement is also true for $r = \frac{s+t}{2}$. Let d_s, d_t, d_r be the points of L corresponding to s, t, r respectively, and set $\eta = \rho(a,d_t)$, $\zeta = \rho(a,d_s)$. Then by assumption

$$\eta^2 = t\alpha^2 + (1-s)\beta^2 - t(1-t)\gamma^2$$

$$\zeta^2 = s\alpha^2 + (1-s)\beta^2 - s(1-s)\gamma^2$$

and, since $\rho(d_t,d_s) = |s-t|\gamma$ and d_r is the midpoint of d_t and d_s on L, by the Parallelogram Axiom we have

$$\rho(a,d_r)^2 = \frac{1}{2}[\zeta^2 + \eta^2] - \frac{1}{4}(s-t)^2\gamma^2$$

$$= \frac{1}{2}[s\alpha^2 + (1-s)\beta^2 - s(1-s)\gamma^2 + t\alpha^2 + (1-t)\beta^2 - t(1-t)\gamma^2] - \frac{1}{4}(s^2 - 2st + t^2)\gamma^2$$

$$= \frac{s+t}{2}\alpha^2 + \frac{(1-s)+(1-t)}{2}\beta^2 - \left[\frac{1}{2}s - \frac{1}{2}s^2 + \frac{1}{2}t - \frac{1}{2}t^2 + \frac{1}{4}s^2 + \frac{1}{4}t^2 - \frac{1}{2}st\right]\gamma^2$$

$$= \frac{s+t}{2}\alpha^2 + \left[1 - \frac{s+t}{2}\right]\beta^2 - \left[\frac{1}{2}(s+t) - \frac{1}{4}s^2 - \frac{1}{4}t^2 - \frac{1}{2}st\right]\gamma^2$$

$$= r\alpha^2 + (1-r)\beta^2 - (r-r^2)\gamma^2$$

so the formula holds for r too. Therefore the formula holds for all dyadic rational t. 👍

VI.3.1.30. Corollary. Let (E, ρ) be a metric space satisfying the Parallelogram Axiom. Let L be a line in E, and a a point of E. Then there is an isometry from $L \cup \{a\}$ to $\mathbb{R}^2$. If $b, c \in L$ and $\{a', b', c'\}$ is a subset of $\mathbb{R}^2$ isometric to $\{a, b, c\}$, then there is an isometry ϕ from $L \cup \{a\}$ to $\mathbb{R}^2$ with $\phi(a) = a'$, $\phi(b) = b'$, $\phi(c) = c'$.

Proof: Let b and c be two distinct points on L. Let $\{a', b', c'\}$ be three points in $\mathbb{R}^2$ isometric to $\{a, b, c\}$, and L' the line in $\mathbb{R}^2$ through b' and c'. If d is any point on L and d' the point on L' with $\rho_2(b', d') = \rho(b, d)$ and $\rho_2(c', d') = \rho(c, d)$, then applying Lemma VI.3.1.29. in both E and $\mathbb{R}^2$ we obtain $\rho_2(a', d') = \rho(a, d)$. Thus $L \cup \{a\}$ is isometric to $L' \cup \{a'\}$ under an isometry ϕ with $\phi(a) = a'$, $\phi(b) = b'$, $\phi(c) = c'$. 👍

VI.3.1.31. Lemma. Let (E, ρ) be a metric space satisfying the Parallelogram Axiom. Let a, b, c be points of E with $a \neq b$ and $a \neq c$, and let L_1 be a line in E through a and b, and L_2 a line in E through a and c. If ϕ is an isometry of $\{a, b, c\}$ with $\{a', b', c'\} \subseteq \mathbb{R}^2$, then ϕ extends uniquely to an isometry from $L_1 \cup L_2$ to $\mathbb{R}^2$.

Proof: Let L_1' be the (unique) line in $\mathbb{R}^2$ through a' and b', and L_2' the line in $\mathbb{R}^2$ through a' and c'. There is a unique isometry ϕ_1 from L_1 to L_1' with $\phi_1(a) = a'$ and $\phi_1(b) = b'$, since if $x \in L_1$, x is uniquely determined by $\rho(x, a)$ and $\rho(x, b)$, and similarly for L_1'. In the same way, there is a unique isometry ϕ_2 from L_2 to L_2' with $\phi_2(a) = a'$, $\phi_2(c) = c'$. It suffices to show that if $x \in L_1$ and $y \in L_2$, and $x' = \phi_1(x)$, $y' = \phi_2(y)$, then $\rho_2(x', y') = \rho(x, y)$. We may assume $x \neq a$ and $y \neq a$.

By VI.3.1.30., there is an isometry ψ from $L_1 \cup \{c\}$ to $\mathbb{R}^2$ with $\psi(a) = a'$, $\psi(b) = b'$, $\psi(c) = c'$, and we must have $\psi|_{L_1} = \phi_1$, so $\psi(x) = x'$. Thus $\rho_2(x', c') = \rho(x, c)$, so by VI.3.1.30. there is an isometry θ from $L_2 \cup \{x\}$ to $\mathbb{R}^2$ with $\theta(a) = a'$, $\theta(c) = c'$, $\theta(x) = x'$. We must have $\theta|_{L_2} = \phi_2$, so $\theta(y) = y'$ and we have $\rho_2(x', y') = \rho(x, y)$. 👍

VI.3.1.32. Corollary. Let (E, ρ) be a metric space satisfying the Parallelogram Axiom, and let a and b be distinct points of E. Then there is at most one line in E through a and b.

Proof: If L_1 and L_2 are lines in E through a and b, set $c = b$ and apply VI.3.1.31.. 👍

This completes the proof of Theorem VI.3.1.27..

Affine Subspaces

It would seem to make sense to define an n-dimensional affine subspace of a metric space to be a subset isometric to $\mathbb{R}^n$. But it will turn out to be more convenient to make the following unorthodox inductive definition, which is equivalent (VI.3.1.35.) in the presence of the Line and Parallelogram Axioms:

VI.3.1.33. Definition. Let (E, ρ) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom.

(i) A *zero-dimensional affine subspace* of E is a single point.

(ii) If $n \geq 1$, an *n-dimensional affine generator* in E is a subset $G(L, a)$ consisting of all points x lying on a line through a and a point of L, where L is an $(n-1)$-dimensional affine subspace of E and a is a point of E not in L.

(iii) If $n \geq 1$, an *n-dimensional affine subspace* of E is the closure of an n-dimensional affine generator in E.

(iv) A set $\{x_0, \ldots, x_n\}$ of $n+1$ points of E is *independent* if it is not contained in any $(n-1)$-dimensional affine subspace of E.

It is easy to verify that the one-dimensional affine subspaces of E are precisely the lines in E.

VI.3.1.34. PROPOSITION. Let (E, ρ) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom, and L a subset of E isometric to $\mathbb{R}^n$. Then L is an n-dimensional affine subspace of E.

PROOF: This is easily seen by induction, since if L is an $(n-1)$-dimensional affine subspace of $\mathbb{R}^n$ (in the usual sense, hence by the inductive hypothesis in the sense of VI.3.1.33.), and a is a point of $\mathbb{R}^n$ not on L, then $G(L, a)$ consists of all points of $\mathbb{R}^n$ not in the affine subspace through a parallel to L, along with a, and this set is dense in $\mathbb{R}^n$. 👍

The next theorem, essentially the converse of VI.3.1.34., is perhaps the most important result of this section.

VI.3.1.35. THEOREM. Let (E, ρ) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom, and $n \in \mathbb{N}$. Then

(i) Every n-dimensional affine subspace of E is isometric to $\mathbb{R}^n$.

(ii) If $\{x_0, \ldots, x_n\}$ is an independent set of $n+1$ points of E, then there is a unique n-dimensional subspace of E containing $\{x_0, \ldots, x_n\}$.

PROOF: The proof is by induction on n. The case $n = 1$ is obvious from VI.3.1.27.. For the inductive step, we suppose that the statement is true for n for any metric space satisfying the hypotheses. Then let (E, ρ) be a metric space satisfying the hypotheses. Proofs of (i) and (ii) for $n+1$ will be done in parallel. To set up the proof, we make notational definitions for the two cases.

To prove (i), let L_{n+1} be an $(n+1)$-dimensional affine subspace of E, and let L_n be an n-dimensional affine subspace of E and a a point of E such that L_{n+1} is the closure of $G(L_n, a)$. To prove (ii), let $\{x_0, \ldots, x_{n+1}\}$ be an independent set of $n+2$ points of E. If $\{x_0, \ldots, x_n\}$ is contained in an affine $(n-1)$-dimensional subspace L_{n-1}, then $\{x_0, \ldots, x_{n+1}\}$ is contained in the n-dimensional affine subspace which is the closure of $G(L_{n-1}, x_{n+1})$, a contradiction. Thus $\{x_0, \ldots, x_n\}$ is independent and contained in a unique n-dimensional affine subspace L_n of E. Let $a = x_{n+1}$ and let L_{n+1} be the closure of $G(L_n, a)$.

In both cases we have an $(n+1)$-dimensional affine subspace L_{n+1} which is the closure of an $(n+1)$-dimensional generator $G(L_n, a)$, where L_n is an n-dimensional affine subspace of E and a a point of E not in

L_n. By the inductive hypothesis, L_n is isometric to $\mathbb{R}^n$, hence is complete and has the Bolzano-Weierstrass property, and satisfies the Line Axiom, i.e. if $b, c \in L_n$, then the unique line in E through b and c is contained in L_n.

By VI.3.1.24., there is a foot f of a on L_n. If g is another foot of a on L_n, let L be the line in L_n through f and g; then both f and g are feet of a on L, impossible by the Pythagorean Axiom if $f \neq g$. Thus f is the unique foot of a on L_n, and f is the foot of a on any line in L_n through f. So if b is any point of L_n, we have

$$\rho(a,b)^2 = \rho(a,f)^2 + \rho(f,b)^2 \ .$$

In $\mathbb{R}^{n+1}$, let L'_n be the n-dimensional subspace of points with last coordinate 0, and let $a' = (0, \ldots, 0, \lambda)$, where $\lambda = \rho(a,f)$. Let ϕ be an isometry of L_n onto L'_n sending f to the origin. Then if $b \in L_{n+1}$, we have

$$\rho_2(a', \phi(b))^2 = \lambda^2 + \rho(0, \phi(b))^2$$

so by the previous calculation $\rho_2(a', \phi(b)) = \rho(a,b)$, and ϕ extends to an isometry from $L_n \cup \{a\}$ to $L'_n \cup \{a'\}$, also denoted ϕ.

We next show that ϕ can be extended to an isometry from $G(L_n, a)$ to $G(L'_n, a')$, which consists of all the points of $\mathbb{R}^{n+1}$ whose last coordinate is not λ, along with a' (i.e. $\mathbb{R}^{n+1}$ with the horizontal hyperplane through a', except a' itself, removed).

Let $x \in G(L_n, a)$, $x \neq a$. Then there is a unique $b \in L_n$ such that x is on the line through a and b. Let $\phi(x)$ be the corresponding point on the line through a' and $\phi(b)$ in $\mathbb{R}^{n+1}$. If y is another point in $G(L_n, a)$, $y \neq a$, let y be on the line through a and $c \in L_n$. If the plane through $a', \phi(b), \phi(c)$ in $\mathbb{R}^{n+1}$ is identified with $\mathbb{R}^2$, then by VI.3.1.31. the isometry ϕ from $\{a, b, c\}$ to $\{a', \phi(b), \phi(c)\}$ extends uniquely to the lines through a and b and through a and c; hence in particular $\rho_2(\phi(x), \phi(y)) = \rho(x,y)$ and the ϕ we have defined is an isometry from $G(L_n, a)$ into $G(L'_n, a')$. To see that it is surjective, let $x' \in G(L'_n, a')$ $x' \neq a'$, and let b' be the point of L'_n on the line through a' and x' (a' and x' have different last coordinates, so there is a unique point on this line with last coordinate 0). Then $b' = \phi(b)$ for some $b \in L_n$, and there is a (unique) point x on the line in E through a and b with $x' = \phi(x)$, and ϕ maps $G(L_n, a)$ onto $G(L'_n, a')$.

Thus ϕ extends to an isometry from L_{n+1} into $\mathbb{R}^{n+1}$, also denoted ϕ. We show that this ϕ is surjective. (This would be immediate if (E, ρ) is complete, but (E, ρ) need not be complete.) We must show that if

$$m' = (\alpha_1, \ldots, \alpha_n, \lambda) \in \mathbb{R}^n \setminus G(L'_n, a')$$

then there is an $m \in E$ with $\phi(m) = m'$. Let $b' = (\alpha_1, \ldots, \alpha_n, 0)$ and $c' = (\alpha_1, \ldots, \alpha_n, 2\lambda)$ in $\mathbb{R}^{n+1}$, and let $b = \phi^{-1}(b')$, $c = \phi^{-1}(c')$, and let m be the midpoint of the line segment in E with endpoints a and b. It suffices to show that $m \in L_{n+1}$, for then necessarily $\phi(m) = m'$. But if $t \in (0,1)$, $t \neq \frac{1}{2}$, then the point $d'_t = (\alpha_1, \ldots, \alpha_n, 2t\lambda) \in G(L'_n, a')$ is between b' and c', hence $d_t = \phi^{-1}(d'_t) \in G(L_n, a)$ is between b and c; hence all the points of the line segment between b and c other than m are in $G(L_n, a)$, so m is in the closure of $G(L_n, a)$.

We have now proved (i) for $n+1$, and proved all of (ii) except uniqueness. For uniqueness, let $\tilde{L}_{n+1}$ be any $(n+1)$-dimensional affine subspace of E containing $\{x_0, \ldots, x_{n+1}\}$. Then by (i) $\tilde{L}_{n+1}$ is isometric to $\mathbb{R}^{n+1}$ and hence satisfies the Line Axiom. Any k-dimensional affine subspace of $\tilde{L}_{n+1}$ ($k \leq n+1$) is a k-dimensional affine subspace of E by VI.3.1.34.. Thus $\{x_0, \ldots, x_n\}$ is independent as a subset of $\tilde{L}_{n+1}$, so by the inductive hypothesis applied to $\tilde{L}_{n+1}$, $\{x_0, \ldots, x_n\}$ is contained in a unique n-dimensional affine subspace of $\tilde{L}_{n+1}$. But L_n is the unique n-dimensional affine subspace of E containing $\{x_0, \ldots, x_n\}$, so $L_n \subseteq \tilde{L}_{n+1}$. Then $G(L_n, x_{n+1}) \subseteq \tilde{L}_{n+1}$ since $\tilde{L}_{n+1}$ contains the line in E through any two of its points.

Thus the closure L_{n+1} of $G(L_n, x_{n+1})$ is contained in $\tilde{L}_{n+1}$ since $\tilde{L}_{n+1}$ is closed in E. But this implies that $L_{n+1} = \tilde{L}_{n+1}$ since $\mathbb{R}^{n+1}$ has no proper $(n+1)$-dimensional affine subspaces.

This completes the proof of Theorem VI.3.1.35.. 👍

VI.3.1.36. Corollary. Let (E, ρ) be a metric space satisfying the Line and Parallelogram (or Pythagorean) Axioms, and $S = \{x_0, \ldots, x_m\}$ a subset of E with $n+1$ points. Then, for some $m \leq n$, there is a (unique) m-dimensional affine subspace of E containing S which is the smallest affine subspace of E containing S, called the affine subspace of E *spanned* by S, denoted $span(S)$; $m = n$ if and only if S is independent. Any subset of an independent set is independent.

Any real inner product space with the induced metric satisfies all the axioms considered so far. Conversely:

VI.3.1.37. Theorem. Let (E, ρ) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom, and let $a \in E$. Then

(i) There is a unique real vector space structure on E with a the zero vector, and a unique inner product on this vector space, such that ρ is the metric induced by the inner product.

(ii) Every subset of E satisfying the Line Axiom (in particular, every n-dimensional affine subspace in the sense of VI.3.1.33.) is an affine subspace in the linear algebra sense ().

(iii) If $S = \{x_0, \ldots, x_n\}$ is a finite subset of E, then the span of S in the sense of VI.3.1.36. is the linear span of S in the vector space structure on E with zero vector x_0. The set S is independent in the sense of VI.3.1.33. if and only if $\{x_1, \ldots, x_n\}$ is linearly independent in this vector space structure on E.

Combining this with VI.3.1.13., we obtain:

VI.3.1.38. Corollary. Let (E, ρ) be a complete metric space satisfying the Between, Extension, and Parallelogram (or Pythagorean) Axioms. Then (E, ρ) is isometric to a real Hilbert space ().

Proof: The proof of VI.3.1.37. follows easily from VI.3.1.35.. Fix $a \in E$. If $b, c \in E$, there is an isometry from the span of $\{a, b, c\}$ to $\mathbb{R}^n$ for some $n \leq 2$ sending a to 0, and we define vector addition, scalar multiplication, norm, and inner product for b, c via this isometry. Since an isometry of $\mathbb{R}^n$ fixing the origin is linear and orthogonal, these operations are well defined. All except the inner product have simple intrinsic characterizations. (The inner product can be recovered from the norm by polarization ().) We have $\|b\| = \rho(a, b)$. If $\alpha \in \mathbb{R}$ and $b \neq a$, then αb is the unique element c on the line in E through a and b with $\rho(a, c) = |\alpha|\rho(a, b)$ and a between c and b if and only if $\alpha < 0$. If m is the midpoint of b and c in E, then $b + c = 2m$. Thus this structure is the only possible vector space structure and norm compatible with ρ.

To verify that E is a normed vector space with these operations, the only axiom which is not obviously satisfied is associativity of vector addition. But if $b, c, d \in E$, then the span of $\{a, b, c, d\}$ is isometric to $\mathbb{R}^n$ for some $n \leq 3$, and the vector addition in E restricted to this subspace agrees with vector addition in this $\mathbb{R}^n$ if an isometry is chosen sending a to the origin, and associativity of vector addition follows.

Parts (ii) and (iii) are immediate from (i). 👍

The Dimension Axiom

Thus, to characterize $\mathbb{R}^n$ as a metric space, we need only one additional axiom insuring that the space is n-dimensional. The simplest way to do this is

VI.3.1.39. Dimension n Axiom: There is an independent set of $n+1$ elements of E which spans E.

The Dimension n Axiom can be rephrased in a purely metric way not requiring reference to the Unique Line Axiom.

VI.3.1.40. Definition. Let (E,ρ) be a metric space, and $(a_0,\dots,a_m)$ an $(m+1)$-tuple of points in E. The *Cayley-Menger determinant* $D(a_0,\dots,a_m)$ of $(a_0,\dots,a_m)$ is the determinant of the $(m+2)\times(m+2)$ matrix

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & \rho(a_0,a_1)^2 & \cdots & \rho(a_0,a_n)^2 \\ 1 & \rho(a_1,a_0)^2 & 0 & \cdots & \rho(a_1,a_m)^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \rho(a_m,a_0)^2 & \rho(a_m,a_1)^2 & \cdots & 0 \end{bmatrix}$$

VI.3.1.41. If $E=\mathbb{R}^n$, there is a constant $c_m=\frac{(-1)^{m+1}}{2^m(m!)^2}$ depending only on m such that $c_m D(a_0,\dots,a_m)$ is the square of the m-dimensional volume of the parallelepiped with edges $a_1-a_0,\dots,a_m-a_0$ (). In particular, it is nonzero if and only if $\{a_1-a_0,\dots,a_m-a_0\}$ is linearly independent. (The constant is not relevant for our uses of the Cayley-Menger determinant in this section, since we will be concerned only whether it is 0.)

VI.3.1.42. n-Determinant Axiom: The space E has an $(n+1)$-tuple $(a_0,\dots,a_n)$ with $D(a_0,\dots,a_n)\neq 0$, but $D(b_0,\dots,b_{n+1})=0$ for any $(n+2)$-tuple $(b_0,\dots,b_{n+1})$ in E.

The Characterization Theorems

The main result of this section is:

VI.3.1.43. Theorem. Let (E,ρ) be a metric space satisfying the Line Axiom and the Parallelogram (or Pythagorean) Axiom. Then the following are equivalent:

(i) (E,ρ) satisfies the Dimension n axiom.

(ii) (E,ρ) satisfies the n-Determinant Axiom.

(iii) (E,ρ) is isometric to $(\mathbb{R}^n,\rho_2)$.

Proof: (i) ⇒ (iii) is VI.3.1.35., (iii) ⇒ (ii) and the equivalence of (i) and (ii) for $\mathbb{R}^n$ is VI.3.1.41.. 👍

Combining this result with VI.3.1.13., we obtain perhaps the nicest metric characterization of n-dimensional Euclidean space:

VI.3.1.44. Corollary. Let (E, ρ) be a complete metric space satisfying the Between, Extension, Parallelogram (or Pythagorean), and n-Determinant Axioms. Then (E, ρ) is isometric to $(\mathbb{R}^n, \rho_2)$.

In fact, the Parallelogram Axiom can be replaced with a more complicated condition involving Cayley-Menger determinants of smaller order. This was the original metric characterization of $\mathbb{R}^n$ obtained by K. Menger in 1928. See [Blu70, Theorem 48.2].

VI.3.2. $\mathbb{R}^n$ is n-Dimensional

Everyone "knows" that $\mathbb{R}^n$ is n-dimensional. But what does this mean? Specifically, what does "dimension" mean? And how can $\mathbb{R}^n$ be distinguished from $\mathbb{R}^m$ for $m \neq n$ by its properties?

$\mathbb{R}^n$ is a set with several types of structure: algebraic, geometric, topological, and analytic. Efforts have been made to define and/or describe the "dimension" of $\mathbb{R}^n$ in terms of each of these types of structure, and various theories of dimension have been studied in different contexts; the only thing these various theories have in common is that they all assign the *same* number n to $\mathbb{R}^n$, justifying the use of the name "dimension" for all of them.

VI.3.2.1. Specifically, we have:

(i) The vector space dimension () of $\mathbb{R}^n$ as a real vector space is n.

(ii) The topological dimension () of $\mathbb{R}^n$ as a topological space is n. (In fact, there are several notions of topological dimension, all of which give dimension n to $\mathbb{R}^n$.)

(iii) The Hausdorff dimension () of $\mathbb{R}^n$ is n.

VI.3.2.2. There is one respect in which all Euclidean spaces are the same, however: cardinality. Contrary to Cantor's expectations (he first thought the cardinality of $\mathbb{R}^n$ should be $\aleph_n$), the cardinality of $\mathbb{R}^n$ is $2^{\aleph_0}$ for all n (). Thus some additional structure beyond set theory is needed to distinguish $\mathbb{R}^n$ from $\mathbb{R}^m$ for $n \neq m$.

VI.3.3. Exercises

VI.3.3.1. Let X be a set, and $a, b \in X$. Define a metric ρ on X by $\rho(a, b) = 2$ and $\rho(x, y) = 1$ for all other distinct pairs x, y. Then every element of $X \setminus \{a, b\}$ is between a and b.

VI.3.3.2. In $(\mathbb{R}^2, \rho_\infty)$, the set of points between $a = (0, 0)$ and $b = (2, 0)$ is

$$\{(t, s) : 0 < t < 2, |s| \leq \min(t, 2 - t)\}$$

which is the "diamond" with corners a, b, $(1, 1)$, and $(1, -1)$.

VI.3.3.3. Let (E, ρ) be the subset

$$\{(t, t) : 0 \leq t \leq 1\} \cup \{(t, t - 2) : 1 < t \leq 2\}$$

of $(\mathbb{R}^2, \rho_\infty)$. If $a = (0, 0)$ and $b = (2, 0)$, then for every t, $0 \leq t \leq 2$, there is a unique c_t between a and b with $\rho(a, c_t) = t$ and $\rho(c_t, b) = 2 - t$ (i.e. E has the Unique Interpolation Property), but there is no line segment in E with endpoints a and b.

VI.3.3.4. Let E be the unit sphere in $\mathbb{R}^3$.

(a) If E is given the induced metric from $\mathbb{R}^3$ (sometimes called the *chord metric*), and $a, b \in E$, then there are no points of E between a and b.

(b) If the distance between points a and b is defined to be the length of the shortest great circle arc through a and b, a metric is obtained called the *arc metric*. If a and b are not diametrically opposed, then there is a unique line segment with endpoints a and b. If a and b are diametrically opposed, every other point of E is between a and b. E is complete and satisfies the Between Axiom, but E contains no lines. It almost satisfies the Extension Axiom, except for pairs of points diametrically opposed.

The same results hold in $\mathbb{R}^n$ for any $n > 1$ (in particular, for the unit circle in $\mathbb{R}^2$).

VI.3.3.5. Let (E, ρ) be a metric space satisfying the Line Axiom. A subset S of E is *collinear* if there is a line in E containing S.

(a) Prove the following theorem (cf. [Blu70, Theorem 21.3]):

THEOREM. (E, ρ) satisfies the Unique Line Axiom if and only if (E, ρ) has the *Two-Triple Property*: whenever $a, b, c, d \in E$, $\{a, b, c\}$ is collinear, and $\{b, c, d\}$ is collinear, then $\{a, b, d\}$ is collinear and $\{a, c, d\}$ is collinear.

It is then a consequence of the theorem that the four lines necessarily coincide, and therefore $\{a, b, c, d\}$ is collinear.

(b) Show directly that if E satisfies the Weak Euclidean Four-Point Axiom, then E has the Two-Triple Property.

VI.3.3.6. Show that $(\mathbb{R}^2, \rho_p)$ () has the Unique Line Property for $1 < p < \infty$, but not the Parallelogram Property (or other properties of VI.3.1.27.) if $p \neq 2$. More generally, any strictly convex normed real vector space () has the Unique Line Property.

VI.3.3.7. Let (E, ρ) be a metric space satisfying the Unique Interpolation and Unique Extrapolation Axioms. Prove that E has the Unique Line Property, as follows. Let $a, b \in E$ with $a \neq b$, and set $\lambda = \rho(a, b)$.

(a) Show the following:

(i) For each $s \geq \lambda$, there is a unique $c \in E$ with $\rho(a, c) = s$ and $\rho(b, c) = s - \lambda$.

(ii) For each s, $0 \leq s \leq \lambda$, there is a unique $c \in E$ with $\rho(a, c) = s$ and $\rho(c, b) = \lambda - s$.

(iii) For each $s < 0$, there is a unique $c \in E$ with $\rho(a, c) = -s$ and $\rho(b, c) = \lambda - s$.

If $s \in \mathbb{R}$, write $\phi(s)$ for the corresponding point c.

(b) Show that $\rho(\phi(r), \phi(s)) = |s - r|$ for all $r, s \in \mathbb{R}$; we may assume $r \leq s$. Consider six separate cases:

(i) $d \leq r \leq s$.

(ii) $0 \leq r \leq d \leq s$.

(iii) $0 \leq r \leq s \leq d$.

(iv) $r \leq 0 < d \leq s$.

(v) $r \leq 0 \leq s \leq d$.

(vi) $r \leq s \leq 0$.

For case (i), consider the unique $b \in E$ with $\rho(a_1, b) = r - \lambda$ and $\rho(b, \phi(s)) = s - r$ by (a). Use the triangle inequality twice to show that $\rho(a_0, b) = r$, so $b = \phi(r)$. Give similar (or simpler) arguments in the other five cases.

VI.3.3.8. In metric geometry (), a stronger version of the Between Axiom is often used.

Midpoint Axiom: Every pair of points in (E, ρ) has a midpoint (VI.3.1.21.).

No uniqueness is assumed. A metric space satisfying the Midpoint Axiom obviously satisfies the Between Axiom.

(a) Prove directly that a complete metric space satisfying the Midpoint Axiom satisfies the Line Segment Axiom, i.e. is strictly intrinsic. (Your proof should not need the AC, but will need the Axiom of Dependent Choice since midpoints are not necessarily unique.)

(b) [Blu70, Theorem 14.1] Adapt this proof to give a proof of Lemma VI.3.1.14. using only DC.

It is known that every Peano space () has a metric satisfying the Midpoint Axiom, or, equivalently, the Between Axiom or the Line Segment Axiom ([Bin49], [Moi49]; note that this result is incorrectly stated in [HY88, p. 130]).



Figure VI.5: A favorite subject in Croatia

VI.4. Length Structures and Metric Geometry

Metric spaces can be regarded as a first approximation to geometry (this is of course not the only interpretation or use for them!) True geometry has at least two other features in addition to a metric structure, i.e. a notion of distance between points:

(i) A notion of "shortest path," curves in the space which at least locally determine distance, i.e. the distance between nearby points is measured along such a path. Such paths, generalizing lines in Euclidean space or great circles on a sphere, are called *geodesics*. For a given point p, any other point q sufficiently close to p should be connected to p by a unique geodesic giving the metric distance between p and q, and geodesics should be extendible.

(ii) A notion of "angle" between geodesics intersecting at a point. Implicit in this is the idea that each geodesic through a point p determines a well-defined "direction" ("tangent") at p. Angles should locally give well-defined notions of area and higher-dimensional volumes, which can be extended globally.

There are at least three approaches possible to defining the additional structure, of increasing generality. First, for Euclidean space there is an "elementary" interpretation of lines and angles (although a fully rigorous treatment is more complicated than it may at first appear; cf. ()), so for "nice" subsets of Euclidean space, e.g. for smooth surfaces in $\mathbb{R}^3$, geodesics and angles can be defined analytically. This approach is described in (). Secondly, the case of smooth surfaces can be abstracted into the notion of smooth manifolds with a Riemannian or similar structure; cf. (). Finally, the notions can be directly developed straight from the general notion of metric space. This is the approach taken in this section.

As one might expect, abstract metric spaces are too general to support geometry with reasonable properties, and additional conditions on the space must be assumed. These additional conditions will be seen as the theory is developed. Ultimately, the additional conditions force the space to essentially be the type of structure from the second approach above; it is significant that the same mathematical structures can be arrived at in these two ways, reinforcing the naturality and importance of the resulting structures.

We will only cover the basic aspects of this approach. A far more extensive and detailed treatment can be found in [BBI01] and in [Gro99] (read and study the first before tackling the second!)

VI.4.1. Arc Length in Metric Spaces

The notion of curves and arc length in Euclidean space () has a straightforward generalization to metric spaces. Throughout this section, we will fix a metric space (X, ρ).

Parametrized Curves

VI.4.1.1. Definition. A *parametrized curve* in X is a continuous function $\gamma : [a, b] \to X$ for some closed bounded interval $[a, b]$ in $\mathbb{R}$. The points $p = \gamma(a)$ and $q = \gamma(b)$ are the *initial* and *final endpoints* of γ respectively, and γ is a *parametrized curve from p to q*.

Note that the interval $[a, b]$ is allowed to be degenerate, i.e. we allow $a = b$. (To be sure, parametrized curves defined on degenerate intervals are not very interesting!)

VI.4.1.2. A parametrized curve has a direction. If $\gamma : [a, b] \to X$ is a parametrized curve from p to q, its *negative* or *reversal* is $-\gamma : [a, b] \to X$ defined by $(-\gamma)(t) = \gamma(a + b - t)$ is a parametrized curve from q to p.

VI.4.1.3. We may think of the range of a parametrized curve as a "curve" in X (although it need not look anything like a geometric curve), and the parametrization as giving "motion" along this curve. This motion may "retrace" parts of the curve in a complicated way.

Restriction and Concatenation

VI.4.1.4. DEFINITION. Let $\gamma : [a,b] \to X$ be a parametrized curve, and $[c,d]$ a closed subinterval of $[a,b]$, i.e. $a \le c \le d \le b$. The parametrized curve $\gamma|_{[c,d]}$ is the *restriction* of γ to $[c,d]$; it is a parametrized curve from $\gamma(c)$ to $\gamma(d)$.

VI.4.1.5. DEFINITION. If $\gamma_1 : [a,b] \to X$ is a parametrized curve from p to q, and $\gamma_2 : [b,c] \to X$ is a parametrized curve from q to r, the *concatenation* is the parametrized curve $\gamma = \gamma_1 + \gamma_2 : [a,c] \to X$ defined by

$$\gamma(t) = \begin{cases} \gamma_1(t) & \text{if} \quad a \le t \le b \\ \gamma_2(t) & \text{if} \quad b \le t \le c \end{cases}$$

The concatenation γ is a parametrized curve from p to r.

We may extend the definition to $\gamma_1 + \cdots + \gamma_n$ if the curves line up properly.

VI.4.1.6. Note the conditions under which the concatenation is defined are rather severe: γ_1 and γ_2 must be parametrized on intervals in $\mathbb{R}$ which concatenate, and the final endpoint of γ_1 must be the initial endpoint of γ_2. Concatenation can be slightly generalized using reparametrizations (VI.4.1.16.). Also note that although additive notation is used, concatenation is not commutative: if $\gamma_1 + \gamma_2$ is defined, then $\gamma_2 + \gamma_1$ is not even defined unless one of the intervals is degenerate.

VI.4.1.7. If $\gamma = \gamma_1 + \gamma_2$ is defined, then the restriction of γ to $[a,b]$ (resp. to $[b,c]$) is γ_1 (resp. γ_2). Conversely, if $\gamma : [a,b]$ is a parametrized curve and $c \in [a,b]$, then $\gamma = \gamma|_{[a,c]} + \gamma|_{[c,b]}$. More generally, if $\mathcal{P} = \{a = t_0, t_1, \ldots, t_n = b\}$ is a partition of $[a,b]$, we have

$$\gamma = \gamma|_{[a,t_1]} + \gamma|_{[t_1,t_2]} + \cdots + \gamma|_{[t_{n-1},b]} \ .$$

Reparametrization

VI.4.1.8. DEFINITION. Let $\gamma : [a,b] \to X$ and $\tilde{\gamma} : [c,d] \to X$ be parametrized curves. Then $\tilde{\gamma}$ is a *reparametrization* of γ if $\tilde{\gamma} = \gamma \circ \phi$, where ϕ is a nondecreasing continuous function from $[c,d]$ onto $[a,b]$ (hence $\phi(c) = a$ and $\phi(d) = b$). If ϕ is strictly increasing, $\tilde{\gamma}$ is a *strict reparametrization* of γ.

VI.4.1.9. A reparametrization of a parametrized curve has the same initial and final endpoints, and the same range. But two parametrized curves with the same range, even also with the same initial and final endpoints, need not be reparametrizations. A reparametrization of a parametrized curve describes "motion" along the range "curve" in the same sense, with the same retracing, but possibly over a different interval of time at a different and varying speed.

VI.4.1.10. Various special kinds of reparametrizations can be considered. The simplest is linear reparametrizations, where ϕ is a linear (affine) function from $[c,d]$ onto $[a,b]$, i.e.

$$\phi(t) = a + \frac{b-a}{d-c}(t-c) \ .$$

One can also consider $\mathcal{C}^r$ reparametrizations, where ϕ is $\mathcal{C}^r$, and strict $\mathcal{C}^r$ reparametrizations, where both ϕ and ϕ^{-1} are $\mathcal{C}^r$ (these are also called smooth reparametrizations). There are also piecewise linear and piecewise smooth reparametrizations.

VI.4.1.11. Reparametrization is not an equivalence relation: it is reflexive and transitive, but not symmetric since the function ϕ need not be one-to-one. Strict reparametrization is symmetric, thus an equivalence relation on parametrized curves. Parametrization does generate an important equivalence relation, which can be described thus: two parametrized curves are equivalent if they are both reparametrizations of the same parametrized curve (VII.1.2.7.).

Nonstationary Curves

VI.4.1.12. Definition. Let $\gamma : [a,b] \to X$ be a parametrized curve, and $[c,d]$ a closed subinterval of $[a,b]$. Then γ is *stationary* on $[c,d]$ if $\gamma(t) = \gamma(c)$ for all $t \in [c,d]$. The curve γ is *nonstationary* if it is not stationary on any (nondegenerate) subinterval of $[a,b]$.

VI.4.1.13. A nonstationary parametrized curve need not be one-to-one; in fact, it can be very noninjective. For example, the space-filling curve of XIII.18.6.14. is nonstationary (the one in XIII.18.6.13. is not nonstationary). For a simpler example, a nonstationary parametrized curve can wrap around a circle in $\mathbb{R}^2$ several times.

VI.4.1.14. The maximal intervals on which a parametrized curve γ is stationary form a collection of disjoint closed subintervals of $[a,b]$. Every parametrized curve is a reparametrization of a nonstationary parametrized curve, which is unique up to strict reparametrization (Exercise ()), obtained by collapsing each maximal stationary subinterval to a point (there may be infinitely many such intervals!) Two parametrized curves are equivalent in the equivalence relation generated by reparametrization (VI.4.1.11.) if and only if their nonstationary versions are strict reparametrizations of each other.

Paths

VI.4.1.15. Definition. Let $p, q \in X$. A *path* from p to q is an equivalence class of parametrized curves from p to q (using the equivalence relation of VI.4.1.11. or VII.1.2.7.).

We will write $\gamma_1 \approx \gamma_2$ if γ_1 and γ_2 are equivalent, and write $\langle\gamma\rangle$ for the equivalence class of γ (the usual notation $[\gamma]$ for an equivalence class is normally reserved for the homotopy class of γ, which is a quite different thing ()).

VI.4.1.16. Paths can be concatenated: if $\langle\gamma_1\rangle$ is a path from p to q and $\langle\gamma_2\rangle$ is a path from q to r, then γ_1 can be concatenated with a suitable reparametrization $\tilde{\gamma}_2$ of γ_2 (e.g. a linear reparametrization), and the equivalence class of $\gamma_1 + \tilde{\gamma}_2$ is well defined and depends only on the classes of γ_1 and γ_2 (Exercise ()). Thus the concatenation of the paths $\langle\gamma_1\rangle + \langle\gamma_2\rangle$ is well defined as $\langle\gamma_1 + \tilde{\gamma}_2\rangle$.

Restriction of paths is a much dicier matter ().

Arc Length

So far we have not used the metric ρ (except to give the topology on X, i.e. to determine continuity of parametrized curves). We now bring ρ into the picture.

VI.4.1.17. DEFINITION. Let $\gamma : [a,b] \to X$ be a parametrized curve, and $\mathcal{P} = \{a = t_0, t_1, \ldots, t_n = b\}$ a partition of $[a,b]$. The *polygonal approximate length of* γ *from* $\mathcal{P}$ is

$$\ell_\rho(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \rho(\gamma(t_{k-1}), \gamma(t_k)) \ .$$

The *arc length* of γ is

$$\ell_\rho(\gamma) = \sup_{\mathcal{P}} \ell_\rho(\gamma, \mathcal{P}) \ .$$

We often suppress the subscript if ρ is understood and there is no possibility of confusion.

VI.4.1.18. If $\mathcal{Q}$ is a partition of $[a,b]$ which refines $\mathcal{P}$, we have $\ell(\gamma, \mathcal{P}) \le \ell(\gamma, \mathcal{Q})$ by the triangle inequality. Since any two partitions have a common refinement, the polygonal approximations approach the arc length as the partitions become arbitrarily fine. The term "polygonal approximation" should not be taken too literally: there is usually not an actual "polygon" in X approximating the curve. The term is suggestive that $\rho(p,q)$ should be the "straight-line" distance between p and q.

VI.4.1.19. The arc length of a curve can easily be infinite (e.g. VI.4.1.20.). The curve γ is *rectifiable* if $\ell(\gamma)$ is finite. Arc length and rectifiability depend crucially on ρ (VI.4.1.20.(ii)).

VI.4.1.20. EXAMPLES. (i) Let $X = [0,1]$ with the ordinary metric $\rho(x,y) = |x-y|$. A parametrized curve in X is a continuous function $\gamma : [a,b] \to [0,1]$. If $\mathcal{P}$ is a partition of $[a,b]$, then $\ell(\gamma, \mathcal{P})$ is the total variation $V(\gamma, \mathcal{P})$ of γ over $\mathcal{P}$ (XIX.4.1.3.), and $\ell(\gamma)$ is the total variation $V_{[a,b]}(\gamma)$ of γ over $[a,b]$ (XIX.4.1.6.). The curve γ is rectifiable if and only if γ has bounded variation. If γ is a smooth function, then γ is rectifiable and

$$\ell(\gamma) = \int_a^b |\gamma'(t)| \, dt \ .$$

(ii) Let $X = [0,1]$, and let $\gamma : [0,1] \to X$ be the identity function $\gamma(t) = t$. If we give X the usual metric ρ, we have $\ell_\rho(\gamma) = 1$. But if $0 < s < 1$ and we give X the snowflake metric ρ^s, where $\rho^s(x,y) = |x-y|^s$ (), then for the partition $\mathcal{P} = \left\{ \frac{k}{n} : 0 \le k \le n \right\}$ of $[0,1]$, we have

$$\ell_{\rho^s}(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \frac{1}{n^s} = \frac{n}{n^s} = n^{1-s}$$

so $\ell_{\rho^s}(\gamma) = +\infty$, i.e. γ is not rectifiable. In fact, in the metric space (X, ρ^s), no nonconstant parametrized curve is rectifiable.

We now develop basic properties of arc length. First note that arc length is independent of parametrization:

VI.4.1.21. PROPOSITION. Let $\tilde{\gamma} : [c,d] \to X$ be a reparametrization of $\gamma : [a,b] \to X$, i.e. $\tilde{\gamma} = \gamma \circ \phi$, $\phi : [c,d] \to [a,b]$. Then $\ell(\tilde{\gamma}) = \ell(\gamma)$.

PROOF: Let $\mathcal{P} = \{c = t_0, \ldots, t_n = d\}$ be a partition of $[c,d]$. Then $\{a = \phi(t_0), \ldots, \phi(t_n) = b\}$ is not quite a partition of $[a,b]$ since there may be repetitions if ϕ is not injective. But repeated points contribute nothing to the sum $\ell(\tilde{\gamma}, \mathcal{P})$, and if $\mathcal{Q}$ is the partition of $[a,b]$ obtained by deleting repetitions, we have $\ell(\gamma, \mathcal{Q}) = \ell(\tilde{\gamma}, \mathcal{P})$. Every partition of $[a,b]$ arises in this manner. 👍

VI.4.1.22. Thus arc length of paths is well defined.

Although the reversal $-\gamma$ of a parametrized curve is not a reparametrization of γ, we nonetheless have:

VI.4.1.23. PROPOSITION. If $\gamma : [a,b] \to X$ is a parametrized curve in X, then $\ell(-\gamma) = \ell(\gamma)$.

PROOF: If $\mathcal{P}$ is a partition of $[a,b]$, let $\tilde{\mathcal{P}}$ be the partition consisting of $\{a + b - t : t \in \mathcal{P}\}$. Then $\tilde{\mathcal{P}}$ is also a partition of $[a,b]$, and every partition of $[a,b]$ arises in this manner. We have $\ell(-\gamma, \tilde{\mathcal{P}}) = \ell(\gamma, \mathcal{P})$. 👍

VI.4.1.24. If $\gamma : [a,b] \to X$ is a parametrized curve from p to q and $\mathcal{P} = \{a,b\}$ is the trivial partition, we have

$$\rho(p,q) = \ell(\gamma, \mathcal{P}) \leq \ell(\gamma) \ .$$

More generally, we have:

VI.4.1.25. PROPOSITION. Let $\gamma : [a,b] \to X$ be a parametrized curve in X, and $a \leq c \leq d \leq b$, then

$$\rho(\gamma(c), \gamma(d)) \leq \ell(\gamma|_{[c,d]}) \leq \ell(\gamma) \ .$$

PROOF: The first inequality has already been noted. For the second, any partition $\mathcal{P}$ of $[c,d]$ can be expanded to a partition $\mathcal{Q}$ of $[a,b]$ by adding a and b if necessary. We have

$$\ell(\gamma|_{[c,d]}, \mathcal{P}) \leq \ell(\gamma, \mathcal{Q}) \leq \ell(\gamma)$$

and taking the supremum over all $\mathcal{P}$ gives the inequality. 👍

If γ is a constant path, obviously $\ell(\gamma) = 0$. Conversely:

VI.4.1.26. COROLLARY. Let γ be a parametrized curve in X. If $\ell(\gamma) = 0$, then γ is constant.

PROOF: If γ is not constant, there are $c, d \in [a,b]$, $c < d$, with $\gamma(c) \neq \gamma(d)$. Then $0 < \rho(\gamma(c), \gamma(d)) \leq \ell(\gamma)$. 👍

Here is a more precise version of VI.4.1.25.:

VI.4.1.27. PROPOSITION. Let $\gamma : [a,b] \to X$ be a parametrized curve, and $c \in [a,b]$. Then

$$\ell(\gamma) = \ell(\gamma|_{[a,c]}) + \ell(\gamma|_{[c,b]}) \ .$$

PROOF: If $\mathcal{P}$ is a partition of $[a,c]$ and $\mathcal{Q}$ a partition of $[c,b]$, then $\mathcal{P} \cup \mathcal{Q}$ is a partition of $[a,b]$, and

$$\ell(\gamma, \mathcal{P} \cup \mathcal{Q}) = \ell(\gamma|_{[a,c]}, \mathcal{P}) + \ell(\gamma|_{[c,b]}, \mathcal{Q}) \ .$$

Every partition of $[a,b]$ containing c (i.e. every sufficiently fine partition) arises in this manner. 👍

Iterating, we get:

VI.4.1.28. COROLLARY. Let $\gamma : [a,b] \to X$ be a parametrized curve, and $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$ a partition of $[a,b]$. Then

$$\ell(\gamma) = \sum_{k=1}^{n} \ell(\gamma|_{[t_{k-1}t_k]}) \ .$$

Rephrasing VI.4.1.27., we get:

VI.4.1.29. PROPOSITION. Let $\gamma_1 : [a,b] \to X$ and $\gamma_2 : [b,c] \to X$ with $\gamma_1(b) = \gamma_2(b)$, so that the concatenation $\gamma_1 + \gamma_2$ is defined. Then

$$\ell(\gamma_1 + \gamma_2) = \ell(\gamma_1) + \ell(\gamma_2) \ .$$

This can also be iterated to apply to arbitrary finite concatenations.

A crucial property of arc length is lower semicontinuity:

VI.4.1.30. PROPOSITION. For each n, let $\gamma_n : [a,b] \to X$ be a parametrized curve from p_n to q_n. Suppose (γ_n) converges pointwise to a parametrized curve $\gamma : [a,b] \to X$ from p to q, i.e. $\gamma_n(t) \to \gamma(t)$ for each $t \in [a,b]$ (so $p_n \to p$, $q_n \to q$). Then

$$\ell(\gamma) \leq \liminf_{n\to\infty} \ell(\gamma_n) \ .$$

We can have strict inequality, even if $\gamma_n \to \gamma$ uniformly on $[a,b]$ (I.6.1.16.).

PROOF: First note that if $\mathcal{P} = \{a = t_0, \ldots, t_m = b\}$ is any partition of $[a, b]$, we have

$$\ell(\gamma_n, \mathcal{P}) = \sum_{k=1}^{m} \rho(\gamma_n(t_{k-1}), \gamma_n(t_k)) \to \sum_{k=1}^{m} \rho(\gamma(t_{k-1}), \gamma(t_k)) = \ell(\gamma, \mathcal{P})$$

by continuity of ρ.

Suppose $\ell(\gamma) = +\infty$. Fix $M > 0$. Then there is a partition $\mathcal{P}$ of $[a, b]$ with $\ell(\gamma, \mathcal{P}) > M$. Then

$$M < \ell(\gamma, \mathcal{P}) = \lim_{n\to\infty} \ell(\gamma_n, \mathcal{P}) \leq \liminf_{n\to\infty} \ell(\gamma_n) \ .$$

This holds for all $M > 0$, so $\liminf_{n\to\infty} \ell(\gamma_n) = +\infty$ and the inequality holds.

Now suppose γ is rectifiable. Let $\epsilon > 0$. Then there is a partition $\mathcal{P}$ of $[a, b]$ with $\ell(\gamma, \mathcal{P}) > \ell(\gamma) - \epsilon$. Then

$$\ell(\gamma) - \epsilon < \ell(\gamma, \mathcal{P}) = \lim_{n\to\infty} \ell(\gamma_n, \mathcal{P}) \leq \liminf_{n\to\infty} \ell(\gamma_n) \ .$$

This holds for all $\epsilon > 0$, so $\liminf_{n\to\infty} \ell(\gamma_n) \geq \ell(\gamma)$. 👍

VI.4.1.31. Note for future reference that the definitions of $\ell(\gamma, \mathcal{P})$ and $\ell(\gamma)$ make perfect sense for a general function $\gamma : [a, b] \to X$, not necessarily continuous, so "arc length" is defined even for discontinuous functions (although it is typically infinite for very discontinuous functions). None of the proofs of VI.4.1.21.–VI.4.1.30. use that γ is continuous, so these properties hold also for discontinuous γ. (But VI.4.1.33. requires continuity of γ.)

Continuity of Arc Length

VI.4.1.32. Let $\gamma : [a, b] \to X$ be a rectifiable parametrized curve. Define a function $\phi : [a, b] \to [0, \ell(\gamma)]$ by setting

$$\phi(t) = \ell(\gamma|_{[a,t]}) \ .$$

Then ϕ is nondecreasing, $\phi(a) = 0$, $\phi(b) = \ell(\gamma)$. If γ is nonstationary, ϕ is strictly increasing.

VI.4.1.33. PROPOSITION. The function ϕ is continuous.

PROOF: The following lemma will provide left continuity on $[a, b]$. The simplest way to get right continuity is to apply the lemma also to $-\gamma$ and use that $\ell(\gamma|_{[a,t_0]}) = \ell(\gamma) - \ell(\gamma|_{[t_0,b]})$. 👍

VI.4.1.34. LEMMA. Let $\gamma : [a, b] \to X$ be a parametrized curve, not necessarily rectifiable. If (t_n) is an increasing sequence in $[a, b]$ converging to t_0, then

$$\ell(\gamma|_{[a,t_0]}) = \lim_{n\to\infty} \ell(\gamma|_{[a,t_n]}) = \sup_n \ell(\gamma|_{[a,t_n]}) \ .$$

PROOF: Define a sequence (γ_n) of parametrized curves on $[a,t_0]$ by setting $\gamma_n(t)=\gamma(t)$ if $a\le t\le t_n$ and $\gamma_n(t)=\gamma(t_n)$ if $t_n\le t\le t_0$ (i.e. γ_n agrees with γ on $[a,t_n]$ and is stationary on $[t_n,t_0]$). We have

$$\ell(\gamma_n)=\ell(\gamma|_{[a,t_n]})\le\ell(\gamma|_{[a,t_0]})$$

for any n by monotonicity (VI.4.1.25.). On the other hand, since γ is continuous at t_0, $\gamma_n\to\gamma$ pointwise on $[a,t_0]$. So applying VI.4.1.30. we get

$$\sup_n \ell(\gamma|_{[a,t_n]})\le\ell(\gamma|_{[a,t_0]})\le\liminf_{n\to\infty}\ell(\gamma_n)=\liminf_{n\to\infty}\ell(\gamma|_{[a,t_n]})$$

and the result follows. 👍

Parametrization by Arc Length

VI.4.1.35. Let $\gamma:[a,b]\to X$ be a nonstationary rectifiable parametrized curve, and let ϕ be the above function. Then ϕ is strictly increasing and continuous, so has a continuous strictly increasing inverse function $\phi^{-1}:[0,\ell(\gamma)]\to[a,b]$. Set $\tilde\gamma=\gamma\circ\phi^{-1}:[0,\ell(\gamma)]\to X$. Then $\tilde\gamma$ is a strict reparametrization of γ with the property that $\ell(\tilde\gamma|_{[0,t]})=t$ for $0\le t\le\ell(\gamma)$, i.e. $\tilde\gamma$ "traverses the path $\langle\gamma\rangle$ at uniform unit speed." The parametrized curve $\tilde\gamma$ is called the *reparametrization of γ by arc length.* If γ is replaced by any strict reparametrization, the $\tilde\gamma$ is unchanged; thus it depends only on $\langle\gamma\rangle$, i.e. any rectifiable path has a unique canonical parametrization by arc length.

Length of Arcs

If C is a "curve" in X, we would like to be able to say what the intrinsic "arc length" of C is. This is a tricky problem. The first question is what kind of subsets should be called "curves." It is too broad to say that the range of any parametrized curve is a "curve" (). One reasonable possibility is:

VI.4.1.36. DEFINITION. An *arc* in X is a subset of X homeomorphic to $[0,1]$.

VI.4.1.37. A subset C of X is an arc if and only if it is the range of an injective parametrized curve: a homeomorphism γ from $[0,1]$ to C is such a parametrized curve, and conversely an injective parametrized curve $\gamma:[a,b]\to X$ with $C=\gamma([a,b])$ is a homeomorphism by (). An arc thus has a unique equivalence class of oriented injective parametrizations, thus two in all; the arc length of all these parametrized curves is the same and can be taken to be the arc length of C.

VI.4.1.38. Slightly more generally, if C can be written as a finite union of arcs with only endpoints in common, the arc length of C can be taken to be the sum of the arc lengths of these arcs. For more complicated subsets, there are grave difficulties in defining arc length this way.

VI.4.1.39. There is a way to unambiguously define "arc length" for general (Borel) subsets of X: one-dimensional Hausdorff measure (). This "arc length" has the nice properties of a measure, generalizing VI.4.1.28.. Subsets whose one-dimensional Hausdorff measure is positive and finite can be regarded as "one-dimensional" (cf. (); not every "one-dimensional" subset will have these properties).

VI.4.2. Length Metrics

VI.4.2.1. The metric distance between two points in a metric space may not have much relationship to the "traveling distance." Here are some concrete examples, in part adapted from [BBI01].

(1) What is the distance between New York and Sydney? The straight-line distance is about 11,000 km. This number is relevant if one is, say, triangulating observational data to compute the distance from Earth to Mars, but is useless for travel since it would involve tunneling through the earth. The great circle distance, which is about 16,000 km, is much more relevant.

(2) What is the distance between the top of the Eiger and the top of the Jungfrau? It is about 5 km "as the crow flies." But a hiker/climber would have to go down into the valley and around or over the Mönch to get from one to the other, a much longer path. And the proper measure of the length of a route might be the time it takes to traverse it rather than the distance traveled.

(3) What is the distance between Boston and San Francisco? Even the great circle distance is not too relevant if one is driving; the road distance is the important consideration. So only certain paths are allowable for driving. And driving time might again be more important than distance – a longer route on interstate highways might be faster.

(4) On a more abstract level, if one lived in the Cantor set, it would not be possible to "move" or "travel" from one point even to very nearby ones, if time is interpreted as a real number and motion must be continuous. Something like this could happen in the real world if people lived on a chain of small islands and did not have boats or know how to swim.

We will address (1), (2), and (4) in this section, and (3) in the next section.

Intrinsic Metrics

VI.4.2.2. We would like to describe when a metric actually represents "traveling distance" and show that every metric defines a new associated "traveling distance" metric. It will be convenient to work with extended metrics, i.e. ones allowed to take values in $[0, +\infty]$ (); in some of the literature on metric geometry, "metric" means "extended metric" (we will not make this convention). The next definition is reasonable:

VI.4.2.3. Definition. An extended metric ρ on a set X is *strictly intrinsic*, or a *strict length metric*, if, for every $p, q \in X$, there is a path γ from p to q in X with $\ell(\gamma) = \rho(p, q)$.

For technical reasons, it is convenient to slightly broaden the definition:

VI.4.2.4. Definition. An extended metric ρ on a set X is *intrinsic*, or a *length metric*, if, for every $p, q \in X$, we have

$$\rho(p, q) = \inf_{\gamma} \ell(\gamma)$$

where the infimum is over all paths γ from p to q in X (with the convention that $\inf(\emptyset) = +\infty$).

VI.4.2.5. A strict length metric is a length metric (VI.4.1.24.), but not conversely. If ρ is a strict length metric, then (X, ρ) is path-connected. If ρ is a length metric, then (X, ρ) is not necessarily path-connected; but if $p, q \in X$ and $\rho(p, q) = +\infty$, then p and q are in different connected components of (X, ρ) () and hence there is no path in X from p to q. So if ρ is an extended metric which is a length metric, the following are equivalent:

(i) ρ is a metric.

(ii) (X, ρ) is path-connected.

(iii) (X, ρ) is *rectifiably path-connected*, i.e. for any $p, q \in X$, there is a rectifiable path from p to q in X.

In particular, a strictly intrinsic metric must be an actual metric, not an extended metric. The notions of length and strict length metrics are designed for and almost exclusively used in the (rectifiably) path-connected case. In fact, see VI.4.2.8..

VI.4.2.6. The term "intrinsic" is intended to mean that the metric represents "traveling distance within the space." For example, the arc-length metric on a circle is (strictly) intrinsic, but the Euclidean distance on a circle embedded in $\mathbb{R}^2$ is not intrinsic since one must travel "outside the space" to realize the Euclidean distance. Similarly, if the space is the surface of the earth (assumed to be a perfect sphere), the great-circle metric is intrinsic but not the straight-line (Euclidean) metric, which involves traveling through the earth and not within the space.

VI.4.2.7. Examples. (i) The Euclidean metric on $\mathbb{R}^n$ is strictly intrinsic.

(ii) More generally, the norm metric on any normed vector space is strictly intrinsic, as is the restriction to any convex subset.

(iii) As noted above, the arc length metric on a circle, or the great circle metric on a sphere, is strictly intrinsic. But the restriction of the Euclidean metric to an embedded circle or sphere is not intrinsic.

(iv) Let $X = \mathbb{R}^2 \setminus \{(0, 0)\}$. Then the restriction of the usual Euclidean metric to X is intrinsic but not strictly intrinsic: $(1, 0)$ and $(-1, 0)$ cannot be connected by a path of minimum length in X. See VI.4.5.1. for an example of a complete metric which is intrinsic but not strictly intrinsic.

(v) The discrete extended metric ρ on any set, where $\rho(p, q) = +\infty$ if $p \neq q$, is intrinsic.

VI.4.2.8. Proposition. Let ρ be an intrinsic metric on a set X. Then (X, ρ) is locally path-connected. In fact, every open ball in (X, ρ) is path-connected.

Proof: Let $p \in X$ and $\epsilon > 0$. Let $q \in B_\epsilon(p)$. Since $\rho(p, q) < \epsilon$, there is a parametrized curve $\gamma : [a, b] \to X$ from p to q with $\ell(\gamma) < \epsilon$. If $c \in [a, b]$, then by VI.4.1.25.

$$\rho(p, \gamma(c)) \leq \ell(\gamma|_{[a,c]}) \leq \ell(\gamma) < \epsilon$$

so the range of γ is contained in $B_\epsilon(p)$. Thus $B_\epsilon(p)$ is path-connected. 👍

The Length Metric of a Metric

Now suppose ρ is a general metric (or extended metric) on a space X. We want to define a new (extended) metric on X which describes "traveling distance" in (X, ρ).

VI.4.2.9. DEFINITION. Let (X, ρ) be a metric space. If $p, q \in X$, define

$$\rho_l(p, q) = \inf_\gamma \ell(\gamma)$$

where the infimum is over all paths γ from p to q in X (with the convention that $\inf(\emptyset) = +\infty$).

The function ρ_l is called the *length metric* of ρ; the terminology will be justified below.

VI.4.2.10. PROPOSITION. The function ρ_l is an extended metric on X.

PROOF: We need only verify the triangle inequality. Let $p, q, r \in X$; we show $\rho_l(p, r) \leq \rho_l(p, q) + \rho_l(q, r)$. If $\rho_l(p, q)$ or $\rho_l(q, r)$ is $+\infty$, there is nothing to prove, so assume both are finite. Let $\epsilon > 0$. There are paths γ_1 from p to q and γ_2 from q to r with $\ell(\gamma_1) < \rho_l(p, q) + \epsilon$ and $\ell(\gamma_2) < \rho_l(q, r) + \epsilon$. Reparametrizing if necessary, we can concatenate γ_1 and γ_2 (VI.4.1.16.) to get a path $\gamma_1 + \gamma_2$ from p to r. Thus

$$\rho_l(p, r) \leq \ell(\gamma_1 + \gamma_2) = \ell(\gamma_1) + \ell(\gamma_2) < \rho(p, q) + \rho(q, r) + 2\epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, the desired inequality follows. 👍

VI.4.2.11. Even if ρ is a metric, ρ_l is an extended metric in general. For example, if (X, ρ) is not path-connected, then ρ_l is necessarily extended: $\rho_l(p, q) = +\infty$ if p and q are in different path components. Even if (X, ρ) is path-connected, ρ_l can be an extended metric. In fact, ρ_l is a metric if and only if (X, ρ) is rectifiably path-connected.

VI.4.2.12. We have $\rho(p, q) \leq \rho_l(p, q)$ for any p, q by VI.4.1.24., i.e. ρ_l dominates ρ, so the topology of ρ_l is stronger than the topology of ρ. The metrics ρ and ρ_l are equivalent in good cases, but not in general, i.e. the ρ_l topology can be strictly stronger than the ρ-topology. If ρ is intrinsic, then $\rho_l = \rho$; the converse is also true (VI.4.2.14.).

VI.4.2.13. PROPOSITION. Let (X, ρ) be a metric space, and ρ_l the length metric of ρ. Then for any parametrized curve $\gamma : [a, b] \to X$ in (X, ρ), we have $\ell_{\rho_l}(\gamma) = \ell_\rho(\gamma)$.

Note that there is a subtlety: a continuous function from $[a, b]$ to (X, ρ) is not obviously continuous as a function from $[a, b]$ to (X, ρ_l) (and is not continuous in general). But by VII.3.1.15. arc length is also defined for discontinuous γ. In fact, we will not use continuity of γ in the proof.

PROOF: Fix γ. Since $\rho \leq \rho_l$, for any partition $\mathcal{P}$ of $[a, b]$ we have $\ell_\rho(\gamma, \mathcal{P}) \leq \ell_{\rho_l}(\gamma, \mathcal{P})$, and hence $\ell_\rho(\gamma) \leq \ell_{\rho_l}(\gamma)$. Conversely, let $\epsilon > 0$. If $\mathcal{P} = \{a = t_0, \ldots, t_n = b\}$ is a partition of $[a, b]$, set $p_k = \gamma(t_k)$ for $0 \leq k \leq n$. Then

$$\ell_{\rho_l}(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \rho_l(p_{k-1}, p_k) \leq \sum_{k=1}^{n} \ell_\rho(\gamma|_{[t_{k-1}, t_k]}) = \ell_\rho(\gamma)$$

(VI.4.1.28.), and taking the supremum over all $\mathcal{P}$ we obtain $\ell_{\rho_l}(\gamma) \leq \ell_\rho(\gamma)$. 👍

VI.4.2.14. Corollary. Let (X, ρ) be a metric space, and ρ_l the length metric of ρ. Then ρ_l is intrinsic. Thus $(\rho_l)_l = \rho_l$, and $\rho = \rho_l$ if and only if ρ is intrinsic.

VI.4.2.15. It can be shown using VI.4.2.13. and VI.4.1.33. that if γ is a rectifiable parametrized curve in (X, ρ), i.e. γ is a continuous function from $[a, b]$ to (X, ρ) with $\ell_\rho(\gamma) < +\infty$, then γ is also continuous as a function from $[a, b]$ to (X, ρ_l). This can be false if γ is not rectifiable (VI.4.2.16.(iv)).

VI.4.2.16. Examples. (i) If X is a circle embedded in $\mathbb{R}^2$ and ρ is the restriction of the Euclidean metric, then ρ_l is the arc length metric. Similarly, if X is a sphere embedded in $\mathbb{R}^3$ and ρ is the Euclidean metric, then ρ_l is the great circle metric. In these cases $\rho_l \neq \rho$, but the metrics are equivalent.

(ii) Let X be the Warsaw circle (XIII.19.5.6.), and ρ the restriction of the Euclidean metric in $\mathbb{R}^2$. Since (X, ρ) is not locally path-connected and (X, ρ_l) is (VI.4.2.8.), the metrics cannot be equivalent. In fact, (X, ρ_l) is homeomorphic to a half-open interval.

(iii) Let (X, ρ) be a totally disconnected metric space, e.g. the Cantor set with the usual metric. The metric ρ_l is the discrete extended metric (VI.4.2.7.(v)), so the topology from ρ_l is the discrete topology.

(iv) Let X be $[0, 1]$ with the snowflake metric ρ^s for some s, $0 < s < 1$ (VI.4.1.20.(ii)). Then X is path-connected and locally path-connected, but has no nonconstant rectifiable paths. The metric $(\rho^s)_l$ is the discrete extended metric, so the topology from ρ_l is the discrete topology. Thus the topology from ρ_l can depend on the metric ρ and not just on the topology from ρ.

Local Characterization of Intrinsic Metrics

Intrinsic metrics are characterized locally, and are in an appropriate sense the only locally determined metrics.

VI.4.2.17. Definition. Let ρ and σ be extended metrics on a set X. The metric ρ *locally dominates* σ if, for each $x \in X$, there is a ρ-open ball B_x centered at x such that $\sigma(y, z) \leq \rho(y, z)$ for all $y, z \in B_x$. The metric ρ *uniformly locally dominates* σ if there is an $\epsilon > 0$ such that, whenever B is a ball of ρ-radius ϵ in X, $\sigma \leq \rho$ on B.

The extended metric ρ is *locally determined* if, whenever σ is an extended metric which is locally dominated by ρ, then $\sigma \leq \rho$.

VI.4.2.18. Examples. (i) The usual metric on $\mathbb{R}$, or any subset of $\mathbb{R}$ (e.g. $\mathbb{Q}$), is locally determined.

(ii) The arc length metric σ on a circle in $\mathbb{R}^2$ is locally determined. But the Euclidean metric ρ is not, since ρ locally dominates $(1 - \epsilon)\sigma$ for every $\epsilon > 0$ but does not dominate if ϵ is small enough.

(iii) If a small open arc (less than half the circumference) is removed from a circle, the restriction of arc length to the remaining space is not locally determined.

(iv) The discrete metric () on any set X is locally determined.

VI.4.2.19. PROPOSITION. Let (X, ρ) be a metric space. If ρ is intrinsic, it is locally determined.

PROOF: Suppose ρ is intrinsic, and σ is a pseudometric which is locally dominated by ρ. Let $x, y \in X$; we show $\sigma(x, y) \leq \rho(x, y)$. We may assume $\rho(x, y) < +\infty$. Fix $\epsilon > 0$. Let $\gamma : [0, 1] \to X$ be a path from x to y with $\ell(\gamma) < \rho(x, y) + \epsilon$. For $z \in \gamma([0, 1])$, there is an open neighborhood U_z of z with $\sigma(u, v) \leq \rho(u, v)$ for all $u, v \in U_z$. Since $\gamma([0, 1])$ is compact, there are finitely many such U_z which cover it, so there is a partition $\{0 = t_0, \ldots, t_n = 1\}$ of $[0, 1]$ such that, if $x_k = \gamma(t_k)$, consecutive x_k are in the same U_z. Thus we have

$$\sigma(x, y) \leq \sum_{k=1}^{n} \sigma(x_{k-1}, x_k) \leq \sum_{k=1}^{n} \rho(x_{k-1}, x_k) \leq \ell(\gamma) < \rho(x, y) + \epsilon$$

and since $\epsilon > 0$ is arbitrary, $\sigma(x, y) \leq \rho(x, y)$. 👍

VI.4.2.20. The converse is not true, even if ρ is complete (VI.4.2.18.(iv)). To obtain an appropriate converse, we make the following construction. Let ρ be an extended metric on a set X and $\epsilon > 0$. Set

$$\rho_\epsilon(x, y) = \inf \sum_{k=1}^{n} \rho(x_{k-1}, x_k)$$

where the infimum is over all finite chains $\{x = x_0, x_1, \ldots, x_n = y\}$ where $\rho(x_{k-1}, x_k) \leq \epsilon$ for all k (set $\rho_\epsilon(x, y) = +\infty$ if there is no such chain). It is easily proved by an argument almost identical to () that ρ_ϵ is an extended metric, that $\rho \leq \rho_\epsilon \leq \rho_\delta$ if $\delta < \epsilon$. We have that $\rho_\epsilon(x, y) = \rho(x, y)$ if $\rho(x, y) \leq \epsilon$, and in particular ρ uniformly locally dominates ρ_ϵ for any $\epsilon > 0$.

VI.4.2.21. By (), $\rho_{loc} = \sup_{\epsilon > 0} \rho_\epsilon$ is an extended metric, called the *local metric* of ρ. Comparing definitions, we have $\rho_{loc} \leq \rho_l$. Since $(\rho_\epsilon)_\delta = \rho_\delta$ if $\delta \leq \epsilon$, it follows that

$$(\rho_\epsilon)_{loc} = (\rho_{loc})_\epsilon = (\rho_{loc})_{loc} = \rho_{loc}$$

for any $\epsilon > 0$.

The next proposition is an immediate corollary of VI.4.2.19., or it can be easily proved directly.

VI.4.2.22. PROPOSITION. Let (X, ρ) be a metric space. If ρ is intrinsic, then $\rho = \rho_\epsilon$ for all $\epsilon > 0$.

VI.4.2.23. DEFINITION. Let ρ be a metric on a set X. Then ρ is *uniformly locally determined* if $\rho = \rho_\epsilon$ for all $\epsilon > 0$, i.e. if $\rho = \rho_{loc}$.

VI.4.2.24. EXAMPLES. (i) The usual metric on $\mathbb{Q}$ (or any other subset of $\mathbb{R}$) is uniformly locally determined.

(ii) The discrete metric on a set with more than one element is locally determined but not uniformly locally determined.

(iii) For any metric ρ, ρ_{loc} is uniformly locally determined (VI.4.2.21.).

VI.4.2.25. THEOREM. Let (X, ρ) be a complete metric space. Then $\rho_l = \rho_{loc}$.

VI.4.2.26. COROLLARY. Let (X, ρ) be a complete metric space. If ρ is uniformly locally determined, it is intrinsic.

For the proof of the theorem, we will use a variant of the Between Axiom (VI.3.1.11.) which is of independent interest:

VI.4.2.27. DEFINITION. A metric space (X, ρ) has the *approximate midpoint property* if, for every $x, y \in X$ and $\epsilon > 0$ there is a $z \in X$ with

$$\left|\rho(x,z) - \frac{\rho(x,y)}{2}\right| < \epsilon \text{ and } \left|\rho(z,y) - \frac{\rho(x,y)}{2}\right| < \epsilon \ .$$

Such a z is called an *ϵ-midpoint* of x and y.

VI.4.2.28. PROPOSITION. A uniformly locally determined metric has the approximate midpoint property.

PROOF: If (X, ρ) is locally determined, let $x, y \in X$ and $\epsilon > 0$. If $\{x = x_0, \ldots, x_n = y\}$ is a finite chain with $\rho(x_{k-1}, x_k) < \frac{\epsilon}{2}$ for all k and $\sum_k \rho(x_{k-1}, x_k) < \rho(x) + \frac{\epsilon}{2}$, then one of the x_k (for $k \approx \frac{n}{2}$) will be an ϵ-midpoint of x and y. 👍

VI.4.2.29. THEOREM. Let (X, ρ) be a complete metric space with the approximate midpoint property. Then ρ is intrinsic.

PROOF: This is somewhat similar to (), but no form of choice is needed. Let $x, y \in X$ and $\epsilon > 0$. Let $x_{1/2}$ be an $\frac{\epsilon}{4}$-midpoint of x and y. Let $x_{1/4}$ be an $\frac{\epsilon}{16}$-midpoint of $x = x_0$ and $x_{1/2}$, and $x_{3/4}$ an $\frac{\epsilon}{16}$-midpoint of $x_{1/2}$ and $y = x_1$. Continue inductively in this manner to define x_r for dyadic rational $r \in [0, 1]$, using a denominator of 4^n if the denominator of r is 2^n. It is easy to check that $r \mapsto x_r$ is uniformly continuous and thus extends to a map $\gamma : [0, 1] \to X$ since ρ is complete. Then γ is a path from x to y, and $\ell(\gamma) \leq \rho(x, y) + \epsilon$. 👍

Theorem VI.4.2.25., and hence Corollary VI.4.2.26., follow immediately.

VI.4.2.30. The example of $\mathbb{Q}$ (VI.4.2.24.(i)) shows that completeness is necessary in the theorem.

VI.4.3. Length Structures

We now change point of view. Instead of working with arc length defined by a metric, we assume directly that we have a notion of length for certain ("admissible") paths in the space X and from it derive a length metric and thus a topology. We must start with a metric, or at least a topology, on X, in order to have a notion of continuous path in X, but the initial topology will be inconsequential at the end. In order to do this, there must be enough admissible paths, and the length function must have reasonable properties.

Our exposition is largely adapted from [BBI01].

VI.4.3.1. DEFINITION. Let (X, σ) be a metric space. A *length structure* on (X, σ) is a pair $L = (A, \ell)$, where A is a set of parametrized curves in X (called *admissible paths* for L) such that

(1) If $\gamma : [a, b] \to X$ is in A, then $\gamma|_{[c,d]} \in L$ for every subinterval $[c, d]$ of $[a, b]$.

(2) If $\gamma_1, \gamma_2 \in A$ and $\gamma_1 + \gamma_2$ is defined, then $\gamma_1 + \gamma_2 \in A$.

(3) Any linear reparametrization of a parametrized curve in A is also in A. (Combined with (3), any piecewise-linear reparametrization of a parametrized curve in A is then also in A.) Also, if $\gamma \in A$, then its reversal $-\gamma$ is in A too.

and $\ell : A \to [0, +\infty]$ (called the *length function* of L) satisfying

(i) If $\gamma : [a, b] \to X$ is in A and $c \in [a, b]$, then $\ell(\gamma) = \ell(\gamma|_{[a,c]}) + \ell(\gamma_{[c,b]})$ (i.e. ℓ is additive on concatenations).

(ii) If $\gamma : [a, b] \to X$ is in A, then the function $\phi(t) = \ell(\gamma_{[a,t]}$ is continuous on $[a, b]$.

(iii) If $\gamma \in A$ and $\tilde{\gamma}$ is a linear reparametrization of γ, then $\ell(\tilde{\gamma}) = \ell(\gamma)$. (From (i) it then follows that ℓ is invariant under piecewise-linear reparametrizations, i.e. ℓ is really defined on equivalence classes of parametrized curves.) Also, $\ell(-\gamma) = \ell(\gamma)$ for any $\gamma \in A$.

(iv) If $p \in X$ and U is an open neighborhood of p (in the σ-topology), then there is a $\delta > 0$ such that any $\gamma \in A$ from p to a point $q \in U^c$ satisfies $\ell(\gamma) \geq \delta$.

The set A of admissible paths is technically encoded in the length function ℓ as its domain, and some of the literature does not distinguish between L and ℓ; but it seems a little clearer to make a notational distinction.

VI.4.3.2. Condition (iv) is an additional compatibility condition of L with the σ-topology on X besides the requirement that the curves in A are continuous. These two compatibility conditions are the only place the metric σ appears, and only depend on the σ-topology; the metric σ could be dispensed with entirely and just replaced by an initial topology on X (which should be assumed Hausdorff for a reasonable theory, although it could be dispensed with entirely by taking it to be the indiscrete topology).

VI.4.3.3. EXAMPLES. (i) Let (X, ρ) be a metric space, A the collection of all parametrized curves in X, and $\ell(\gamma) = \ell_\rho(\gamma)$. This is the *length structure induced by* ρ, denoted L_ρ.

(ii) More generally, let (X, ρ) be a metric space and $Y \subseteq X$. Let A be the set of parametrized curves in X whose range is contained in Y, and for $\gamma \in A$ set $\ell(\gamma) = \ell_\rho(\gamma)$. Then (A, ℓ) is a length structure on (X, ρ), as well as on (Y, ρ). This was the main motivating example for the theory of length structures, with $X = \mathbb{R}^n$, ρ the Euclidean metric, and Y a smooth submanifold.

(iii) Let (X, ρ) be $\mathbb{R}^n$ with the Euclidean metric, and take A to be either the set of piecewise-linear curves or piecewise-smooth curves in $\mathbb{R}^n$, with $\ell = \ell_\rho|_A$. These length structures are not essentially different from the length structure using all parametrized curves. One can also consider the restrictions to suitable subsets as in (ii).

(iv) Let (X, ρ) be $\mathbb{R}^n$ with the Euclidean metric, and take A to be the set of all parametrized curves in $\mathbb{R}^n$. Set $\ell(\gamma) = \ell_\rho(\gamma)$ if γ is piecewise-linear and $\ell(\gamma) = +\infty$ otherwise.

(v) Let $X = \mathbb{R}^2$, σ the Euclidean metric. Let A be the collection of all parametrized curves whose image consists of a union of finitely many horizontal and vertical line segments. For $\gamma \in A$, set $\ell(\gamma) = \ell_\sigma(\gamma)$.

(vi) Here is a concrete example, stated somewhat imprecisely. Let X be the United States, and A the set of driving routes (so all points, including endpoints, are points on roads). The length $\ell(\gamma)$ of a driving route is the driving time for the route. (Driving routes are taken as paths, i.e. equivalence classes of parametrized curves, and driving time is not connected with parametrization.) This must be idealized to become precise and give a length structure: driving time must be well defined and be independent of direction, and it must be assumed that a vehicle can reverse direction instantaneously to make driving time additive on curves which retrace routes.

In this example, not all points of the space are connected by admissible paths. To rectify this, we could add to A the set of all routes traversable by surface transportation, including walking (assuming that every point of the country is within walking distance of a road).

(vii) As a precise theoretical version of (vi), let X be $\mathbb{R}^n$ (or a smooth submanifold) and $f : X \to (0, +\infty)$ a continuous function (continuity can be somewhat relaxed). Let A be the set of piecewise-smooth parametrized curves in X, and for $\gamma : [a, b] \to X$ in A, set

$$\ell(\gamma) = \int_\gamma f = \int_a^b f(\gamma(t))|\gamma'(t)|\, dt$$

(this is just the integral of f along the path defined by γ ()). In the case of (v), X could be the set of all points on roads, and $f(\boldsymbol{x})$ the reciprocal of the maximum speed one can (or is allowed to) drive at the point $\boldsymbol{x}$.

(viii) As a generalization, let X be $\mathbb{R}^n$ (or a smooth submanifold) and $f : X \times \mathbb{R}^n \to (0, +\infty)$ a continuous function (continuity can be somewhat relaxed). Let A be the set of piecewise-smooth parametrized curves in X, and for $\gamma : [a, b] \to X$ in A, set

$$\ell(\gamma) = \int_a^b f(\gamma(t), \gamma'(t))\, dt\ .$$

To make ℓ invariant under linear reparametrizations, we need $f(\boldsymbol{x}, \alpha\boldsymbol{y}) = |\alpha| f(\boldsymbol{x}, \boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y}, \alpha$. Such a length structure is called a *Finslerian length structure*. The idea is that the speed of travel may depend not only on the point, but also on the direction of travel (although the speed in diametrically opposite directions must be the same, not necessarily a realistic assumption in applications).

The Length Metric of a Length Structure

VI.4.3.4. Definition. Let $L = (A, \ell)$ be a length structure on a metric space (X, σ). Define an extended metric ρ_L by

$$\rho_L(p, q) = \inf_\gamma \ell(\gamma)$$

for $p \neq q$, where the infimum is over all paths $\gamma \in A$ from p to q (with the convention that $\inf(\emptyset) = +\infty$).

The proof that ρ_L is an extended metric is essentially identical to the proof of VI.4.2.10., using that a concatenation of admissible paths is admissible. The extended metric ρ_L is a metric if and only if any two points of X can be connected by an admissible path of finite length. The ρ_L-topology on X is stronger than the initial σ-topology by condition (iv); it is often strictly stronger.

VI.4.3.5. PROPOSITION. Let $L = (A, \ell)$ be a length structure on (X, σ), and ρ_L the corresponding extended metric. Then ρ_L is a length metric, and $\ell_{\rho_L}(\gamma) \leq \ell(\gamma)$ for all $\gamma \in A$.

We can have strict inequality (VI.4.3.6.(iv)).

PROOF: For the second statement, let $\gamma : [a, b] \to X$ be in A. If $\mathcal{P} = \{a = t_0, \ldots, t_n = b\}$ is any partition of $[a, b]$, we have

$$\ell_{\rho_L}(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \rho_L(\gamma(t_{k-1}), \gamma(t_k)) \leq \sum_{k=1}^{n} \ell(\gamma|_{[t_{k-1}, t_k]}) = \ell(\gamma)$$

since $\gamma|_{[t_{k-1}, t_k]} \in A$ by (1), and taking the supremum over all $\mathcal{P}$ we obtain the inequality.
To show ρ_L is a length metric, let $p, q \in X$. We show that

$$\rho_L(p, q) = \lambda := \inf_{\gamma} \ell_{\rho_L}(\gamma)$$

with the infimum over all parametrized curves from p to q. We need only show $\rho_L(p, q) \geq \lambda$, since $\leq$ is automatic (VI.4.1.24.). If $\rho_L(p, q) = +\infty$, there is nothing to prove, so we may suppose $\rho_L(p, q)$ is finite. Let $\epsilon > 0$. There is a $\gamma \in A$ from p to q with $\ell(\gamma) < \rho_L(p, q) + \epsilon$. Then

$$\lambda \leq \ell_{\rho_L}(\gamma) \leq \ell(\gamma) < \rho_L(p, q) + \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, the inequality follows. 👍

VI.4.3.6. EXAMPLES. (i) If (X, ρ) is a metric space and L_ρ is the induced length structure (VI.4.3.3.(i)), then $\rho_{L_\rho} = \rho_l$.

(ii) In the situation of VI.4.3.3.(ii), if L is regarded as a length structure on (Y, ρ), then $\rho_L = \rho_l$ as in (i). If L is regarded as a length structure on (X, ρ), then ρ_L is an extended metric on X, with $\rho_L(p, q) = +\infty$ unless both p and q are in Y. In the ρ_L topology on X, Y is a clopen set and each point of $X \setminus Y$ is an isolated point.

(iii) In VI.4.3.3.(iii), the induced metric ρ_L on $\mathbb{R}^n$ is just the Euclidean metric, since the set of admissible curves includes a curve of minimal length between any two points. Thus restricting the set of admissible curves does not necessarily change the length metric, so long as the admissible curves include arbitrarily short paths (i.e. close to the minimum for all paths) between any pair of points.

(iv) In VI.4.3.3.(iv), the induced metric ρ_L on $\mathbb{R}^n$ is again just the Euclidean metric. Thus, even if the set of admissible curves is the same, changing the length function ℓ does not necessarily change the length metric, if there are short paths of (approximately) the same length under the two length functions.

(v) In VI.4.3.3.(v), the length metric is not the Euclidean metric $\rho = \rho_2$, but rather the 1-metric ρ_1 (). Thus if the set of admissible curves is changed, the length metric can change if the set of admissible curves does not contain arbitrarily short paths between points.

Recovering a Length Structure from its Metric

VI.4.3.7. If ρ is a length metric, then the induced metric from the length structure L_ρ is ρ itself (cf. VI.4.3.6.(i)). How about the converse: if $L = (A, \ell)$ is a length structure and ρ_L the induced length metric, is $L = L_{\rho_L}$? In one sense, the answer is usually no: the set of admissible paths for L_{ρ_L} is often larger than A (cf. Examples VI.4.3.6.(iii) and (v)). But ℓ_{ρ_L} may not agree with ℓ even on A (VI.4.3.6.(iv); the problem is lower semicontinuity (VI.4.3.9.)). It would be useful to know when ℓ_{ρ_L} agrees with ℓ on A, since ℓ is often given by a formula making it easier to calculate than ℓ_{ρ_L} (e.g. in VI.4.3.3.(vii)–(viii)).

VI.4.3.8. The length function ℓ_{ρ_L} is always lower semicontinuous (VI.4.1.30.), so a necessary condition that $\ell = \ell_{\rho_L}|_A$ for a length structure $L = (A, \ell)$ is that ℓ is lower semicontinuous on A. But ℓ is not necessarily lower semicontinuous, e.g. the ℓ of Example VI.4.3.3.(iv) (the ℓ of Example VI.4.3.3.(viii) is also not necessarily lower semicontinuous, cf. ()). Lower semicontinuity is the only constraint:

VI.4.3.9. PROPOSITION. Let $L = (A, \rho)$ be a length structure on (X, σ), and ρ_L the corresponding length metric. Then $\ell = \ell_{\rho_L}|_A$ if and only if ℓ is lower semicontinuous on A.

VI.4.4. Metric Geometry

VI.4.5. Exercises

VI.4.5.1. For each n let I_n be a closed interval in $\mathbb{R}$ of length $1 + \frac{1}{n}$ with endpoints a_n and b_n. Glue the I_n together by identifying all the a_n to a point a, and all the b_n to a point b. Let X be the quotient space, and ρ the length metric on X.

(a) Show that (X, ρ) is a complete metric space.

(b) Show that $\rho(a, b) = 1$, but there is no path from a to b of length 1. Thus (X, ρ) is not complete as a length space, i.e. ρ is not strictly intrinsic.

(c) Realize X concretely as a subset of $\mathbb{R}^2$ consisting of a union of circular arcs in the upper half plane through $(0, 0)$ and $(1, 0)$. Show that ρ is the length metric defined by the restriction σ of the Euclidean metric on $\mathbb{R}^2$, and that ρ is equivalent (but not uniformly equivalent) to the incomplete metric σ.

VI.4.5.2. Let (X, ρ) be a metric space, and ρ_l the associated length metric on X.

(a) Show that if ρ is complete, then ρ_l is also complete.

(b) If ρ_l is complete, then ρ is not necessarily complete, even if (X, ρ) is rectifiably path-connected and locally path-connected. [Let X be the topologist's sine curve (XIII.19.5.5.) with the points on the y-axis removed, with ρ the ordinary metric from $\mathbb{R}^2$. $\mathbb{Q}$ with the ordinary metric is an example if the space is not required to be path-connected.]

Chapter VII

Curves

> "Everyone knows what a curve is, until he has studied enough mathematics to become confused through the countless number of possible exceptions."
>
> *Felix Klein*

What is a curve? Almost everyone has a good intuitive idea of what a curve should be, but it is a surprisingly delicate matter to make a precise definition which includes all the things which should be curves and excludes everything that should not.

At a minimum, a curve should be a one-dimensional subset of a Euclidean space. Most people would agree that the following are curves:

Graphs of continuous functions, or at least graphs of differentiable functions, from $\mathbb{R}$ to $\mathbb{R}$.

Lines, circles, and polygons.

A figure 8, figure 6, figure α, and the like.

Finite separated unions of the above.

On the other hand, the following are clearly not curves:

Anything zero-dimensional, e.g. the Cantor set ().

Anything more than one-dimensional, e.g. a (solid) disk or square, or a ball or sphere in $\mathbb{R}^n$, $n > 2$.

Anything not at least locally compact.

In addition, there are many things which are in a gray area, e.g. a figure X, figure Y, figure $\otimes$, and many of the the one-dimensional subsets of $\mathbb{R}^2$ and $\mathbb{R}^3$ discussed in (), such as the Topologist's Sine Curve, the Hawaiian Earring, the Sierpiński gasket and carpet, solenoids, the pseudo-arc, even wildly embedded arcs.

We will begin by defining curves via parametrizations, although we will distinguish between curves and parametrized curves. Parametrizations are natural since they describe "motion along the curve." But defining a curve as just a continuous image of an interval is too broad: for example, there are "space-filling curves," continuous functions from $[0, 1]$ onto a square $[0, 1]^2$. In fact, the Banach-Mazurkiewicz Theorem () says

that every compact metrizable topological space which is connected and locally connected, and in particular any connected, locally connected closed bounded subset of a Euclidean space, is a continuous image of $[0,1]$. Allowing continuous images of non-closed intervals, e.g. $\mathbb{R}$, admits even more pathology: a continuous image of $\mathbb{R}$, even a one-to-one continuous image, need not even be locally compact (). On the other hand, requiring the parametrization to be a homeomorphism eliminates such a natural curve as a circle, and to relax the requirement to being a local homeomorphism rules out curves like a figure 8. So we will have to be more careful.

VII.1. Curves and Parametrized Curves in $\mathbb{R}^n$

VII.1.1. Parametrized Curves

VII.1.1.1. DEFINITION. A *parametrized curve* in $\mathbb{R}^n$ is a continuous function $\gamma : [a,b] \to \mathbb{R}^n$ for some closed bounded interval $[a,b]$ in $\mathbb{R}$. The points $p = \gamma(a)$ and $q = \gamma(b)$ are the *initial* and *final endpoints* of γ respectively, and γ is a *parametrized curve from p to q*.

Note that the interval $[a,b]$ is allowed to be degenerate, i.e. we allow $a = b$. (To be sure, parametrized curves defined on degenerate intervals are not very interesting!)

VII.1.1.2. A parametrized curve has a direction. If $\gamma : [a,b] \to \mathbb{R}^n$ is a parametrized curve from p to q, its *negative* or *reversal* is $-\gamma : [a,b] \to \mathbb{R}^n$ defined by $(-\gamma)(t) = \gamma(a+b-t)$ is a parametrized curve from q to p.

VII.1.1.3. We may think of the range of a parametrized curve as a "curve" in $\mathbb{R}^n$ (although it need not look anything like a geometric curve), and the parametrization as giving "motion" along this curve. This motion may "retrace" parts of the curve in a complicated way.

Restriction and Concatenation

VII.1.1.4. DEFINITION. Let $\gamma : [a,b] \to \mathbb{R}^n$ be a parametrized curve, and $[c,d]$ a closed subinterval of $[a,b]$, i.e. $a \le c \le d \le b$. The parametrized curve $\gamma|_{[c,d]}$ is the *restriction* of γ to $[c,d]$; it is a parametrized curve from $\gamma(c)$ to $\gamma(d)$.

VII.1.1.5. DEFINITION. If $\gamma_1 : [a,b] \to \mathbb{R}^n$ is a parametrized curve from p to q, and $\gamma_2 : [b,c] \to \mathbb{R}^n$ is a parametrized curve from q to r, the *concatenation* is the parametrized curve $\gamma = \gamma_1 + \gamma_2 : [a,c] \to \mathbb{R}^n$ defined by

$$\gamma(t) = \begin{cases} \gamma_1(t) & \text{if} \quad a \le t \le b \\ \gamma_2(t) & \text{if} \quad b \le t \le c \end{cases}$$

The concatenation γ is a parametrized curve from p to r.

We may extend the definition to $\gamma_1 + \cdots + \gamma_n$ if the curves line up properly.

VII.1.1.6. Note the conditions under which the concatenation is defined are rather severe: γ_1 and γ_2 must be parametrized on intervals in $\mathbb{R}$ which concatenate, and the final endpoint of γ_1 must be the initial endpoint of γ_2. Concatenation can be slightly generalized using reparametrizations (VI.4.1.16.). Also note that although additive notation is used, concatenation is not commutative: if $\gamma_1 + \gamma_2$ is defined, then $\gamma_2 + \gamma_1$ is not even defined unless one of the intervals is degenerate.

VII.1.1.7. If $\gamma = \gamma_1 + \gamma_2$ is defined, then the restriction of γ to $[a, b]$ (resp. to $[b, c]$) is γ_1 (resp. γ_2). Conversely, if $\gamma : [a, b]$ is a parametrized curve and $c \in [a, b]$, then $\gamma = \gamma|_{[a,c]} + \gamma|_{[c,b]}$. More generally, if $\mathcal{P} = \{a = t_0, t_1, \ldots, t_n = b\}$ is a partition of $[a, b]$, we have

$$\gamma = \gamma|_{[a,t_1]} + \gamma|_{[t_1,t_2]} + \cdots + \gamma|_{[t_{n-1},b]} \ .$$

VII.1.2. Reparametrization

VII.1.2.1. Definition. Let $\gamma : [a, b] \to \mathbb{R}^n$ and $\tilde{\gamma} : [c, d] \to \mathbb{R}^n$ be parametrized curves. Then $\tilde{\gamma}$ is a *reparametrization* of γ if $\tilde{\gamma} = \gamma \circ \phi$, where ϕ is a nondecreasing continuous function from $[c, d]$ onto $[a, b]$ (hence $\phi(c) = a$ and $\phi(d) = b$). If ϕ is strictly increasing, $\tilde{\gamma}$ is a *strict reparametrization* of γ.

VII.1.2.2. A reparametrization of a parametrized curve has the same initial and final endpoints, and the same range. But two parametrized curves with the same range, even also with the same initial and final endpoints, need not be reparametrizations. A reparametrization of a parametrized curve describes "motion" along the range "curve" in the same sense, with the same retracing, but possibly over a different interval of time at a different and varying speed.

VII.1.2.3. Various special kinds of reparametrizations can be considered. The simplest is linear reparametrizations, where ϕ is a linear (affine) function from $[c, d]$ onto $[a, b]$, i.e.

$$\phi(t) = a + \frac{b - a}{d - c}(t - c) \ .$$

One can also consider $\mathcal{C}^r$ reparametrizations, where ϕ is $\mathcal{C}^r$, and strict $\mathcal{C}^r$ reparametrizations, where both ϕ and ϕ^{-1} are $\mathcal{C}^r$ (these are also called smooth reparametrizations). There are also piecewise linear and piecewise smooth reparametrizations.

VII.1.2.4. Reparametrization is not an equivalence relation: it is reflexive and transitive, but not symmetric since the function ϕ need not be one-to-one. Strict reparametrization is symmetric, thus an equivalence relation on parametrized curves. Parametrization does generate an important equivalence relation, which can be described thus: two parametrized curves are equivalent if they are both reparametrizations of the same parametrized curve (VII.1.2.7.).

Nonstationary Curves

VII.1.2.5. Definition. Let $\gamma : [a, b] \to \mathbb{R}^n$ be a parametrized curve, and $[c, d]$ a closed subinterval of $[a, b]$. Then γ is *stationary* on $[c, d]$ if $\gamma(t) = \gamma(c)$ for all $t \in [c, d]$. The curve γ is *nonstationary* if it is not stationary on any (nondegenerate) subinterval of $[a, b]$.

VII.1.2.6. A nonstationary parametrized curve need not be one-to-one; in fact, it can be very noninjective. For example, the space-filling curve of XIII.18.6.14. is nonstationary (the one in XIII.18.6.13. is not nonstationary). For a simpler example, a nonstationary parametrized curve can wrap around a circle in $\mathbb{R}^2$ several times.

VII.1.2.7. The maximal intervals on which a parametrized curve γ is stationary form a collection of disjoint closed subintervals of $[a, b]$. Every parametrized curve is a reparametrization of a nonstationary parametrized curve, which is unique up to strict reparametrization (Exercise ()), obtained by collapsing each maximal stationary subinterval to a point (there may be infinitely many such intervals!) Two parametrized curves are equivalent in the equivalence relation generated by reparametrization (VII.1.2.4.) if and only if their nonstationary versions are strict reparametrizations of each other.

VII.1.3. Paths

VII.1.3.1. Definition. Let $p, q \in \mathbb{R}^n$. A *path* from p to q is an equivalence class of parametrized curves from p to q (using the equivalence relation of VI.4.1.11. or VII.1.2.7.).

We will write $\gamma_1 \approx \gamma_2$ if γ_1 and γ_2 are equivalent, and write $\langle\gamma\rangle$ for the equivalence class of γ (the usual notation $[\gamma]$ for an equivalence class is normally reserved for the homotopy class of γ, which is a quite different thing ()).

VII.1.3.2. Paths can be concatenated: if $\langle\gamma_1\rangle$ is a path from p to q and $\langle\gamma_2\rangle$ is a path from q to r, then γ_1 can be concatenated with a suitable reparametrization $\tilde{\gamma}_2$ of γ_2 (e.g. a linear reparametrization), and the equivalence class of $\gamma_1 + \tilde{\gamma}_2$ is well defined and depends only on the classes of γ_1 and γ_2 (Exercise ()). Thus the concatenation of the paths $\langle\gamma_1\rangle + \langle\gamma_2\rangle$ is well defined as $\langle\gamma_1 + \tilde{\gamma}_2\rangle$.

Restriction of paths is a much dicier matter ().

VII.2. Smooth and Piecewise Smooth Curves in $\mathbb{R}^n$

Roughly speaking, a smooth curve in $\mathbb{R}^n$ is the image of a smooth function from $\mathbb{R}$, or a subinterval, to $\mathbb{R}^n$. But this is not exactly correct: we have to make some additional restrictions.

VII.2.1. Smooth Functions from $\mathbb{R}$ to $\mathbb{R}^n$

VII.2.1.1. Definition. Let γ be a function from an interval I to $\mathbb{R}^n$, with coordinate functions $\gamma_1, \ldots, \gamma_n$, and let $1 \leq r \leq \infty$. Then γ is $\mathcal{C}^r$ on I if each γ_k is $\mathcal{C}^r$ on I, i.e. has continuous derivatives of orders $\leq r$ on I.

This definition generalizes V.3.7.2., and is a special case of VIII.3.3.4., i.e. the definitions are consistent. We will generically, and imprecisely, use the term *smooth* to mean a function which is at least $\mathcal{C}^1$, and usually $\mathcal{C}^\infty$; we will try to be more specific when necessary or important.

The Derivative of a Vector-Valued Function

DEFINITION. Let $\gamma = (\gamma_1, \ldots, \gamma_n)$ be a function from an interval I in $\mathbb{R}$ to $\mathbb{R}^n$, and let a be an interior point of I. Then γ is *differentiable* at a if each γ_k is differentiable at a. If γ is differentiable at a, set

$$\gamma'(a) = (\gamma_1'(a), \ldots, \gamma_n'(a)) \ .$$

One can also define one-sided derivatives, especially at endpoints of I.

VII.2.1.2. Thus if γ is $\mathcal{C}^1$ on I, then it is differentiable at every (interior) point of I and one-sided differentiable at endpoints, and the function γ' is continuous on I (in the sense of ()), and conversely.

VII.2.1.3. This definition is (essentially) compatible with VIII.3.1.4., with slightly differing language: in the language there, γ is differentiable at a if and only if the limit

$$\lim_{h \to 0} \frac{1}{h}[\gamma(a+h) - \gamma(a)]$$

exists and equals $\gamma'(a)$, or equivalently

$$\lim_{h \to 0} \frac{\gamma(a+h) - \gamma(a) - h\gamma'(a)}{|h|} = \mathbf{0} \ .$$

Geometric Interpretation of the Derivative

VII.2.1.4. Suppose $\gamma : I \to \mathbb{R}^n$ is a function which is differentiable at an interior point a of I, and set $\boldsymbol{p} = \gamma(a)$. If $\boldsymbol{v} = \gamma'(a) \neq \mathbf{0}$, we want to interpret $\boldsymbol{v}$ as a "tangent vector" to the curve γ at $\boldsymbol{p}$. We have not yet defined what this means, so we could, and will, just define the tangent line to γ at $\boldsymbol{p}$ to be the line through $\boldsymbol{p}$ parallel to $\boldsymbol{v}$. But let us first see that this definition is reasonable.

VII.2.1.5. EXAMPLE. Consider first the case where f is a differentiable function from an open interval I to $\mathbb{R}$, and set

$$\gamma(t) = (t, f(t)) \in \mathbb{R}^2$$

for $t \in I$. Then γ is a differentiable function from I to $\mathbb{R}^2$ whose image is the graph of f, and for $t \in I$ we have $\gamma'(t) = (1, f'(t))$. If $a \in I$, $\boldsymbol{p} = (a, f(a)) = \gamma(a)$, and $\boldsymbol{v} = (1, f'(a)) = \gamma'(a)$, then the line through $\boldsymbol{p}$ parallel to $\boldsymbol{v}$ has slope $f'(a)$, and is exactly the tangent line to the graph of f at $\boldsymbol{p}$, as defined in V.3.1.10..

VII.2.1.6. More generally, if $n = 2$ and γ is a $\mathcal{C}^1$ function from I to $\mathbb{R}^2$ with $\gamma'(a) \neq \mathbf{0}$, the situation can be reduced to the one of VII.2.1.5. by reparametrization (Exercise VII.2.3.1.).

VII.2.1.7. Here is a better argument which is valid for any n and which does not require γ' to be continuous or even defined anywhere but a, i.e. in the general situation of VII.2.1.4.. The geometric picture is very similar to the way the tangent line to the graph of a function from $\mathbb{R}$ to $\mathbb{R}$ is defined in V.3.1.10., although we cannot define the slope of a line in $\mathbb{R}^n$ and must instead define the line itself. Fix the point $\boldsymbol{p} = \gamma(a)$, and for t near a let $\boldsymbol{q}$ be the variable point $\gamma(t) = \gamma(a+\Delta t)$. Write $\Delta\gamma = \boldsymbol{q} - \boldsymbol{p}$, and set $\boldsymbol{w} = \frac{1}{\Delta t}\Delta\gamma$ (Figure ()):
We have

$$\lim_{\Delta t \to 0} \frac{\Delta\gamma}{\Delta t} = \boldsymbol{v}$$

i.e. we can write $\boldsymbol{v} = \frac{d\gamma}{dt}(a)$.

VII.2.1.8. In the physical situation where n is 2 or 3, we may consider the motion of a particle in $\mathbb{R}^n$ whose position at time t is $\gamma(t)$. Then the vector $\boldsymbol{w} = \frac{\Delta\gamma}{\Delta t}$ can be interpreted as the "average velocity" of the particle between $t = a$ and $t = a + \Delta t$, and $\boldsymbol{v} = \gamma'(a)$ is the "instantaneous velocity" of the particle at time a. We can also interpret $\left\| \frac{\Delta\gamma}{\Delta t} \right\|$ as the "average speed" of the particle between $t = a$ and $t = a + \Delta t$ and $\|\boldsymbol{v}\|$ as the "instantaneous speed" at a.

VII.2.1.9. Here is another way to express this. Define

$$\beta(t) = \boldsymbol{p} + (t - a)\boldsymbol{v}$$

for any $t \in \mathbb{R}$. Then $\beta : \mathbb{R} \to \mathbb{R}^n$, $\beta(a) = \boldsymbol{p}$, and β is differentiable everywhere ($\mathcal{C}^\infty$) with $\beta'(t) = \boldsymbol{v}$ for all t (this is easily verified by considering the coordinate functions of β, which are linear or affine). In fact, β parametrizes motion along the line through $\boldsymbol{p}$ parallel to $\boldsymbol{v}$ at constant speed $\|\boldsymbol{v}\|$. If t is close to a, $\beta(t)$ closely approximates $\gamma(t)$; in fact,

$$\lim_{t \to a} \frac{\|\gamma(t) - \beta(t)\|}{|t - a|} = \lim_{h \to 0} \frac{\|\gamma(a + h) - \gamma(a) - h\gamma'(a)\|}{|h|} = 0$$

or, in other words, for any $\epsilon > 0$ there is a $\delta > 0$ such that

$$\|\gamma(t) - \beta(t)\| < \epsilon|t - a|$$

for $|t - a| < \delta$, i.e. γ and β are virtually indistinguishable for t near a. Thus we can not only interpret the line L parametrized by β as the "tangent line" to γ at p, but we can also interpret $\boldsymbol{v}$ as pointing in the "direction of motion" of γ at a and $\|\boldsymbol{v}\|$ as the "instantaneous speed" of γ at a.

VII.2.2. Smooth Curves in $\mathbb{R}^n$

When $\gamma'(a) = \mathbf{0}$ in VII.2.1.4., all bets are off. Even if γ is $\mathcal{C}^\infty$ on I, the image need not be a smooth curve with a tangent line at a: it can be very pathological, as the following examples show.

VII.2.2.1. Examples. (i) Define $\gamma : \mathbb{R} \to \mathbb{R}^2$ by $\gamma(t) = (t^3, t^2)$. Then γ is $\mathcal{C}^\infty$, and $\gamma'(t) = (3t^2, 2t)$. So $\gamma'(0) = \mathbf{0}$. The image of γ is the graph of $f(t) = t^{2/3}$, which has a cusp at $\gamma(0) = \mathbf{0}$ and is not smooth there.

(ii) The situation can be far worse. Define $\gamma : \mathbb{R} \to \mathbb{R}^2$ by

$$\gamma(t) = \begin{cases} (t^2 \cos \frac{1}{t}, t^2 \sin \frac{1}{t}) & \text{if} \quad t \neq 0 \\ (0,0) & \text{if} \quad t = 0 \end{cases} .$$

Then γ is differentiable at 0 and $\gamma'(0) = \mathbf{0}$. However, as $t \to 0$ the curve spirals around the origin infinitely many times, so cannot have anything like a tangent line at $\mathbf{0}$.

This example can be modified to make it $\mathcal{C}^\infty$ by replacing t^2 by e^{-1/t^2} in the definition.

(iii) Of course, the curve *can* be smooth at a point where $\gamma'(a) = \mathbf{0}$: consider $\gamma(t) = (t^3, t^3)$. But there is no straightforward way to tell from the derivative when this is the case.

We thus make the following definition:

VII.2.2.2. DEFINITION. Let $1 \leq r \leq \infty$. A $\mathcal{C}^r$ *parametrized curve* in $\mathbb{R}^n$ is a $\mathcal{C}^r$ function γ from an interval I to $\mathbb{R}^n$ (one-sided derivatives at endpoints of I) such that $\gamma'(t) \neq 0$ for all $t \in I$.

We will imprecisely use the term *smooth parametrized curve* to denote a $\mathcal{C}^r$ parametrized curve for sufficiently large r depending on context (usually $r = \infty$).

Piecewise-Smooth Parametrized Curves

VII.2.2.3. We also define a piecewise $\mathcal{C}^r$ parametrized curve in the obvious way: a parametrized curve $\gamma : I \to \mathbb{R}^n$ is *piecewise* $\mathcal{C}^r$ *on* I if γ is continuous on I and there is a partition of I into a finite number $I_1, \ldots, I_m$ of nonoverlapping subintervals such that γ is a $\mathcal{C}^r$ parametrized curve on each I_k. Thus if a is a common endpoint for I_k and I_{k+1}, γ must have left and right derivatives at a of orders $\leq r$, but they need not coincide, although we need that $\lim_{t \to a-} \gamma(t) = \lim_{t \to a+} \gamma(t) = \gamma(a)$ and that the first derivatives are everywhere nonzero. The image of γ is thus a smooth curve on each I_k but may have an "angle" or "cusp" at a common endpoint.

VII.2.2.4. It turns out that any piecewise $\mathcal{C}^r$ parametrized curve has a reparametrization by a function which is $\mathcal{C}^r$ everywhere, but with zero derivative at the common endpoints (Exercise VII.2.3.2.). However, the image of a $\mathcal{C}^r$ function does not necessarily have a reparametrization making it a piecewise $\mathcal{C}^r$ parametrized curve (Example VII.2.2.1.(ii)).

VII.2.2.5. It is quite useful to have a notion of piecewise-smooth parametrized curves, for many reasons, including:

(i) It is frequently important to consider curves, e.g. polygonal curves, which are piecewise-smooth but not smooth.

(ii) A concatenation of smooth parametrized curves is not smooth in general, but is piecewise-smooth. In fact, a concatenation of piecewise-smooth parametrized curves is piecewise-smooth.

VII.2.3. Exercises

VII.2.3.1. Prove the statement in VII.2.1.6.. [Without loss of generality, suppose $\gamma_1'(a) \neq 0$. If $b = \gamma_1(a)$, then γ_1 has a smooth inverse function β in an interval around b. Let $\tilde{\gamma} = \gamma \circ \beta$.]

VII.2.3.2. Prove the statement in VII.2.2.4.. [For each k, let ϕ_k be a strictly increasing $\mathcal{C}^\infty$ function from I_k onto I_k with all derivatives 0 at the endpoints, with ϕ_k' nonzero on the interior of I_k, cf. V.17.2.5..]

VII.3. Rectifiable Curves and Arc Length

Throughout this section, ρ will denote the Euclidean distance function on $\mathbb{R}^d$. (We use d instead of n so that n is available as a counting parameter for partitions.)

VII.3.1. Arc Length

VII.3.1.1. DEFINITION. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a parametrized curve, and $\mathcal{P} = \{a = t_0, t_1, \ldots, t_n = b\}$ a partition of $[a,b]$. The *polygonal approximate length of γ from $\mathcal{P}$* is

$$\ell(\gamma, \mathcal{P}) = \sum_{k=1}^{n} \rho(\gamma(t_{k-1}), \gamma(t_k)) \ .$$

The *arc length* of γ is

$$\ell(\gamma) = \sup_{\mathcal{P}} \ell(\gamma, \mathcal{P}) \ .$$

VII.3.1.2. If $\mathcal{Q}$ is a partition of $[a,b]$ which refines $\mathcal{P}$, we have $\ell(\gamma, \mathcal{P}) \le \ell(\gamma, \mathcal{Q})$ by the triangle inequality. Since any two partitions have a common refinement, the polygonal approximations approach the arc length as the partitions become arbitrarily fine. The term "polygonal approximation" should not be taken too literally: there is usually not an actual "polygon" in $\mathbb{R}^d$ approximating the curve. The term is suggestive that $\rho(p,q)$ should be the "straight-line" distance between p and q.

The arc length of a curve can easily be infinite (e.g. VI.4.1.20.).

VII.3.1.3. DEFINITION. The curve γ is *rectifiable* if $\ell(\gamma)$ is finite.

VII.3.1.4. EXAMPLES. (i) A parametrized curve in $\mathbb{R}^1$ is a continuous function $\gamma : [a,b] \to \mathbb{R}$. If $\mathcal{P}$ is a partition of $[a,b]$, then $\ell(\gamma, \mathcal{P})$ is the total variation $V(\gamma, \mathcal{P})$ of γ over $\mathcal{P}$ (XIX.4.1.3.), and $\ell(\gamma)$ is the total variation $V_{[a,b]}(\gamma)$ of γ over $[a,b]$ (XIX.4.1.6.). The curve γ is rectifiable if and only if γ has bounded variation. If γ is a smooth function, then γ is rectifiable and (VII.3.4.6.)

$$\ell(\gamma) = \int_a^b \|\gamma'(t)\| \, dt \ .$$

We now develop basic properties of arc length. First note that arc length is independent of parametrization:

VII.3.1.5. PROPOSITION. Let $\tilde{\gamma} : [c,d] \to \mathbb{R}^d$ be a reparametrization of $\gamma : [a,b] \to \mathbb{R}^d$, i.e. $\tilde{\gamma} = \gamma \circ \phi$, $\phi : [c,d] \to [a,b]$. Then $\ell(\tilde{\gamma}) = \ell(\gamma)$.

PROOF: Let $\mathcal{P} = \{c = t_0, \ldots, t_n = d\}$ be a partition of $[c,d]$. Then $\{a = \phi(t_0), \ldots, \phi(t_n) = b\}$ is not quite a partition of $[a,b]$ since there may be repetitions if ϕ is not injective. But repeated points contribute nothing to the sum $\ell(\tilde{\gamma}, \mathcal{P})$, and if $\mathcal{Q}$ is the partition of $[a,b]$ obtained by deleting repetitions, we have $\ell(\gamma, \mathcal{Q}) = \ell(\tilde{\gamma}, \mathcal{P})$. Every partition of $[a,b]$ arises in this manner. 👍

VII.3.1.6. Thus arc length of paths is well defined. Actually, the use of a parametrization can be regarded as merely a convenient way of "partitioning" the curve into small segments: if $\gamma : [a,b] \to \mathbb{R}^d$ is a parametrized curve, with $\boldsymbol{p} = \gamma(a)$ and $\boldsymbol{q} = \gamma(b)$, so the image of γ is a "curve" from $\boldsymbol{p}$ to $\boldsymbol{q}$, a partition $\mathcal{P} = \{a = t_0, \ldots, t_n = b\}$ of $[a,b]$ gives a "partition" $\{\boldsymbol{p} = \boldsymbol{x}_0, \ldots, \boldsymbol{x}_n = \boldsymbol{q}\}$ of the curve, where $\boldsymbol{x}_k = \gamma(t_k)$. The arc length of γ is

$$\ell(\gamma) = \sup \sum_{k=1}^{n} \rho(\boldsymbol{x}_{k-1}, \boldsymbol{x}_k)$$

where the supremum is over all such "partitions."

Although the reversal $-\gamma$ of a parametrized curve is not a reparametrization of γ, we nonetheless have:

VII.3.1.7. PROPOSITION. If $\gamma : [a,b] \to \mathbb{R}^d$ is a parametrized curve in X, then $\ell(-\gamma) = \ell(\gamma)$.

PROOF: If $\mathcal{P}$ is a partition of $[a,b]$, let $\tilde{\mathcal{P}}$ be the partition consisting of $\{a + b - t : t \in \mathcal{P}\}$. Then $\tilde{\mathcal{P}}$ is also a partition of $[a,b]$, and every partition of $[a,b]$ arises in this manner. We have $\ell(-\gamma, \tilde{\mathcal{P}}) = \ell(\gamma, \mathcal{P})$. 👍

VII.3.1.8. If $\gamma : [a,b] \to X$ is a parametrized curve from p to q and $\mathcal{P} = \{a,b\}$ is the trivial partition, we have

$$\rho(p,q) = \ell(\gamma, \mathcal{P}) \leq \ell(\gamma) \ .$$

More generally, we have:

VII.3.1.9. PROPOSITION. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a parametrized curve in $\mathbb{R}^d$, and $a \leq c \leq d \leq b$, then

$$\rho(\gamma(c), \gamma(d)) \leq \ell(\gamma|_{[c,d]}) \leq \ell(\gamma) \ .$$

PROOF: The first inequality has already been noted. For the second, any partition $\mathcal{P}$ of $[c,d]$ can be expanded to a partition $\mathcal{Q}$ of $[a,b]$ by adding a and b if necessary. We have

$$\ell(\gamma|_{[c,d]}, \mathcal{P}) \leq \ell(\gamma, \mathcal{Q}) \leq \ell(\gamma)$$

and taking the supremum over all $\mathcal{P}$ gives the inequality. 👍

If γ is a constant path, obviously $\ell(\gamma) = 0$. Conversely:

VII.3.1.10. COROLLARY. Let γ be a parametrized curve in $\mathbb{R}^d$. If $\ell(\gamma) = 0$, then γ is constant.

PROOF: If γ is not constant, there are $c, d \in [a, b]$, $c < d$, with $\gamma(c) \neq \gamma(d)$. Then $0 < \rho(\gamma(c), \gamma(d)) \leq \ell(\gamma)$. 👍

Here is a more precise version of VII.3.1.9.:

VII.3.1.11. PROPOSITION. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a parametrized curve, and $c \in [a, b]$. Then

$$\ell(\gamma) = \ell(\gamma|_{[a,c]}) + \ell(\gamma|_{[c,b]}) \ .$$

PROOF: If $\mathcal{P}$ is a partition of $[a, c]$ and $\mathcal{Q}$ a partition of $[c, b]$, then $\mathcal{P} \cup \mathcal{Q}$ is a partition of $[a, b]$, and

$$\ell(\gamma, \mathcal{P} \cup \mathcal{Q}) = \ell(\gamma|_{[a,c]}, \mathcal{P}) + \ell(\gamma|_{[c,b]}, \mathcal{Q}) \ .$$

Every partition of $[a, b]$ containing c (i.e. every sufficiently fine partition) arises in this manner. 👍

Iterating, we get:

VII.3.1.12. COROLLARY. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a parametrized curve, and $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$ a partition of $[a, b]$. Then

$$\ell(\gamma) = \sum_{k=1}^{n} \ell(\gamma|_{[t_{k-1} t_k]}) \ .$$

Rephrasing VII.3.1.11., we get:

VII.3.1.13. PROPOSITION. Let $\gamma_1 : [a, b] \to \mathbb{R}^d$ and $\gamma_2 : [b, c] \to \mathbb{R}^d$ with $\gamma_1(b) = \gamma_2(b)$, so that the concatenation $\gamma_1 + \gamma_2$ is defined. Then

$$\ell(\gamma_1 + \gamma_2) = \ell(\gamma_1) + \ell(\gamma_2) \ .$$

This can also be iterated to apply to arbitrary finite concatenations.

A crucial property of arc length is lower semicontinuity:

VII.3.1.14. PROPOSITION. For each n, let $\gamma_n : [a,b] \to \mathbb{R}^d$ be a parametrized curve from p_n to q_n. Suppose (γ_n) converges pointwise to a parametrized curve $\gamma : [a,b] \to X$ from p to q, i.e. $\gamma_n(t) \to \gamma(t)$ for each $t \in [a,b]$ (so $p_n \to p$, $q_n \to q$). Then

$$\ell(\gamma) \leq \liminf_{n\to\infty} \ell(\gamma_n) \ .$$

We can have strict inequality, even if $\gamma_n \to \gamma$ uniformly on $[a,b]$ (I.6.1.16.).

PROOF: First note that if $\mathcal{P} = \{a = t_0, \ldots, t_m = b\}$ is any partition of $[a,b]$, we have

$$\ell(\gamma_n, \mathcal{P}) = \sum_{k=1}^{m} \rho(\gamma_n(t_{k-1}), \gamma_n(t_k)) \to \sum_{k=1}^{m} \rho(\gamma(t_{k-1}), \gamma(t_k)) = \ell(\gamma, \mathcal{P})$$

by continuity of ρ.
Suppose $\ell(\gamma) = +\infty$. Fix $M > 0$. Then there is a partition $\mathcal{P}$ of $[a,b]$ with $\ell(\gamma, \mathcal{P}) > M$. Then

$$M < \ell(\gamma, \mathcal{P}) = \lim_{n\to\infty} \ell(\gamma_n, \mathcal{P}) \leq \liminf_{n\to\infty} \ell(\gamma_n) \ .$$

This holds for all $M > 0$, so $\liminf_{n\to\infty} \ell(\gamma_n) = +\infty$ and the inequality holds.
Now suppose γ is rectifiable. Let $\epsilon > 0$. Then there is a partition $\mathcal{P}$ of $[a,b]$ with $\ell(\gamma, \mathcal{P}) > \ell(\gamma) - \epsilon$. Then

$$\ell(\gamma) - \epsilon < \ell(\gamma, \mathcal{P}) = \lim_{n\to\infty} \ell(\gamma_n, \mathcal{P}) \leq \liminf_{n\to\infty} \ell(\gamma_n) \ .$$

This holds for all $\epsilon > 0$, so $\ell(\gamma) \leq \liminf_{n\to\infty} \ell(\gamma_n)$. 👍

VII.3.1.15. Note for future reference that the definitions of $\ell(\gamma, \mathcal{P})$ and $\ell(\gamma)$ make perfect sense for a general function $\gamma : [a,b] \to \mathbb{R}^d$, not necessarily continuous, so "arc length" is defined even for discontinuous functions (although it is typically infinite for very discontinuous functions). None of the proofs of VII.3.1.5.–VII.3.1.14. use that γ is continuous, so these properties hold also for discontinuous γ. (But VII.3.2.2. requires continuity of γ.)

VII.3.2. Rectifiable Curves

Continuity of Arc Length

VII.3.2.1. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a rectifiable parametrized curve. Define a function $\phi : [a,b] \to [0, \ell(\gamma)]$ by setting

$$\phi(t) = \ell(\gamma|_{[a,t]}) \ .$$

Then ϕ is nondecreasing, $\phi(a) = 0$, $\phi(b) = \ell(\gamma)$. If γ is nonstationary, ϕ is strictly increasing.

VII.3.2.2. PROPOSITION. The function ϕ is continuous.

PROOF: The following lemma will provide left continuity on $[a,b]$. The simplest way to get right continuity is to apply the lemma also to $-\gamma$ and use that $\ell(\gamma|_{[a,t_0]}) = \ell(\gamma) - \ell(\gamma|_{[t_0,b]})$. 👍

VII.3.2.3. Lemma. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a parametrized curve, not necessarily rectifiable. If (t_n) is an increasing sequence in $[a,b]$ converging to t_0, then

$$\ell(\gamma|_{[a,t_0]}) = \lim_{n\to\infty} \ell(\gamma|_{[a,t_n]}) = \sup_n \ell(\gamma|_{[a,t_n]}) \ .$$

Proof: Define a sequence (γ_n) of parametrized curves on $[a,t_0]$ by setting $\gamma_n(t) = \gamma(t)$ if $a \le t \le t_n$ and $\gamma_n(t) = \gamma(t_n)$ if $t_n \le t \le t_0$ (i.e. γ_n agrees with γ on $[a,t_n]$ and is stationary on $[t_n,t_0]$). We have

$$\ell(\gamma_n) = \ell(\gamma|_{[a,t_n]}) \le \ell(\gamma|_{[a,t_0]})$$

for any n by monotonicity (VII.3.1.9.). On the other hand, since γ is continuous at t_0, $\gamma_n \to \gamma$ pointwise on $[a,t_0]$. So applying VII.3.1.14. we get

$$\sup_n \ell(\gamma|_{[a,t_n]}) \le \ell(\gamma|_{[a,t_0]}) \le \liminf_{n\to\infty} \ell(\gamma_n) = \liminf_{n\to\infty} \ell(\gamma|_{[a,t_n]})$$

and the result follows. 👍

VII.3.2.4. Corollary. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a rectifiable curve. Then for any $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $a \le c < d \le b$ and $d - c < \delta$, then $\ell(\gamma|_{[c,d]}) < \epsilon$.

Proof: The function ϕ is uniformly continuous on $[a,b]$ (). 👍

Estimation of Arc Length

The next result is strongly reminiscent of the equivalence of Definitions 1A and 1B of the Riemann integral (IX.2.3.1.), and proved in a similar way.

VII.3.2.5. Theorem. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a rectifiable curve. Then for any $\epsilon > 0$ there is a $\delta > 0$ such that $\ell(\gamma,\mathcal{P}) > \ell(\gamma) - \epsilon$ for any partition $\mathcal{P}$ of $[a,b]$ with $\|\mathcal{P}\| < \delta$.

Proof: Let $\mathcal{Q} = \{a = t_0, \dots, t_n = b\}$ be a partition of $[a,b]$ with $\ell(\gamma,\mathcal{Q}) > \ell(\gamma) - \frac{\epsilon}{2}$. Since γ is uniformly continuous (), there is a $\delta > 0$ such that if $a \le c < d \le b$ and $d - c < \delta$, $\rho(\gamma(c),\gamma(d)) < \frac{\epsilon}{4n}$. We may assume $3\delta < \|\mathcal{Q}\|$.
Let $\mathcal{P} = \{a = s_0, \dots, s_m = b\}$ be a partition of $[a,b]$ with $\|\mathcal{P}\| < \delta$. For $1 \le k \le n-1$ let j_k be the largest j such that $s_j \le t_k$. We then have

$$\ell(\gamma,\mathcal{P}) = \sum_{j=1}^{m} \rho(\gamma(s_{j-1}),\gamma(s_j))$$

$$= \sum_{j=1}^{j_1} \rho(\gamma(s_{j-1}),\gamma(s_j)) + \sum_{k=1}^{n-2} \sum_{j=j_k+2}^{j_{k+1}} \rho(\gamma(s_{j-1}),\gamma(s_j)) + \sum_{j=j_{n-1}+2}^{m} \rho(\gamma(s_{j-1}),\gamma(s_j)) + \sum_{k=1}^{m-1} \rho(\gamma(s_{j_k}),\gamma(s_{j_k+1})) \ .$$

By the triangle inequality, we have

$$\sum_{j=1}^{j_1} \rho(\gamma(s_{j-1}),\gamma(s_j)) \geq \rho(\gamma(a),\gamma(t_1)) - \rho(\gamma(s_{j_1}),\gamma(t_1)) > \rho(\gamma(a),\gamma(t_1)) - \frac{\epsilon}{4n} \ .$$

Similarly, for $1 \leq k \leq n-2$, we have

$$\sum_{j=j_k+2}^{j_{k+1}} \rho(\gamma(s_{j-1}),\gamma(s_j)) \geq \rho(\gamma(t_k),t_{k+1})) - \rho(\gamma(t_k),\gamma(s_{j_k+1}) - \rho(\gamma(s_{j_{k+1}}),\gamma(t_{k+1})) > \rho(\gamma(t_k),t_{k+1})) - \frac{\epsilon}{2n}$$

$$\sum_{j=j_{n-1}+2}^{m} \rho(\gamma(s_{j-1}),\gamma(s_j)) > \rho(\gamma(t_{n-1}),\gamma(b)) - \frac{\epsilon}{4n} \ .$$

Thus we have

$$\ell(\gamma,\mathcal{P}) = \sum_{j=1}^{m} \rho(\gamma(s_{j-1}),\gamma(s_j))$$

$$\geq \sum_{j=1}^{j_1} \rho(\gamma(s_{j-1}),\gamma(s_j)) + \sum_{k=1}^{n-2} \sum_{j=j_k+2}^{j_{k+1}} \rho(\gamma(s_{j-1}),\gamma(s_j)) + \sum_{j=j_{n-1}+2}^{m} \rho(\gamma(s_{j-1}),\gamma(s_j))$$

$$> \sum_{k=1}^{n} \rho(\gamma(t_{k-1}),\gamma(t_k)) - (n-2)\frac{\epsilon}{2n} - 2\frac{\epsilon}{4n} = \ell(\gamma,\mathcal{Q}) - \frac{\epsilon}{2} > \ell(\gamma) - \epsilon \ .$$



Parametrization by Arc Length

VII.3.2.6. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a nonstationary rectifiable parametrized curve, and let ϕ be the above function. Then ϕ is strictly increasing and continuous, so has a continuous strictly increasing inverse function $\phi^{-1} : [0,\ell(\gamma)] \to [a,b]$. Set $\tilde{\gamma} = \gamma \circ \phi^{-1} : [0,\ell(\gamma)] \to X$. Then $\tilde{\gamma}$ is a strict reparametrization of γ with the property that $\ell(\tilde{\gamma}|_{[0,t]}) = t$ for $0 \leq t \leq \ell(\gamma)$, i.e. $\tilde{\gamma}$ "traverses the path $\langle\gamma\rangle$ at uniform unit speed." The parametrized curve $\tilde{\gamma}$ is called the *reparametrization of γ by arc length.* If γ is replaced by any strict reparametrization, the $\tilde{\gamma}$ is unchanged; thus it depends only on $\langle\gamma\rangle$, i.e. any rectifiable path has a unique canonical parametrization by arc length.

Differentiability of Rectifiable Curves

A rectifiable curve has "bounded variation" and hence should be differentiable almost everywhere, and in fact is:

VII.3.2.7. THEOREM. Let $\gamma : [a,b] \to \mathbb{R}^n$ be a rectifiable curve. Then γ is differentiable a.e.

PROOF: Let $\gamma_1,\ldots,\gamma_n$ be the coordinate functions of γ. Then each γ_k has bounded variation $V_{[a,b]}(\gamma_k) \leq \ell(\gamma)$ (VII.3.5.1.(a)). Thus γ_k is differentiable on $[a,b]$ except on a set A_k of measure 0 (XIX.4.2.4.). So γ is differentiable on $[a,b]$ except on $A_1 \cup \cdots \cup A_n$, a set of measure 0. 

A rectifiable curve can fail to be differentiable at a dense set of points, however ().

VII.3.2.8. If γ is rectifiable, one might expect that its arc length can be computed by integration of the length of the derivative. However, this fails in general since γ is not necessarily absolutely continuous (). For an extreme example, if γ is parametrized on $[0,1]$ and g is the Cantor function (), then $\gamma \circ g$ is a reparametrization of γ, and $(\gamma \circ g)' = \mathbf{0}$ a.e. But if γ is parametrized by an absolutely continuous function, it is true. This result is surprisingly hard to

(i) prove

(ii) find in books

perhaps because it falls through the crack between real analysis and differential geometry. Our proof is adapted from [Pel77], where a more general result is proved. If γ is smooth, there is an (arguably simpler) alternate proof, which we give in a more general setting in VII.3.4.6..

VII.3.2.9. THEOREM. Let $\gamma : [a,b] \to \mathbb{R}^n$ be a (necessarily rectifiable) curve, with absolutely continuous coordinate functions. Then γ is differentiable a.e., and

$$\ell(\gamma) = \int_a^b \|\gamma'(t)\| \, d\lambda(t)$$

(where the integral is a Lebesgue integral). The length function $\phi(x) = \ell(\gamma|_{[a,x]})$ is absolutely continuous, hence differentiable a.e., and $\phi'(t) = \|\gamma'(t)\|$ a.e.

The proof will mostly consist of three lemmas. The first is of independent interest, that a rectifiable curve can be "unwound" to a curve with nondecreasing coordinate functions without changing the arc length or length of the derivative (a.e.):

VII.3.2.10. LEMMA. Let $\gamma : [a,b] \to \mathbb{R}^n$ be a rectifiable curve with coordinate functions $\gamma_1, \ldots, \gamma_n$ of bounded variation. For $1 \le k \le n$ let $v_k = T\gamma_k$ be the total variation function (XIX.4.2.1.) of γ_k on $[a,b]$, i.e. $v_k(t) = V_{[a,t]}(\gamma_1)$ (the v_k are continuous by XIX.4.2.6.). Let $\bar{\gamma} : [a,b] \to \mathbb{R}^n$ have coordinate functions $v_1, \ldots, v_n$. Then $\ell(\bar{\gamma}) = \ell(\gamma)$.

PROOF: We have, for any k and any $s, t \in [a,b]$,

$$|\gamma_k(t) - \gamma_k(s)| \le |v_k(t) - v_k(s)|$$

and it follows that for any partition $\mathcal{P}$ of $[a,b]$, $\ell(\gamma, \mathcal{P}) \le \ell(\bar{\gamma}, \mathcal{P})$. Taking the supremum over all $\mathcal{P}$, we obtain $\ell(\gamma) \le \ell(\bar{\gamma})$.

For the opposite inequality, suppose $a \le c < d \le b$ and $\mathcal{P} = \{c = t_0, \ldots, t_m = d\}$ is a partition of $[c,d]$. Using the triangle inequality in $\mathbb{R}^n$ we obtain

$$\ell(\gamma|_{[c,d]}, \mathcal{P}) = \sum_{j=1}^{m} \sqrt{\sum_{k=1}^{n} (\gamma_k(t_j) - \gamma_k(t_{j-1}))^2}$$

$$\ge \sqrt{\sum_{k=1}^{n} \left[\sum_{j=1}^{m} |\gamma_k(t_j) - \gamma_k(t_{j-1})| \right]^2} = \sqrt{\sum_{k=1}^{n} [V(\gamma_k, \mathcal{P})]^2} \, .$$

Taking the supremum over all $\mathcal{P}$ we obtain

$$\ell(\gamma|_{[c,d]}) \geq \sqrt{\sum_{k=1}^{n} [V_{[c,d]}(\gamma_k)]^2} = \sqrt{\sum_{k=1}^{n} (v_k(d) - v_k(c))^2} \ .$$

If $\mathcal{Q} = \{a = s_0, \ldots, s_r = b\}$ is any partition of $[a, b]$, applying this inequality over each subinterval $[s_{j-1}, s_j]$, we obtain

$$\ell(\gamma) = \sum_{j=1}^{r} \ell(\gamma|_{[s_{j-1},s_j]}) \geq \sum_{j=1}^{r} \sqrt{\sum_{k=1}^{n} (v_k(s_j) - v_k(s_{j-1}))^2} = \ell(\bar{\gamma}, \mathcal{Q})$$

and it follows by taking the supremum over all $\mathcal{Q}$ that $\ell(\gamma) \geq \ell(\bar{\gamma})$. 👍

VII.3.2.11. If $\gamma_1, \ldots, \gamma_n$ are absolutely continuous, so are $v_1, \ldots, v_n$, and $v_k' = |\gamma_k'|$ a.e. for each k, so $\|\bar{\gamma}'\| = \|\gamma'\|$ a.e. Thus to prove VII.3.2.9. we may work with $\bar{\gamma}$, i.e. we may assume that $\gamma_k' \geq 0$ a.e. for each k.

The other two lemmas are very general:

VII.3.2.12. Lemma. Let $(X, \mathcal{A}, \mu)$ be a measure space and $f_1, \ldots, f_n$ nonnegative measurable functions. Then, for any $A \in \mathcal{A}$,

$$\sqrt{\sum_{k=1}^{n} \left[\int_A f_k \, d\mu\right]^2} \leq \int_A \sqrt{\sum_{k=1}^{n} f_k^2} \ d\mu \ .$$

Proof: This is a good example of bootstrapping. First suppose the f_k are simple functions, and are all constant on sets $E_1, \ldots, E_m$, with f_k taking value $\alpha_{k,j}$ on E_j. Then by the triangle inequality on $\mathbb{R}^n$,

$$\sqrt{\sum_{k=1}^{n} \left[\int_A f_k \, d\mu\right]^2} = \sqrt{\sum_{k=1}^{n} \left[\sum_{j=1}^{m} \int_{A\cap E_j} f_k \, d\mu\right]^2} = \sqrt{\sum_{k=1}^{n} \left[\sum_{j=1}^{m} \alpha_{k,j} \mu(A \cap E_j)\right]^2}$$

$$\leq \sum_{j=1}^{m} \sqrt{\sum_{k=1}^{n} [\alpha_{k,j} \mu(A \cap E_j)]^2} = \sum_{j=1}^{m} \left[\sqrt{\sum_{k=1}^{n} \alpha_{k,j}^2} \ \mu(A \cap E_j)\right] = \sum_{j=1}^{m} \left[\int_{A\cap E_j} \sqrt{\sum_{k=1}^{n} f_k^2} \ d\mu\right] = \int_A \sqrt{\sum_{k=1}^{n} f_k^2} \ d\mu \ .$$

For general f_k, take increasing sequences of nonnegative simple functions and apply the Monotone Convergence Theorem. 👍

VII.3.2.13. LEMMA. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f, f_1, \ldots, f_n$ nonnegative measurable functions on X. If

$$\left[\int_A f\,d\mu\right]^2 \geq \sum_{k=1}^n \left[\int_A f_k\,d\mu\right]^2$$

for all $A \in A$, then $f \geq \sqrt{\sum_{k=1}^n f_k^2}$ a.e.

PROOF: Suppose $f_1, \ldots, f_n$ are simple and all constant on $E_1, \ldots, E_m$, and f_k takes the constant value $\alpha_{k,j}$ on E_j. If $A \subseteq E_j$, then

$$\int_A f\,d\mu \geq \sqrt{\sum_{k=1}^n \left[\int_A f_k\,d\mu\right]^2} = \sqrt{\sum_{k=1}^n [\alpha_{k,j}\mu(A)]^2} = \mu(A)\sqrt{\sum_{k=1}^n \alpha_{k,j}^2}\ .$$

This is true for all $A \subseteq E_j$, which implies that

$$f \geq \sqrt{\sum_{k=1}^n \alpha_{k,j}^2} = \sqrt{\sum_{k=1}^n f_k^2}$$

a.e. on E_j. Since $\cup_{j=1}^m E_j = X$, we have $f \geq \sqrt{\sum_{k=1}^n f_k^2}$ a.e. on X.

For general f_k, take increasing sequences of nonnegative simple functions converging to the f_k. 👍

VII.3.2.14. We now give the proof of Theorem VII.3.2.9..

PROOF: We may assume γ_k' is nonnegative for all k (VII.3.2.11.). If $\mathcal{P} = \{a = t_0, \ldots, t_m = b\}$ is a partition of $[a, b]$, we have

$$\ell(\gamma, \mathcal{P}) = \sum_{j=1}^m \|\gamma(t_j) - \gamma(t_{j-1})\| = \sum_{j=1}^m \sqrt{\sum_{k=1}^n \left[\int_{t_{j-1}}^{t_j} \gamma_k'\,d\lambda\right]^2}$$

$$\leq \sum_{j=1}^m \left[\int_{t_{j-1}}^{t_j} \sqrt{\sum_{k=1}^n (\gamma_k')^2}\ d\lambda\right] = \sum_{j=1}^m \left[\int_{t_{j-1}}^{t_j} \|\gamma'\|\,d\lambda\right] = \int_a^b \|\gamma'\|\,d\lambda$$

by Lemma VII.3.2.12.. Taking the supremum over all $\mathcal{P}$, we obtain

$$\ell(\gamma) \leq \int_a^b \|\gamma'\|\,d\lambda\ .$$

For the reverse inequality, note that, if $\phi(t) = \ell(\gamma|_{[a,t]})$ for $a \leq t \leq b$, then ϕ is absolutely continuous (VII.3.5.1.(b)). If $a \leq c < d \leq b$, we have

$$\left[\int_c^d \phi'\,d\lambda\right]^2 = [\phi(d) - \phi(c)]^2 = [\ell(\gamma|_{[c,d]})]^2$$

$$\geq \|\gamma(d) - \gamma(c)\|^2 = \sum_{k=1}^{n} \left[\int_c^d \gamma_k' \, d\lambda \right]^2$$

and thus (), for every measurable A, we have

$$\left[\int_A \phi' \, d\lambda \right]^2 \geq \sum_{k=1}^{n} \left[\int_A \gamma_k' \, d\lambda \right]^2 .$$

So by Lemma VII.3.2.13. we have

$$\phi' \geq \sqrt{\sum_{k=1}^{n} [\gamma_k']^2} = \|\gamma'\| \text{ a.e.}$$

and hence

$$\ell(\gamma) = \int_a^b \phi' \, d\lambda \geq \int_a^b \|\gamma'\| \, d\lambda .$$

We have, for each $x \in [a, b]$,

$$\phi(x) = \int_a^x \phi' \, d\lambda = \int_a^x \|\gamma'\| \, d\lambda$$

so if $f = \phi' - \|\gamma'\|$, then f is an integrable function on $[a, b]$ and $\int_a^x f \, d\lambda = 0$ for all $x \in [a, b]$, so $f = 0$ a.e. (XVIII.1.8.14.). 👍

The following special case of VII.3.2.9. is especially noteworthy. The result is probably not surprising, but it is definitely nontrivial (cf. VII.3.2.16.). See () for a direct proof valid in general metric spaces.

VII.3.2.15. Corollary. Let $\gamma : [0, b] \to \mathbb{R}^n$ be a rectifiable curve, parametrized by arc length. Then γ is absolutely continuous and hence is differentiable a.e., and $\|\gamma'(t)\| = 1$ for almost all t.

Proof: Since γ is parametrized by arc length, it is a Lipschitz function with constant 1, so the coordinate functions are absolutely continuous and VII.3.2.9. applies; and ϕ' is the constant function 1. 👍

VII.3.2.16. Example. One might naïvely think that if γ is a rectifiable curve, parametrized by arc length, then $\|\gamma'(t)\| = 1$ whenever it is defined. But this is false. Here is an example, adapted from ().

Let c be the curve in $\mathbb{R}^2$ consisting of semicircular arcs between $\frac{1}{n}$ and $\frac{1}{n+1}$ for each $n \in \mathbb{N}$, alternating above and below the x-axis to make the curve differentiable on $(0, 1]$, reflected on the interval $[-1, 0)$, along with the origin. See Figure (). Then c is the graph of a function on $[-1, 1]$. Parametrize c by arc length γ, starting at $(-1, 0)$. Then $\ell(\gamma) = \pi$. We have that $\gamma\left(\frac{\pi}{2}\right) = (0, 0)$. Note that γ is differentiable everywhere except possibly at $\frac{\pi}{2}$. But it is differentiable there also, as the following calculation shows. Write γ_1 and γ_2 for the coordinate functions of γ.

The height of the semicircle between $\frac{1}{n}$ and $\frac{1}{n+1}$ is $O\left(\frac{1}{n^2}\right)$, and it follows easily that $\gamma_2'\left(\frac{\pi}{2}\right) = 0$. If $\frac{\pi}{2} + \frac{\pi}{2(n+1)} \leq t \leq \frac{\pi}{2} + \frac{\pi}{2n}$, we have $\frac{1}{n+1} \leq \gamma_1(t) \leq \frac{1}{n}$, and thus

$$\frac{2}{\pi} \cdot \frac{n}{n+1} = \frac{\frac{1}{n+1}}{\frac{\pi}{2n}} \leq \frac{\gamma_1(t) - \gamma_1\left(\frac{\pi}{2}\right)}{t - \frac{\pi}{2}} \leq \frac{\frac{1}{n}}{\frac{\pi}{2(n+1)}} = \frac{2}{\pi} \cdot \frac{n+1}{n}$$

and so by the squeeze theorem

$$\lim_{t\to\frac{\pi}{2}+}\frac{\gamma_1(t)-\gamma_1\left(\frac{\pi}{2}\right)}{t-\frac{\pi}{2}}$$

exists and equals $\frac{2}{\pi}$. The limit from the left similarly equals $\frac{2}{\pi}$. So $\gamma_1'\left(\frac{\pi}{2}\right)$ exists and equals $\frac{2}{\pi}$.

Thus γ is differentiable at $\frac{\pi}{2}$ and $\gamma'\left(\frac{\pi}{2}\right)=\left(\frac{2}{\pi},0\right)$, so

$$\left\|\gamma'\left(\frac{\pi}{2}\right)\right\|=\frac{2}{\pi}<1\ .$$

Replacing the semicircles by curves c_n of arc length $\frac{2}{\sqrt{n}}-\frac{2}{\sqrt{n+1}}=O(n^{-3/2})$ lying in the square

$$\left[\frac{1}{n+1},\frac{1}{n}\right]\times\left[-\frac{1}{n^2},\frac{1}{n^2}\right]$$

(and similarly on the left) we can get a similar example where $\gamma'=\mathbf{0}$ at the origin. Much more pathological examples, where $\|\gamma'(t)\|<1$ for many t, can be constructed similarly.

VII.3.3. Length of Arcs

If C is a "curve" in $\mathbb{R}^d$, we would like to be able to say what the intrinsic "arc length" of C is. This is a tricky problem. The first question is what kind of subsets should be called "curves." It is too broad to say that the range of any parametrized curve is a "curve" (). One reasonable possibility is:

VII.3.3.17. Definition. An *arc* in $\mathbb{R}^d$ is a subset of $\mathbb{R}^n$ homeomorphic to $[0,1]$.

VII.3.3.18. A subset C of $\mathbb{R}^d$ is an arc if and only if it is the range of an injective parametrized curve: a homeomorphism γ from $[0,1]$ to C is such a parametrized curve, and conversely an injective parametrized curve $\gamma:[a,b]\to\mathbb{R}^n$ with $C=\gamma([a,b])$ is a homeomorphism by (). An arc thus has a unique equivalence class of oriented injective parametrizations, thus two in all; the arc length of all these parametrized curves is the same and can be taken to be the arc length of C.

VII.3.3.19. Slightly more generally, if C can be written as a finite union of arcs with only endpoints in common, the arc length of C can be taken to be the sum of the arc lengths of these arcs. For more complicated subsets, there are grave difficulties in defining arc length this way.

VII.3.3.20. There is a way to unambiguously define "arc length" for general (Borel) subsets of $\mathbb{R}^d$: one-dimensional Hausdorff measure (). This "arc length" has the nice properties of a measure, generalizing VII.3.1.12.. Subsets whose one-dimensional Hausdorff measure is positive and finite can be regarded as "one-dimensional" (cf. (); such a set need not be topologically one-dimensional, and not every "one-dimensional" subset will have these properties).

VII.3.4. Integrating a Function Along a Curve

VII.3.4.1. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a parametrized curve. Suppose f is a real-valued function on the image C of γ. We want to define the integral

$$\int_\gamma f$$

of f along γ. To get a reasonable definition, we need to assume $f \circ \gamma$ is continuous on $[a,b]$ (it suffices that f is continuous on C) and that γ is rectifiable. The prototype physical interpretation is to think of C as a thin wire of variable density $f(\boldsymbol{x})$ at the point $\boldsymbol{x}$; the integral should be the total mass of the wire. There are many other similar interpretations and applications of this type of integral.

VII.3.4.2. Definition. Let $\gamma : [a,b] \to \mathbb{R}^d$ be a (rectifiable) curve in $\mathbb{R}^d$ with image C, and f a real-valued function on C. If $\dot{\mathcal{P}} = \{a = t_0, \dots, t_n = b; s_1, \dots, s_n\}$ is a tagged partition of $[a,b]$ with $\Delta\boldsymbol{x}_j = \gamma(t_j) - \gamma(t_{j-1})$, set

$$\mathcal{R}(f,\gamma,\dot{\mathcal{P}}) = \sum_{j=1}^{n} f(\gamma(s_j))\|\Delta\boldsymbol{x}_j\| \ .$$

Set

$$\int_\gamma f = \lim_{\|\mathcal{P}\|\to 0} \mathcal{R}(f,\gamma,\dot{\mathcal{P}})$$

if the limit exists (and is finite).

VII.3.4.3. Rectifiability of γ is not a necessary part of the definition, but the integral will rarely exist (and will not have a reasonable interpretation) if γ is not rectifiable. Note that if γ is rectifiable and f is the constant function 1, then $\int_\gamma 1$ exists and equals $\ell(\gamma)$ by definition (and the integral will not exist if γ is not rectifiable). One would also not expect the limit to exist unless f is (at least almost) continuous. The case of continuous f is almost exclusively the one used in applications anyway.

An integral of this type is sometimes called a *line integral*, but this term is more frequently reserved for an integral of the type described in (). The term *curvilinear integral* is also used. Note that this integral is an *unoriented integral*, in contrast to the one in (): we have

$$\int_{-\gamma} f = \int_\gamma f$$

for any γ and f (for which the integral is defined).

VII.3.4.4. The integral of f over γ is clearly independent of the parametrization of γ. In fact, as in VII.3.1.6., using a parametrization at all is just a convenient way of describing partitions of the curve. If $\{\boldsymbol{p} = \boldsymbol{x}_0, \dots, \boldsymbol{x}_n = \boldsymbol{q}\}$ is a "partition" of the curve corresponding to a partition $\mathcal{P} = \{a = t_0, \dots, t_n = b\}$ as in VII.3.1.6., and for each j we choose a "tag" $\boldsymbol{y}_j$ in the j'th "segment" of the curve by taking $\boldsymbol{y}_j = \gamma(s_j)$ for some $s_j \in [t_{j-1}, t_j]$, then we have a "Riemann sum" estimate

$$\sum_{j=1}^{n} f(\boldsymbol{y}_j)\|\Delta\boldsymbol{x}_j\|$$

where $\Delta \boldsymbol{x}_j = \boldsymbol{x}_j - \boldsymbol{x}_{j-1}$, so everything is expressible without reference to a parametrization (although defining what constitutes a suitable "partition" is tricky without using a parametrization unless the curve is an arc). However, it is tricky to express the limit as $\|\mathcal{P}\| \to 0$ without reference to parametrizations unless the curve is an arc, although it is independent of the parametrization chosen (it was easier for arc length since the limit there was just a supremum, but this is not necessarily true for integrating a function along the curve).

Our main result gives a way of computing the integral by an ordinary one-dimensional integral:

VII.3.4.5. THEOREM. Let $\gamma : [a, b] \to \mathbb{R}^n$ be a piecewise-smooth ($\mathcal{C}^1$) curve in $\mathbb{R}^n$ with image C, and f a real-valued function on C such that $f \circ \gamma$ is continuous. Then $\int_\gamma f$ exists and

$$\int_\gamma f = \int_a^b f(\gamma(t)) \, \|\gamma'(t)\| \, dt \ .$$

VII.3.4.6. COROLLARY. Let $\gamma : [a, b] \to \mathbb{R}^n$ be a piecewise-smooth ($\mathcal{C}^1$) curve in $\mathbb{R}^n$. Then

$$\ell(\gamma) = \int_a^b \|\gamma'(t)\| \, dt \ .$$

We now prove Theorem VII.3.4.5..

PROOF: By reducing to subintervals we may assume γ is $\mathcal{C}^1$ on $[a, b]$. Let $\gamma_1, \ldots, \gamma_n$ be the coordinate functions of γ. Let $\mathcal{P} = \{a = t_0, \ldots, t_m = b\}$ be a partition of $[a, b]$. For each j and k, by the MVT there is an $s_j^{(k)} \in (t_{j-1}, t_j)$ such that

$$\gamma_k(t_j) - \gamma_k(t_{j-1}) = \gamma_k' \left(s_j^{(k)} \right) \Delta t_j$$

where $\Delta t_j = t_j - t_{j-1}$. If we could choose all the $s_j^{(k)}$ to be the same s_j for each j, the proof would be simple: if we take $\dot{\mathcal{P}} = \{a = t_0, \ldots, t_m = b; s_1, \ldots, s_m\}$ we would have

$$\mathcal{R}(f, \gamma, \dot{\mathcal{P}}) = \sum_{j=1}^m f(\gamma(s_j)) \sqrt{\gamma_1'(s_j)^2 + \cdots + \gamma_n'(s_j)^2} \, \Delta t_j$$

and as $\|\mathcal{P}\| \to 0$ these Riemann sums converge to

$$\int_a^b f(\gamma(t)) \, \|\gamma'(t)\| \, dt$$

since the integrand is continuous on $[a, b]$ ().

However, the $s_j^{(k)}$ cannot be all chosen the same for all k in general, complicating the proof. First, since $f \circ \gamma$ is continuous on $[a,b]$, it is bounded; fix M with $|f(\gamma(t))| \leq M$ for $t \in [a,b]$. Then, if $\epsilon > 0$, since each γ_k' is uniformly continuous on $[a,b]$, there is a $\delta > 0$ such that if $\|\mathcal{P}\| < \delta$, then

$$|\gamma_k'(t) - \gamma_k'(s)| < \frac{\epsilon}{M}$$

for $s, t \in [t_{j-1}, t_j]$ for all j.

Now fix an $s_j \in [t_{j-1}, t_j]$ for each j, and set $\dot{\mathcal{P}} = \{a = t_0, \ldots, t_m = b; s_1, \ldots, s_m\}$. If the $s_j^{(k)}$ are as above, then

$$\mathcal{R}(f, \gamma, \dot{\mathcal{P}}) = \sum_{j=1}^{m} f(\gamma(s_j)) \, \|\Delta \boldsymbol{x}_j\| = \sum_{j=1}^{m} f(\gamma(s_j)) \sqrt{\gamma_1'(s_j^{(1)})^2 + \cdots + \gamma_n'(s_j^{(n)})^2} \, \Delta t_j \ .$$

Set

$$\mathcal{S}(f, \gamma, \dot{\mathcal{P}}) = \sum_{j=1}^{m} f(\gamma(s_j)) \sqrt{\gamma_1'(s_j)^2 + \cdots + \gamma_n'(s_j)^2} \, \Delta t_j \ .$$

Then

$$\lim_{\|\mathcal{P}\| \to 0} \mathcal{S}(f, \gamma, \dot{\mathcal{P}}) = \int_a^b f(\gamma(t)) \, \|\gamma'(t)\| \, dt$$

as above. But if $\|\mathcal{P}\| < \delta$, we have

$$|\mathcal{R}(f, \gamma, \dot{\mathcal{P}}) - \mathcal{S}(f, \gamma, \dot{\mathcal{P}})| = \left| \sum_{j=1}^{m} f(\gamma(s_j)) \left[\sqrt{\gamma_1'(s_j^{(1)})^2 + \cdots + \gamma_n'(s_j^{(n)})^2} - \sqrt{\gamma_1'(s_j)^2 + \cdots + \gamma_n'(s_j)^2} \right] \Delta t_j \right|$$

$$\leq \sum_{j=1}^{m} |f(\gamma(s_j))| \left| \sqrt{\gamma_1'(s_j^{(1)})^2 + \cdots + \gamma_n'(s_j^{(n)})^2} - \sqrt{\gamma_1'(s_j)^2 + \cdots + \gamma_n'(s_j)^2} \right| \Delta t_j$$

$$\leq \sum_{j=1}^{m} M \left[\sum_{k=1}^{n} |\gamma_k'(s_j^{(k)}) - \gamma_k'(s_j)| \right] \Delta t_j \leq \sum_{j=1}^{m} Mn \frac{\epsilon}{M} \Delta t_j = n\epsilon(b-a) \ .$$

Since $\epsilon > 0$ is arbitrary, we also have

$$\lim_{\|\mathcal{P}\| \to 0} \mathcal{R}(f, \gamma, \dot{\mathcal{P}}) = \int_a^b f(\gamma(t)) \, \|\gamma'(t)\| \, dt \ .$$



VII.3.4.7. There is a generalization along the lines of VII.3.2.9., but it is somewhat delicate. Details are left to the interested reader.

VII.3.5. Exercises

VII.3.5.1. (a) Let $\gamma : [a, b] \to \mathbb{R}^n$ be a parametrized curve, with coordinate functions $\gamma_1, \ldots, \gamma_n$. Show that the total variation (XIX.4.1.6.) satisfies $V_{[a,b]}(\gamma_k) \leq \ell(\gamma) \leq \sum_{j=1}^{n} V_{[a,b]}(\gamma_j)$ for any k, and thus γ is rectifiable if and only if all coordinate functions of γ have bounded variation. [For each k and each partition $\mathcal{P}$ of $[a, b]$, we have $V(\gamma_k, \mathcal{P}) \leq \ell(\gamma, \mathcal{P}) \leq \ell(\gamma)$.]

(b) Let $\gamma : [a, b] \to \mathbb{R}^n$ be a rectifiable parametrized curve, and ϕ the function defined in VII.3.2.2.. Show that ϕ is absolutely continuous if and only if each coordinate function of γ is absolutely continuous. [If each γ_k is absolutely continuous, and $\epsilon > 0$, for each k choose δ_k which works in the definition of absolute continuity for γ_k for $\frac{\epsilon}{n}$. Show that $\delta = \min_k(\delta_k)$ works in the definition for ϕ and ϵ.]

VII.3.5.2. Prove the following result complementary to VII.3.2.5.:

THEOREM. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a parametrized curve. If $\ell(\gamma) = \infty$, then for any $M > 0$ there is a $\delta > 0$ such that $\ell(\gamma, \mathcal{P}) > M$ for any partition $\mathcal{P}$ of $[a, b]$ with $\|\mathcal{P}\| < \delta$.

VII.3.5.3. Carry out the details of the last construction in VII.3.2.16..

VII.4. Curves in $\mathbb{R}^2$

VII.5. Curves in $\mathbb{R}^3$

VII.5.1. Exercises

Chapter VIII

Multivariable and Vector Calculus

In this chapter, we discuss calculus for functions from one Euclidean space $\mathbb{R}^n$ (or a subset) to another Euclidean space $\mathbb{R}^m$. The terms *multivariable calculus* and *vector calculus* used are roughly synonymous, but "vector calculus" is slightly more general: "multivariable calculus" refers to the case where $n > 1$ (and $m \geq 1$), while "vector calculus" also includes the case where $n = 1$ and $m > 1$, which is exactly the case of curves in $\mathbb{R}^m$ discussed in Chapter (). (In developing the theories, we work with general n and m, not even excluding the case $n = m = 1$.) There is also a slight difference in point of view: multivariable calculus is generally thought of as consideration of real-valued functions, or systems of real-valued functions, of several real variables, while vector calculus is thought of as being the study of functions whose input and output are vectors in Euclidean spaces, not necessarily of the same dimension. (The two points of view are complementary and both are needed to get the full picture of the subject.)

Consideration of functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ has distinct and complicating features over the case $n = m = 1$ arising from having either $n > 1$ or $m > 1$ (or both). The complications caused by considering the case $m > 1$ are quite modest, however. If $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is a function, write

$$\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$$

where $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are the coordinate functions of $\boldsymbol{f}$ (technically, $f_k = \pi_k^{(m)} \circ \boldsymbol{f}$, where $\pi_k^{(m)} : \mathbb{R}^m \to \mathbb{R}$ is projection onto the k'th coordinate, i.e. $\pi_k^{(m)}(x_1, \ldots, x_m) = x_k$), then analysis of $\boldsymbol{f}$ can be largely reduced to analysis of the f_k. There are situations where it is undesirable to reduce to the coordinate functions, though, such as in the study of curves, and in results like the Chain Rule and the Implicit and Inverse Function Theorems, where the interplay of coordinates is crucial. Roughly speaking, the multivariable calculus approach is to reduce to coordinate functions and the vector calculus approach is to avoid doing this.

The complications arising from having $n > 1$ are essential, however, and mostly cannot be handled by considering coordinates separately (although, with considerable care, in good cases much can be reduced to one-variable calculus). Thus multivariable calculus (and vector calculus) has a quite different nature than one-variable calculus.

We also make a (somewhat arbitrary and nonstandard) distinction between *vector calculus*, the material covered in this chapter, and *vector analysis*, which includes results related to divergence and curl, as well as properties of curves, surfaces, and their generalizations, and differential forms (really the beginning of differential geometry).

VIII.1. Limits and Continuity for Vector Functions

Defining limits and continuity for functions of more than one variable can be very easy or very tricky, depending whether one takes the proper point of view.

VIII.1.1. Graphs of Functions of Several Variables

VIII.1.2. Limits of Sequences in $\mathbb{R}^n$

VIII.1.2.1. The first observation is that there is much more "room" in $\mathbb{R}^n$, $n \geq 2$, for sequences to behave in a complicated way than in $\mathbb{R} = \mathbb{R}^1$. A sequence in $\mathbb{R}$ can approach a limit only from the left or right, or more generally jump back and forth from left or right, i.e. can be decomposed into at most two subsequences which approach from the left and right respectively. But even in $\mathbb{R}^2$, a sequence can approach a limit in a very complicated way: it can approach from any direction or along any curve, can spiral around the limit wildly, etc., and cannot generally be decomposed into subsequences which approach in a controlled way (things can be even wilder in $\mathbb{R}^3$). Trying to make a definition which covers all the possibilities may seem like a daunting task.

VIII.1.2.2. However, there is a very simple solution. The definition of limit of a sequence in $\mathbb{R}$ can be used almost verbatim in $\mathbb{R}^n$ and gives exactly the right thing. The definition only needs slight redecoration to work perfectly:

VIII.1.2.3. Definition. Let $(\boldsymbol{x}_k)$ be a sequence in $\mathbb{R}^n$, and $\boldsymbol{x} \in \mathbb{R}^n$. Then $\boldsymbol{x}_k \to \boldsymbol{x}$, or $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}$, if for every $\epsilon > 0$ there is a K such that

$$\|\boldsymbol{x}_k - \boldsymbol{x}\| < \epsilon$$

for all $k \geq K$.

This is a special case of redecorating the definition in $\mathbb{R}$ to one valid in a general metric space ().

The Coordinatewise Criterion

There is one basic fact about limits in $\mathbb{R}^n$ which will appear in various forms, which allows a reduction to the case of limits in $\mathbb{R}$ in some cases, including sequences:

VIII.1.2.4. Proposition. [Coordinatewise Criterion for Sequences in $\mathbb{R}^n$] Let $(\boldsymbol{x}_k)$ be a sequence in $\mathbb{R}^n$, and $\boldsymbol{x} \in \mathbb{R}^n$. Write

$$\boldsymbol{x}_k = (x_k^{(1)}, \ldots, x_k^{(n)}), \quad \boldsymbol{x} = (x^{(1)}, \ldots, x^{(n)})$$

in coordinates. Then $\boldsymbol{x}_k \to \boldsymbol{x}$ in $\mathbb{R}^n$ if and only if $x_k^{(j)} \to x^{(j)}$ in $\mathbb{R}$ for each j, $1 \leq j \leq n$.

Proof: Note that for each m we have

$$|x_k^{(m)} - x^{(m)}| \leq \|\boldsymbol{x}_k - \boldsymbol{x}\| \leq \sum_{j=1}^{n} |x_k^{(j)} - x^{(j)}| \ .$$

Suppose $\boldsymbol{x}_k \to \boldsymbol{x}$, and fix j. Let $\epsilon > 0$. There is a K such that

$$|x_k^{(j)} - x^{(j)}| \leq \|\boldsymbol{x}_k - \boldsymbol{x}\| < \epsilon$$

for all $k \geq K$; thus $x_k^{(j)} \to x^{(j)}$. Conversely, suppose $x_k^{(j)} \to x^{(j)}$ for all j. Let $\epsilon > 0$. For each j choose K_j such that

$$|x_k^{(j)} - x^{(j)}| < \frac{\epsilon}{n}$$

for all $k \geq K_j$. Set $K = \max(K_1, \ldots, K_n)$. Then, if $k \geq K$,

$$\|\boldsymbol{x}_k - \boldsymbol{x}\| \leq \sum_{j=1}^{n} |x_k^{(j)} - x^{(j)}| < n \cdot \frac{\epsilon}{n} = \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $\boldsymbol{x}_k \to \boldsymbol{x}$. 👍

The Bolzano-Weierstrass Theorem in $\mathbb{R}^n$

The Bolzano-Weierstrass Theorem holds also for sequences of vectors in $\mathbb{R}^n$:

VIII.1.2.5. THEOREM. [BOLZANO-WEIERSTRASS THEOREM IN $\mathbb{R}^n$] Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.

PROOF: Let $(\boldsymbol{x}_k) = ((x_k^{(1)}, \ldots, x_k^{(n)}))$ be a bounded sequence in $\mathbb{R}^n$. Then $(x_k^{(1)})$ is a bounded sequence in $\mathbb{R}$, hence has a convergent subsequence $(x_{k_{m,1}}^{(1)})$. The sequence $(x_{k_{m,1}}^{(2)})$ is bounded, so has a convergent subsequence $(x_{k_{m,2}}^{(2)})$; we also have that $(x_{k_{m,2}}^{(1)})$ converges since it is a subsequence of $(x_{k_{m,1}}^{(1)})$. Continuing n times, we get a subsub$\cdots$subsequence $(k_{m,n})$ of $\mathbb{N}$ such that $(x_{k_{m,n}}^{(j)})$ converges for $1 \leq j \leq n$, Thus the subsequence $(\boldsymbol{x}_{k_{m,n}})$ of $(\boldsymbol{x}_k)$ converges by the coordinatewise criterion. 👍

VIII.1.3. Multivariable Limits

We now look at limits for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. We first consider the case $m = 1$.

Limits for Real-valued Functions of Several Variables

VIII.1.3.1. Suppose f is a real-valued function defined on a deleted neighborhood U of $\boldsymbol{a} \in \mathbb{R}^n$. We want to say what $\lim_{\boldsymbol{x} \to \boldsymbol{a}} f(\boldsymbol{x})$ means. We first consider some examples with $n = 2$, where we change notation slightly by writing (x, y) for $\boldsymbol{x}$.

VIII.1.3.2. EXAMPLES. (i) Let

$$f(x,y) = \frac{xy}{x^2+y^2}$$

for $(x,y) \neq (0,0)$. Then f is defined in a deleted neighborhood of $(0,0)$ in $\mathbb{R}^2$. What should $\lim_{(x,y)\to(0,0)} f(x,y)$ be?

If $(0,0)$ is approached along either the x- or y-axis, the function is identically 0, so the limit could only reasonably be 0. On the other hand, $f(t,t) = \frac{1}{2}$ for all $t \neq 0$, so if we approach the origin along the line $x = y$ the function value approaches $\frac{1}{2}$. Since the limit should be unique and we approach different values from different directions, we should say the limit does not exist.

Geometrically, the graph is so wrinkled near the origin that no point can be added there to make the graph a reasonable unbroken surface (Figure VIII.1).

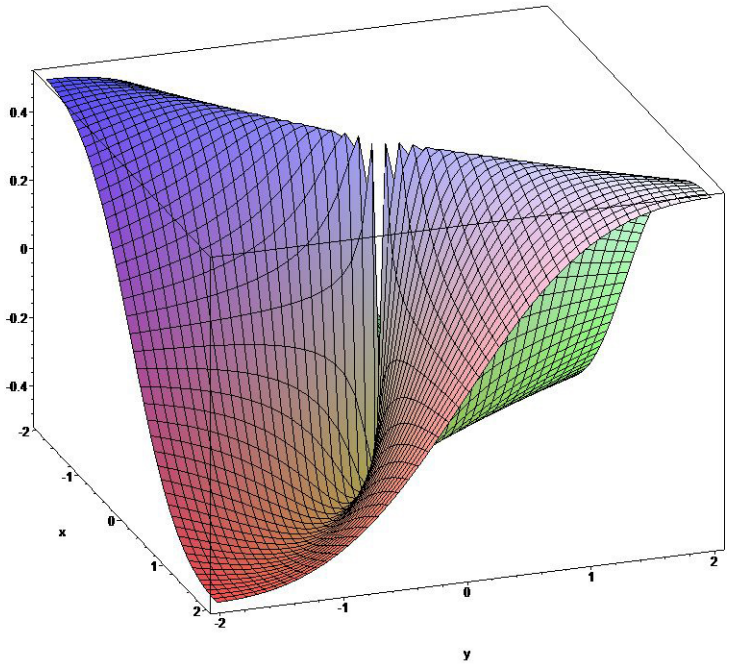


Figure VIII.1: Maple plot of $f(x,y) = \frac{xy}{x^2+y^2}$

(ii) Here is a more striking version of the same phenomenon: Let

$$f(x,y) = \frac{x^2 y}{x^4 + y^2}$$

for $(x,y) \neq (0,0)$. Then f is defined in a deleted neighborhood of $(0,0)$ in $\mathbb{R}^2$. What should $\lim_{(x,y)\to(0,0)} f(x,y)$ be?

Again, if $(0,0)$ is approached along either the x- or y-axis, the function is identically 0, so the limit could only reasonably be 0. And if we approach along the line $y = mx$, we have

$$\lim_{t\to 0} f(t,mt) = \lim_{t\to 0} \frac{mt^3}{t^4 + m^2 t^2} = 0$$

thus the function approaches 0 as $(0,0)$ is approached radially from any direction. Does this mean that we should say the limit is 0?

Suppose we approach along the parabola $y = x^2$. We then have

$$\lim_{t\to 0} f(t,t^2) = \lim_{t\to 0} \frac{t^4}{2t^4} = \frac{1}{2}$$

so the limit cannot exist. This time the crinkle in the graph is curved (Figure VIII.2).

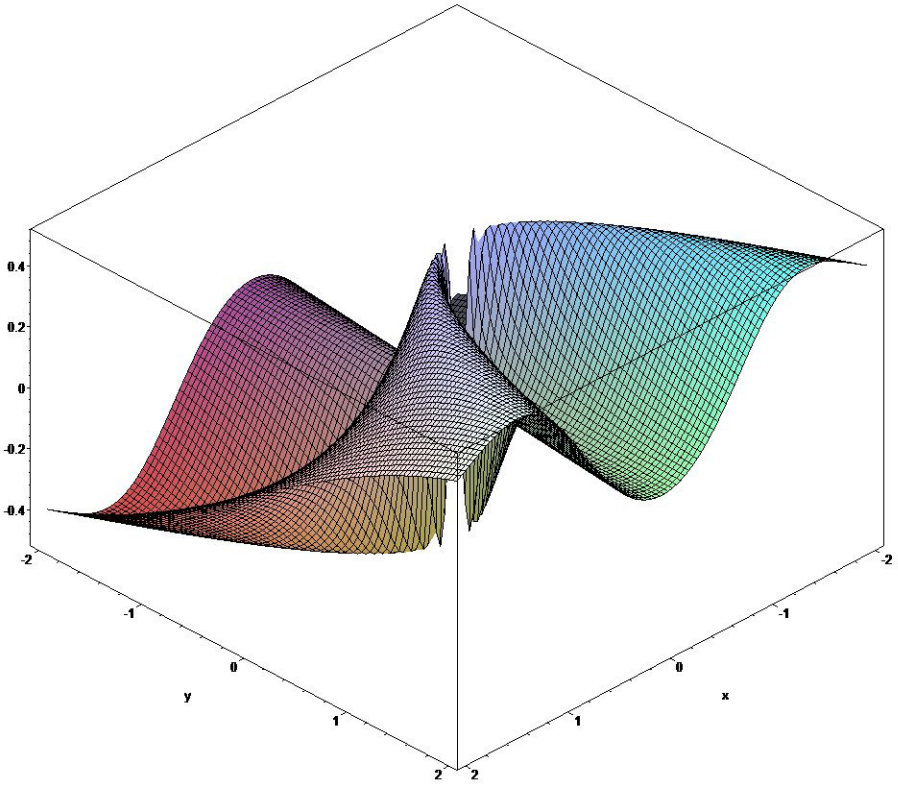


Figure VIII.2: Maple plot of $f(x,y) = \frac{x^2 y}{x^4+y^2}$

To make a good definition, we can again simply redecorate the one-dimensional definition:

VIII.1.3.3. Definition. Let $\boldsymbol{f}$ be a function from a deleted neighborhood U of $\boldsymbol{a} \in \mathbb{R}^n$ to $\mathbb{R}^m$, $\boldsymbol{L} \in \mathbb{R}^m$. Then

$$\lim_{\boldsymbol{x} \to \boldsymbol{a}} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{L}$$

if, for every $\epsilon > 0$, there is a $\delta > 0$ such that

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{L}\| < \epsilon$$

whenever $\boldsymbol{x} \in U$ and $0 < \|\boldsymbol{x} - \boldsymbol{a}\| < \delta$.

As in the case of sequences, limits can be taken coordinatewise. The proof of the next proposition is almost identical to the proof of VIII.1.2.4., and is left to the reader:

VIII.1.3.4. Proposition. [Coordinatewise Criterion for Functions from $\mathbb{R}^n$ to $\mathbb{R}^m$] Let $\boldsymbol{f} = (f_1, \ldots, f_m)$ be a function from a deleted neighborhood U of $\boldsymbol{a} \in \mathbb{R}^n$ to $\mathbb{R}^m$, $\boldsymbol{L} = (L_1, \ldots, L_m) \in \mathbb{R}^m$. Then

$$\lim_{\boldsymbol{x} \to \boldsymbol{a}} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{L}$$

if and only if

$$\lim_{\boldsymbol{x} \to \boldsymbol{a}} f_k(\boldsymbol{x}) = L_k$$

for $1 \leq k \leq m$.

VIII.1.3.5. Note that the coordinatewise criterion is in the *range* space $\mathbb{R}^m$, not the domain space. The previous examples show that it is not adequate to work coordinatewise in the domain. In fact, the limits in those examples do not exist under our definition.

Many properties of limits in the one-dimensional case hold for multidimensional limits, with essentially the same (redecorated) proofs, such as:

VIII.1.3.6. Proposition. Limits are unique: if $\boldsymbol{f}$ is a function from a deleted neighborhood U of $\boldsymbol{a} \in \mathbb{R}^n$ to $\mathbb{R}^m$, $\boldsymbol{L}, \boldsymbol{M} \in \mathbb{R}^m$, and

$$\lim_{\boldsymbol{x} \to \boldsymbol{a}} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{L} \text{ and } \lim_{\boldsymbol{x} \to \boldsymbol{a}} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{M}$$

then $\boldsymbol{L} = \boldsymbol{M}$.

Continuity

We may make the same definition of (pointwise) continuity in the multivariable case as in one dimension:

VIII.1.3.7. DEFINITION. Let $\boldsymbol{f}$ be a function from a neighborhood U of $\boldsymbol{a} \in \mathbb{R}^n$ to $\mathbb{R}^m$, $\boldsymbol{L} \in \mathbb{R}^m$. Then $\boldsymbol{f}$ is *continuous at* $\boldsymbol{a}$ if, for every $\epsilon > 0$, there is a $\delta > 0$ such that

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\| < \epsilon$$

whenever $\boldsymbol{x} \in U$ and $\|\boldsymbol{x} - \boldsymbol{a}\| < \delta$.

VIII.1.3.8. As in the one-dimensional case, $\boldsymbol{f}$ is continuous at $\boldsymbol{a}$ if and only if

$\boldsymbol{f}(\boldsymbol{a})$ exists.

$\lim_{\boldsymbol{x} \to \boldsymbol{a}} \boldsymbol{f}(\boldsymbol{x})$ exists.

$\lim_{\boldsymbol{x} \to \boldsymbol{a}} \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{a})$.

Continuity on a Set

Sequential Criteria for Limits and Continuity

VIII.2. Partial and Directional Derivatives

If f is a function of several variables, one simple way to partially analyze the behavior of f is to hold all but one of the variables fixed, so that f becomes simply a function of the one remaining variable. This leads to the notion of partial derivative. Partial derivatives turn out to give considerably more information about the function than might be at first expected, at least in good cases. The notion can then be generalized to directional derivatives in any direction.

VIII.2.1. Partial Derivatives

VIII.2.1.1. DEFINITION. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$ a function. Let $\boldsymbol{a} = (a_1, \ldots, a_n) \in U$. For $1 \leq k \leq n$, we define the *partial derivative*

$$\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{a}) = \lim_{h \to 0} \frac{\boldsymbol{f}(a_1, \ldots, a_{k-1}, a_k + h, a_{k+1}, \ldots, a_n) - \boldsymbol{f}(a_1, \ldots, a_n)}{h}$$

$$= \lim_{h \to 0} \frac{1}{h}[\boldsymbol{f}(a_1, \ldots, a_{k-1}, a_k + h, a_{k+1}, \ldots, a_n) - \boldsymbol{f}(a_1, \ldots, a_n)]$$

if the limit exists.

In other words, $\frac{\partial \boldsymbol{f}}{\partial x_k}$ is the derivative of $\boldsymbol{f}$ when it is thought of as only a function of x_k by holding the other variables fixed: if

$$\boldsymbol{g}(x) = \boldsymbol{f}(a_1, \ldots, a_{k-1}, x, a_{k+1}, \ldots, a_n)$$

then

$$\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{a}) = \boldsymbol{g}'(a_k) \ .$$

VIII.2.1.2. If $\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{a})$ exists, it is a vector in $\mathbb{R}^m$. Partial derivatives are most commonly discussed in the case $m = 1$, in which case $\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{a}) \in \mathbb{R}$. The relation with the general case is that if $(f_1, \ldots, f_m)$ are the coordinate functions of $\boldsymbol{f}$, then $\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{a})$ exists if and only if $\frac{\partial f_j}{\partial x_k}(\boldsymbol{a})$ exists for all $1 \leq j \leq m$, in which case

$$\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{a}) = \left(\frac{\partial f_1}{\partial x_k}(\boldsymbol{a}), \ldots, \frac{\partial f_m}{\partial x_k}(\boldsymbol{a}) \right) \ .$$

VIII.2.1.3. The partial derivative symbol ∂ appears to be an italic Cyrillic d. However, the origin of the notation is somewhat obscure (). Alternate notations are sometimes used, especially if the variables are denoted $x, y, z, \ldots$: if f is a function of, say, x, y, and z, $\frac{\partial f}{\partial x}$ is sometimes denoted f_x or f_1, $\frac{\partial f}{\partial y}$ is denoted f_y or f_2, etc. We will mostly stick to the $\frac{\partial}{\partial x}$, etc., notation, except in ().

VIII.2.1.4. It is easy to get caught in confusing or ambiguous notation concerning partial derivatives if one is not careful, and many references use sloppy notation in places. For example, suppose $u = f(x, y, z)$ and $z = g(x, y)$; then what does $\frac{\partial u}{\partial x}$ mean? As a variation, if $u = f(x, y)$ and $y = g(x)$, then one can make sense of both $\frac{\partial u}{\partial x}$ and $\frac{du}{dx}$. To avoid such nasty ambiguities, we will sometimes seem to be overly pedantic in our notation. See also Exercise VIII.2.4.2..

"All notations for the partial derivative omit crucial information: which variables are being kept constant. In modeling real phenomena, it can be difficult even to know what all the variables are. But if you don't, your partial derivatives may be meaningless."

A. Douady[1]

VIII.2.1.5. Like ordinary derivatives, partial derivatives are defined pointwise for each fixed $\boldsymbol{a} \in U$. However, after the fact each partial derivative $\frac{\partial \boldsymbol{f}}{\partial x_k}$ may be thought of as a function from U, or the subset on which it is defined, to $\mathbb{R}^m$.

VIII.2.1.6. When a function f of two or more variables is defined by a formula, partial derivatives can be calculated by the usual rules of differentiation, treating the remaining variables as constants. For example, if

$$f(x,y,z) = x^2y^3 + \frac{\cos yz}{z^2+1} - e^{xyz}$$

we have

$$\frac{\partial f}{\partial x}(x,y,z) = 2xy^3 - yze^{xyz}$$

$$\frac{\partial f}{\partial y}(x,y,z) = 3x^2y^2 - \frac{z\sin yz}{z^2+1} - xze^{xyz}$$

$$\frac{\partial f}{\partial z}(x,y,z) = \frac{(z^2+1)(-y\sin yz) - (\cos yz)(2z)}{(z^2+1)^2} - xye^{xyz} \ .$$

Showing that such calculations are valid takes a little (easy) work, though, a point overlooked in many references; see Exercise VIII.2.4.1..

Partial Derivatives and Continuity

VIII.2.1.7. A differentiable function from $\mathbb{R}$ to $\mathbb{R}$ is continuous. But existence of all partial derivatives does not imply continuity in the higher-dimensional case. For example, let

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{If} \quad (x,y) \neq (0,0) \\ 0 & \text{If} \quad (x,y) = (0,0) \end{cases} \ .$$

Then f has partial derivatives with respect to both x and y at all points of $\mathbb{R}^2$. But f is not continuous at $(0,0)$ (VIII.1.3.2.). See VIII.3.7.8. for a more dramatic example.

However, continuity of all partial derivatives does imply continuity of the function (VIII.3.1.14., VIII.3.3.1.) (note that the partials of this f are discontinuous at $(0,0)$).

There is one special case which is elementary and worth noting:

[1] [HH99, p. 103].

VIII.2.1.8. PROPOSITION. Let U be a nonempty connected open set in $\mathbb{R}^n$, and $f : U \to \mathbb{R}$ a function. If all partial derivatives of f exist at all points of U and are identically zero, then f is constant on U.

PROOF: First suppose U is an open ball (or, more generally, convex). It follows from () that f must be constant on the intersection with U of any line parallel to a coordinate axis. Since any two points of U can be connected by a finite sequence of such line segments within U, f is constant on U.

Now suppose U is just connected. It follows that f is locally constant, i.e. for each $\lambda \in \mathbb{R}$ the set

$$V_\lambda = \{\boldsymbol{x} \in U : f(\boldsymbol{x}) = \lambda\}$$

is an open subset of U. If we knew that f were continuous, it would follow that V_λ is relatively closed in U, hence V_λ would have to be $\emptyset$ or all of U. But we cannot assume that f is continuous, so we argue as follows. Let $\boldsymbol{a} \in U$, and $f(\boldsymbol{a}) = \lambda_0$. Then V_{λ_0} is an open subset of U containing $\boldsymbol{a}$. We have

$$W := U \setminus V_{\lambda_0} = \bigcup_{\lambda \neq \lambda_0} V_\lambda$$

so W is a union of open sets, hence open, and U is the disjoint union of V_{λ_0} and W; thus $W = \emptyset$ since U is connected. 👍

VIII.2.1.9. COROLLARY. Let U be a nonempty connected open set in $\mathbb{R}^n$, and $f, g : U \to \mathbb{R}$ two functions. If all partial derivatives of f and g exist at all points of U and $\frac{\partial f}{\partial x_k}(\boldsymbol{x}) = \frac{\partial g}{\partial x_k}(\boldsymbol{x})$ for all k and all $\boldsymbol{x} \in U$, then $f - g$ is constant on U.

Higher-Order Partials

VIII.2.1.10. If $\boldsymbol{f} : U \to \mathbb{R}^m$ is a function, and the partial derivatives $\frac{\partial \boldsymbol{f}}{\partial x_k}$ are defined on U, then partial derivatives of these partial derivatives are also (potentially) defined. For example, if $\frac{\partial \boldsymbol{f}}{\partial x_k}$ is defined in a neighborhood of $\boldsymbol{a} \in U$, then for $1 \leq j \leq n$ we can define

$$\frac{\partial^2 \boldsymbol{f}}{\partial x_j \partial x_k}(\boldsymbol{a}) = \frac{\partial \left(\frac{\partial \boldsymbol{f}}{\partial x_k}\right)}{\partial x_j} = \lim_{h \to 0} \frac{\frac{\partial \boldsymbol{f}}{\partial x_k}(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots, a_n) - \frac{\partial \boldsymbol{f}}{\partial x_k}(a_1, \ldots, a_n)}{h}$$

if the limit exists.

Note the order of differentiation: $\frac{\partial^2 \boldsymbol{f}}{\partial x_j \partial x_k}(\boldsymbol{a})$ and $\frac{\partial^2 \boldsymbol{f}}{\partial x_k \partial x_j}(\boldsymbol{a})$ have different definitions if $j \neq k$. It turns out (VIII.2.2.2.) that under mild continuity hypotheses these two partial derivatives are numerically equal. ✋

VIII.2.1.11. We may keep going, defining third-order partial derivatives $\frac{\partial^3 \boldsymbol{f}}{\partial x_i \partial x_j \partial x_k}$, etc. The i, j, k need not be distinct.

VIII.2.1.12. In the definition, we may have $j = k$ or $j \neq k$. Thus, if f is a function $f(x, y)$ of two variables, there are four second-order partials: $\frac{\partial^2 \boldsymbol{f}}{\partial x \partial x}$ (usually denoted $\frac{\partial^2 \boldsymbol{f}}{\partial x^2}$), $\frac{\partial^2 \boldsymbol{f}}{\partial x \partial y}$, $\frac{\partial^2 \boldsymbol{f}}{\partial y \partial x}$, and $\frac{\partial^2 \boldsymbol{f}}{\partial y^2}$. The middle two, called *mixed partials*, are often but not always equal. The variety increases rapidly with higher derivatives and/or more variables.

VIII.2.1.13. We may also use the alternate notation f_{xx} or f_{11} for $\frac{\partial^2 \boldsymbol{f}}{\partial x^2}$, etc. When the alternate notation is used, it is conventional to take derivatives from left to right, unlike in the Leibniz notation where derivatives go from right to left: thus f_{xy} or f_{12} denotes $\frac{\partial^2 f}{\partial y \partial x}$, i.e. the function is first differentiated with respect to x and then with respect to y, not $\frac{\partial^2 f}{\partial x \partial y}$ which is denoted f_{yx} or f_{21}. (The distinction is academic in good cases.)

VIII.2.2. Equality of Mixed Partials

In general, the order in which the derivatives are taken in a higher-order partial derivative makes a difference. But under a mild continuity hypothesis, the order does not matter.

VIII.2.2.1. EXAMPLE. Let

$$f(x,y) = \begin{cases} \frac{xy(x^2-y^2)}{x^2+y^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases} .$$

Then f is $\mathcal{C}^1$ on $\mathbb{R}^2$: we have

$$\frac{\partial f}{\partial x}(x,y) = \begin{cases} \frac{y(x^4+4x^2y^2-y^4)}{(x^2+y^2)^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases}$$

$$\frac{\partial f}{\partial y}(x,y) = \begin{cases} \frac{-x(y^4+4x^2y^2-x^4)}{(x^2+y^2)^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases} .$$

The absolute value of the numerators are bounded by $6\|(x,y)\|^5$, so the partials are continuous at $(0,0)$. Also, $\frac{\partial^2 f}{\partial x\,\partial y}$ and $\frac{\partial^2 f}{\partial y\,\partial x}$ are defined everywhere on $\mathbb{R}^2$, and

$$\frac{\partial^2 f}{\partial x\,\partial y}(0,0) = \lim_{h\to 0} \frac{\frac{\partial f}{\partial y}(h,0) - \frac{\partial f}{\partial y}(0,0)}{h} = 1$$

$$\frac{\partial^2 f}{\partial y\,\partial x}(0,0) = \lim_{k\to 0} \frac{\frac{\partial f}{\partial x}(0,k) - \frac{\partial f}{\partial x}(0,0)}{k} = -1$$

so the second-order mixed partials exist at $(0,0)$ are defined but unequal.

However, $\frac{\partial^2 f}{\partial x\,\partial y}$ and $\frac{\partial^2 f}{\partial y\,\partial x}$ are not continuous at $(0,0)$.

Actually, f has continuous partials of all orders everywhere except at the origin, i.e. f is $\mathcal{C}^\infty$ on $\mathbb{R}^2\backslash\{(0,0)\}$.

The graph has a subtle crinkle near the origin (Figure VIII.3).

VIII.2.2.2. THEOREM. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^m$, $\boldsymbol{a} \in U$. Suppose, for some i and j, $\frac{\partial \boldsymbol{f}}{\partial x_i}$ and $\frac{\partial \boldsymbol{f}}{\partial x_j}$ exist everywhere on U, and $\frac{\partial^2 \boldsymbol{f}}{\partial x_i\,\partial x_j}$ exists everywhere on U and is continuous at $\boldsymbol{a}$. Then $\frac{\partial^2 \boldsymbol{f}}{\partial x_j\,\partial x_i}(\boldsymbol{a})$ is defined and

$$\frac{\partial^2 \boldsymbol{f}}{\partial x_j\,\partial x_i}(\boldsymbol{a}) = \frac{\partial^2 \boldsymbol{f}}{\partial x_i\,\partial x_j}(\boldsymbol{a}) .$$

PROOF: There is no content to the statement if $i = j$, so we assume $i \neq j$.

We first make two reductions. By considering the coordinate functions, it suffices to assume $m = 1$. And, since only the coordinates i and j are varied in computing the partials, we may assume $n = 2$. So we suppose

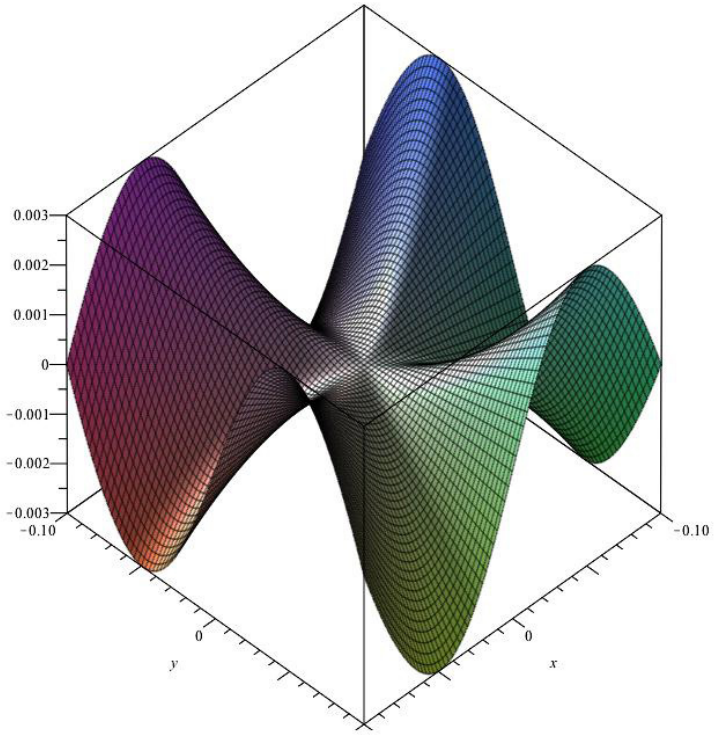


Figure VIII.3: Maple plot of graph of $f(x,y) = \frac{xy(x^2-y^2)}{x^2+y^2}$

$U \subseteq \mathbb{R}^2$, $f : U \to \mathbb{R}$, $(a,b) \in U$, that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist on U, and $\frac{\partial^2 f}{\partial x\,\partial y}$ is defined on U and continuous at (a,b). We want to show that $\frac{\partial^2 f}{\partial y\,\partial x}(a,b)$ is defined and equal to $c := \frac{\partial^2 f}{\partial x\,\partial y}(a,b)$.

Let $\epsilon > 0$, and fix $\delta > 0$ such that if $|h| < \delta$ and $|k| < \delta$, then $(a+h, b+k) \in U$ and

$$\left| \frac{\partial^2 f}{\partial x\,\partial y}(a+h, b+k) - c \right| < \epsilon \ .$$

Set $V = \{(h,k) : |h| < \delta, |k| < \delta\}$. Then for $(h,k) \in V$, set

$$\Delta(h,k) = f(a+h, b+k) - f(a, b+k) - f(a+h, b) + f(a,b) \ .$$

If $(h,k) \in V$ is fixed, for $0 \le t \le 1$ set

$$\phi(t) = f(a+h, b+tk) - f(a, b+tk) \ .$$

Then ϕ is differentiable on $[0,1]$ and by the Chain Rule

$$\phi'(t) = \frac{\partial f}{\partial y}(a+h, b+tk) \cdot k - \frac{\partial f}{\partial y}(a, b+tk) \cdot k \ .$$

Apply the MVT to get a $t \in (0,1)$ such that

$$k\frac{\partial f}{\partial y}(a+h, b+tk) - k\frac{\partial f}{\partial y}(a, b+tk) = \phi'(t)(1-0) = \phi(1) - \phi(0) = \Delta(h,k) \ .$$

Fix such a t, and for $0 \le s \le 1$ set

$$\psi(s) = k\frac{\partial f}{\partial y}(a+sh, b+tk) \ .$$

Then ψ is differentiable on $[0,1]$ and

$$\psi'(s) = k\frac{\partial^2 f}{\partial x\,\partial y}(a+sh, b+tk)\cdot h \ .$$

Again use the MVT to obtain an $s \in (0,1)$ such that

$$hk\frac{\partial^2 f}{\partial x\,\partial y}(a+sh, b+tk) = \psi'(s)(1-0) = \psi(1) - \psi(0) = \Delta(h,k) \ .$$

For every $(h,k) \in V$ there is such a pair $(s,t) \in (0,1)^2$ (which depends on (h,k)), and $(sh, tk) \in V$, so if $(h,k) \in V$, $h \neq 0$, $k \neq 0$, we have

$$\left| \frac{\Delta(h,k)}{hk} - c \right| < \epsilon \ .$$

Now fix k, $0 < |k| < \delta$. We have

$$\lim_{h\to 0} \frac{\Delta(h,k)}{hk} = \frac{1}{k}\lim_{h\to 0}\left[\frac{f(a+h,b+k) - f(a,b+k)}{h} - \frac{f(a+h,b) - f(a,b)}{h}\right] = \frac{1}{k}\left[\frac{\partial f}{\partial x}(a,b+k) - \frac{\partial f}{\partial x}(a,b)\right]$$

so the limit exists and we have that

$$\left| \frac{1}{k}\left[\frac{\partial f}{\partial x}(a,b+k) - \frac{\partial f}{\partial x}(a,b)\right] - c \right| \le \epsilon$$

if $0 < |k| < \delta$. Since $\epsilon > 0$ is arbitrary, we conclude that

$$\frac{\partial^2 f}{\partial y\,\partial x}(a,b) = \lim_{k\to 0} \frac{\frac{\partial f}{\partial x}(a,b+k) - \frac{\partial f}{\partial x}(a,b)}{k}$$

exists and equals c. 👍

VIII.2.2.3. REMARK. (i) In VIII.2.2.2., the hypothesis that $\frac{\partial f}{\partial x_j}$ exists on U is implicit and automatic in the hypothesis that $\frac{\partial^2 f}{\partial x_i\,\partial x_j}$ exists on U.

(ii) Let g be a continuous but nowhere differentiable function from $\mathbb{R}$ to $\mathbb{R}$, and let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x,y) = g(x)$. Then f is continuous, and $\frac{\partial f}{\partial y}$ exists everywhere and is identically zero; thus $\frac{\partial^2 f}{\partial x\,\partial y}$ also exists everywhere and is identically zero, hence continuous. But $\frac{\partial f}{\partial x}$ does not exist anywhere, so $\frac{\partial^2 f}{\partial y\,\partial x}$ does not exist anywhere. So the hypothesis that $\frac{\partial f}{\partial x_i}$ exists, which is necessary for the existence of $\frac{\partial^2 f}{\partial x_j\,\partial x_i}$, is not automatic under the other hypotheses.

VIII.2.2.4. COROLLARY. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ a $\mathcal{C}^2$ function from U to $\mathbb{R}^m$. Then, for any i and j,

$$\frac{\partial^2 \boldsymbol{f}}{\partial x_j \, \partial x_i} = \frac{\partial^2 \boldsymbol{f}}{\partial x_i \, \partial x_j}$$

on U.

VIII.2.2.5. VIII.2.2.4. is called *Clairaut's Theorem*. The version in VIII.2.2.2. can be found in [Str81, p. 377]. Statements of the result in some real analysis texts have unnecessary extra hypotheses, and the "proofs" in many books have flaws. See VIII.4.1.7. for a variation, and IX.3.12.1. for a simple alternate proof of VIII.2.2.4. based on Fubini's Theorem.

VIII.2.2.6. COROLLARY. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ a $\mathcal{C}^r$ function from U to $\mathbb{R}^m$ $(r > 1)$. Then any two mixed partials of $\boldsymbol{f}$ of total order $\leq r$ in which each coordinate direction appears the same number of times are equal on U.

To show this, apply VIII.2.2.4. repeatedly to $\boldsymbol{f}$ and its partial derivatives. For example, if $n \geq 5$ and $r \geq 4$, to show that

$$\frac{\partial^4 \boldsymbol{f}}{\partial x_2 \, \partial x_5 \, \partial x_3 \, \partial x_1} = \frac{\partial^4 \boldsymbol{f}}{\partial x_2 \, \partial x_3 \, \partial x_5 \, \partial x_1}$$

apply VIII.2.2.4. to $\frac{\partial \boldsymbol{f}}{\partial x_1}$ to conclude that

$$\frac{\partial^3 \boldsymbol{f}}{\partial x_5 \, \partial x_3 \, \partial x_1} = \frac{\partial^3 \boldsymbol{f}}{\partial x_3 \, \partial x_5 \, \partial x_1}$$

and then take the partial derivative of this function with respect to x_2. Any permutation of the order of derivatives can be built up from interchanges of successive derivatives, which can be justified in this way.

VIII.2.2.7. This result leads to some convenient notation. A *multi-index* of dimension n is an n-tuple $\alpha = (r_1, \ldots, r_n)$, where the r_k are in $\mathbb{N} \cup \{0\}$. The *total degree* of α is

$$|\alpha| = r_1 + \cdots r_n \ .$$

If α is a multi-index of dimension n and total degree r, U an open subset of $\mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$ is $\mathcal{C}^r$, write

$$\frac{\partial^\alpha \boldsymbol{f}}{\partial x^\alpha} = \frac{\partial^r \boldsymbol{f}}{\partial x_1^{r_1} \, \partial x_2^{r_2} \cdots \partial x_n^{r_n}} \ .$$

By VIII.2.2.6., the successive partial derivatives may be taken in any order.

There is a "converse" to Clairaut's Theorem which is important in vector field theory and in differential equations (compare with IX.3.12.1.):

VIII.2.2.8. THEOREM. Let U be an open rectangle or open disk in $\mathbb{R}^2$, and g and h continuous functions from U to $\mathbb{R}$ such that $\frac{\partial g}{\partial y}$ and $\frac{\partial h}{\partial x}$ are defined, continuous, and equal everywhere on U. Then there is a $\mathcal{C}^1$ function f on U, unique up to adding a constant, with $g = \frac{\partial f}{\partial x}$ and $h = \frac{\partial f}{\partial y}$.

PROOF: For uniqueness, if f and F satisfy the conclusion, then $\phi = F - f$ is $\mathcal{C}^1$ and $D\phi = 0$, so ϕ is constant. For existence, let (a, b) be the center point of U. For $(x, y) \in U$, define

$$f_1(x, y) = \int_b^y h(a, s)\, ds + \int_a^x g(t, y)\, dt$$

$$f_2(x, y) = \int_a^x g(t, b)\, dt + \int_b^y h(x, s)\, ds$$

(note that the paths lie in U). By the Fundamental Theorem of Calculus (FTC1), $\frac{\partial f_1}{\partial x} = g$ and $\frac{\partial f_2}{\partial y} = h$. So it suffices to show $f_1 = f_2$.

Fix $(x, y) \in U$. Then, by the Fundamental Theorem of Calculus (FTC2),

$$f_1(x, y) - f_2(x, y) = \int_a^x [g(t, y) - g(t, b)]\, dt - \int_b^y [h(x, s) - h(a, s)]\, ds$$

$$= \int_a^x \int_b^y \frac{\partial g}{\partial y}(t, s)\, ds\, dt - \int_b^y \int_a^x \frac{\partial h}{\partial x}(t, s)\, dt\, ds = 0$$

by Fubini's Theorem (IX.3.4.10. or IX.3.5.2.) and the fact that $\frac{\partial g}{\partial y} = \frac{\partial h}{\partial x}$. 👍

The equality of f_1 and f_2 is a special case of Green's Theorem ().

VIII.2.2.9. The result can be extended to hold on a general simply connected open set (). But it fails in general on open sets which are not simply connected (). Using more advanced integration, the continuity hypotheses can be relaxed, but it is clumsy to make a precise statement. Of course, if g and h are $\mathcal{C}^r$, the f is $\mathcal{C}^{r+1}$ (but the uniqueness is still among $\mathcal{C}^1$ functions).

VIII.2.2.10. Note that under the hypotheses of VIII.2.2.8., the f need not be more than $\mathcal{C}^1$ since, although the second-order mixed partials of f exist and are continuous in U, $\frac{\partial^2 f}{\partial x^2} = \frac{\partial g}{\partial x}$ and $\frac{\partial^2 f}{\partial y^2} = \frac{\partial h}{\partial y}$ need not exist (anywhere!) For example, let ϕ be a continuous function from $\mathbb{R}$ to $\mathbb{R}$ which is not differentiable anywhere, and set $g(x, y) = \phi(x)$ and $h(x, y) = \phi(y)$. Then g and h satisfy the hypotheses, and $f(x, y) = \Phi(x) + \Phi(y)$, where Φ is an antiderivative for ϕ, is only $\mathcal{C}^1$.

VIII.2.3. Directional Derivatives

VIII.2.3.1. The definition of partial derivatives can be rephrased in vector notation. We write $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ for the standard basis vectors of $\mathbb{R}^n$, i.e. $\mathbf{e}_k$ has 1 in the k'th place and 0's elsewhere (the notation is a little ambiguous since n does not appear; we could write $\mathbf{e}_k^{(n)}$ if it is necessary to specify). If $n = 3$, $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ are usually written $\mathbf{i}, \mathbf{j}, \mathbf{k}$ respectively.

With this notation, if $U \subseteq \mathbb{R}^n$ and $\boldsymbol{f} : U \to \mathbb{R}^m$ is a function (we normally only consider the case $m = 1$), then we have, for $\boldsymbol{a} \in U$ and $1 \leq k \leq n$,

$$\frac{\partial \boldsymbol{f}}{\partial x_k}(\boldsymbol{a}) = \lim_{h \to 0} \frac{\boldsymbol{f}(\boldsymbol{a} + h\mathbf{e}_k) - \boldsymbol{f}(\boldsymbol{a})}{h}$$

(provided the limit exists).

We may generalize this definition (we only write the generalization in the case $m = 1$, although it can be done for general m):

VIII.2.3.2. Definition. Let $U \subseteq \mathbb{R}^n$, and $f : U \to \mathbb{R}$ a function. Let $\boldsymbol{u}$ be a nonzero vector in $\mathbb{R}^n$ (which is usually taken to be a unit vector, i.e. $\|\boldsymbol{u}\| = 1$). The *directional derivative* of f at $\boldsymbol{a} \in U$ with respect to $\boldsymbol{u}$, or in the direction $\boldsymbol{u}$ if $\boldsymbol{u}$ is a unit vector, is

$$\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a}) = \lim_{h \to 0} \frac{f(\boldsymbol{a} + h\boldsymbol{u}) - f(\boldsymbol{a})}{h}$$

(provided the limit exists).

Partial derivatives are the special case $\boldsymbol{u} = \mathbf{e}_k$ for some k (note that $\mathbf{e}_k$ is a unit vector):

$$\frac{\partial f}{\partial x_k}(\boldsymbol{a}) = \frac{\partial f}{\partial \mathbf{e}_k}(\boldsymbol{a}) \ .$$

VIII.2.3.3. There is a nice geometric interpretation of directional derivatives (and in particular partial derivatives). If $f : \mathbb{R}^n \to \mathbb{R}$, the graph of f can be regarded as a hypersurface in $\mathbb{R}^n \times \mathbb{R} \cong \mathbb{R}^{n+1}$. If $\boldsymbol{u}$ is a unit vector in $\mathbb{R}^n$, there is a vertical plane through $(\boldsymbol{a}, f(\boldsymbol{a}))$ parallel to the plane spanned by $\boldsymbol{u}$ and $\mathbf{e}_{n+1}$. The intersection of this plane with the graph is a curve. If the plane is identified with $\mathbb{R}^2$ by taking $(\boldsymbol{a}, f(\boldsymbol{a}))$ as the origin and $\boldsymbol{u}$ and $\mathbf{e}_{n+1}$ as the coordinate directions, the curve becomes a plane curve in $\mathbb{R}^2$ through the origin, and $\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a})$ is just the slope of the tangent line to this curve at the origin (Figure ()). It also has a more elementary interpretation as the rate of increase of f as one moves from $\boldsymbol{a}$ with unit speed along the line in the direction $\boldsymbol{u}$.

VIII.2.3.4. In the case where $\boldsymbol{u}$ is not a unit vector (e.g. in X.1.4.1.), let $\boldsymbol{v} = \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}$ be a unit vector in the direction of $\boldsymbol{u}$. Then we have, writing $t = \|\boldsymbol{u}\|h$,

$$\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a}) = \lim_{h \to 0} \frac{\boldsymbol{f}(\boldsymbol{a} + h\boldsymbol{u}) - \boldsymbol{f}(\boldsymbol{a})}{h} = \|\boldsymbol{u}\| \lim_{t \to 0} \frac{\boldsymbol{f}(\boldsymbol{a} + t\boldsymbol{v}) - \boldsymbol{f}(\boldsymbol{a})}{t} = \|\boldsymbol{u}\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{a}) \ .$$

VIII.2.4. Exercises

VIII.2.4.1. (a) Prove the following rules of partial differentiation, carefully stating each as a theorem. For each, U is an open set in $\mathbb{R}^n$, f and g are real-valued functions on U which are differentiable at $\boldsymbol{a} \in U$, $h : \mathbb{R} \to \mathbb{R}$ is differentiable at $f(\boldsymbol{a})$, $c \in \mathbb{R}$ (also denoting the constant function c from $\mathbb{R}^n$ to $\mathbb{R}$), and $1 \leq j, k \leq n$. π_j denotes the j'th coordinate function, i.e. $\pi_j(x_1, \ldots, x_n) = x_j$.

$$\frac{\partial(c)}{\partial x_k}(\boldsymbol{a}) = 0 \ .$$

$$\frac{\partial(\pi_j)}{\partial x_k}(\boldsymbol{a}) = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases} \ .$$

$$\frac{\partial(f+g)}{\partial x_k}(\boldsymbol{a}) = \frac{\partial f}{\partial x_k}(\boldsymbol{a}) + \frac{\partial g}{\partial x_k}(\boldsymbol{a}) \ .$$

$$\frac{\partial(cf)}{\partial x_k}(\boldsymbol{a}) = c\frac{\partial f}{\partial x_k}(\boldsymbol{a}) \ .$$

$$\frac{\partial(fg)}{\partial x_k}(\boldsymbol{a}) = \frac{\partial f}{\partial x_k}(\boldsymbol{a})g(\boldsymbol{a}) + f(\boldsymbol{a})\frac{\partial g}{\partial x_k}(\boldsymbol{a}) \ .$$

$$\frac{\partial(f/g)}{\partial x_k}(\boldsymbol{a}) = \frac{\frac{\partial f}{\partial x_k}(\boldsymbol{a})g(\boldsymbol{a}) - f(\boldsymbol{a})\frac{\partial g}{\partial x_k}(\boldsymbol{a})}{[g(\boldsymbol{a})]^2} \text{ if } g(\boldsymbol{a}) \neq 0 \ .$$

$$\frac{\partial(h \circ f)}{\partial x_k}(\boldsymbol{a}) = h'(f(\boldsymbol{a}))\frac{\partial f}{\partial x_k}(\boldsymbol{a}) \ .$$

(b) Use these rules to justify calculating partial derivatives as in VIII.2.1.6..

VIII.2.4.2. Discuss the validity of the following computations.
Let $x = r\cos\theta$, $y = r\sin\theta$, so that $r = \sqrt{x^2+y^2}$. Then $\frac{\partial x}{\partial r} = \cos\theta$, and $\frac{\partial r}{\partial x} = \frac{x}{r} = \cos\theta$. Thus

$$\frac{\partial x}{\partial r} = \frac{\partial r}{\partial x} \neq \frac{1}{\frac{\partial x}{\partial r}}$$

unless $\theta = 0$ or $\theta = \pi$ (or πn for some n).

VIII.3. Differentiability of Multivariable Functions

VIII.3.1. Differentiability and the Derivative

VIII.3.1.1. What should it mean for a function from one Euclidean space to another to be differentiable at a point, and what kind of object should the derivative be? In the one-dimensional case, f is differentiable at a if the limit

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}$$

exists; if it does, the limit is called $f'(a)$. But if $\boldsymbol{f}:\mathbb{R}^n\to\mathbb{R}^m$ and $\boldsymbol{a}\in\mathbb{R}^n$, $n>1$, the analogous expression

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}}\frac{\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\boldsymbol{f}(\boldsymbol{a})}{\boldsymbol{h}}$$

does not even make sense since division by a vector is not defined.

VIII.3.1.2. We can change the one-dimensional definition into an equivalent one which does make sense when redecorated in this way, however. The following statements are equivalent by trivial calculation:

$$\lim_{h\to 0}\frac{f(a+h)-f(a)}{h}=f'(a)$$

$$\lim_{h\to 0}\left[\frac{f(a+h)-f(a)}{h}-f'(a)\right]=0$$

$$\lim_{h\to 0}\frac{f(a+h)-f(a)-f'(a)h}{h}=0$$

$$\lim_{h\to 0}\frac{f(a+h)-f(a)-f'(a)h}{|h|}=0$$

(for the last equivalence, consider left and right limits).

VIII.3.1.3. This last expression is almost in the form needed to make a reasonable definition in the vector case. However, it may be unclear how to modify the term $f'(a)h$ in the numerator. A key to the proper approach comes from differential approximation: the best affine approximation to f near a is the function $\phi(a+h)=f(a)+df(a)(h)=f(a)+f'(a)h$, so the numerator is $f(a+h)-\phi(a+h)$, and the statement that the limit is 0 states that f can be closely approximated by this affine function.

A reasonable geometric criterion for differentiability at $\boldsymbol{a}$ in the vector case is that the function must be closely approximable near $\boldsymbol{a}$ by an affine function, i.e. there is an affine function $\phi:\mathbb{R}^n\to\mathbb{R}^m$ (if $\boldsymbol{f}:U\subseteq\mathbb{R}^n\to\mathbb{R}^m$) such that

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}}\frac{\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\phi(\boldsymbol{a}+\boldsymbol{h})}{\|\boldsymbol{h}\|}=\boldsymbol{0}$$

where the fraction is interpreted as the vector

$$\frac{1}{\|\boldsymbol{h}\|}\left[\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})-\phi(\boldsymbol{a}+\boldsymbol{h})\right]\in\mathbb{R}^m\ .$$

The affine function ϕ is of the form $\phi(\boldsymbol{x})=\boldsymbol{c}+T(\boldsymbol{x}-\boldsymbol{a})$, where $T:\mathbb{R}^n\to\mathbb{R}^m$ is a linear function and $\boldsymbol{c}$ is a constant. Since we must have $\phi(\boldsymbol{a})=\boldsymbol{f}(\boldsymbol{a})$, $\boldsymbol{c}=\boldsymbol{f}(\boldsymbol{a})$. So we conclude that the following definition is reasonable:

VIII.3.1.4. DEFINITION. Let $U \subseteq \mathbb{R}^n$ be open, and $\boldsymbol{f} : U \to \mathbb{R}^m$, $\boldsymbol{a} \in U$. Then $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$ if there is a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ for which

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - T(\boldsymbol{h})}{\|\boldsymbol{h}\|} = \boldsymbol{0} \ .$$

VIII.3.1.5. We will see that the linear transformation T, if it exists, is unique (this can be proved directly; cf. Exercise ()). This linear transformation will be called the *total differential* of $\boldsymbol{f}$ at $\boldsymbol{a}$, denoted $d\boldsymbol{f}(\boldsymbol{a})$, and its matrix the *total derivative matrix* of $\boldsymbol{f}$ at $\boldsymbol{a}$, denoted $D\boldsymbol{f}(\boldsymbol{a})$. Note that $d\boldsymbol{f}(\boldsymbol{a})$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $D\boldsymbol{f}(\boldsymbol{a})$ is an $m \times n$ matrix. $D\boldsymbol{f}(\boldsymbol{a})$ has a simple formula in terms of partial derivatives of the coordinate functions of $\boldsymbol{f}$ (VIII.3.1.10.).

VIII.3.1.6. Since a linear transformation from $\mathbb{R}$ to $\mathbb{R}$ is of the form $T(x) = cx$ for some constant c, we recover the usual one-dimensional definition as modified in VIII.3.1.2., and $df(a)$ is the linear function $df(a)(x) = f'(a)x$ and $Df(a)$ is the 1×1 matrix $[f'(a)]$. So our definitions are (up to rephrasing) consistent.

VIII.3.1.7. Note that the definition of differentiability can be rephrased to require that

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{\|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - T(\boldsymbol{h})\|}{\|\boldsymbol{h}\|} = 0$$

since a limit $\lim_{\boldsymbol{x}\to\boldsymbol{b}} \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{0}$ of a vector function $\boldsymbol{g}$ holds if and only if $\lim_{\boldsymbol{x}\to\boldsymbol{b}} \|\boldsymbol{g}(\boldsymbol{x})\| = 0$ (). It will sometimes be more convenient to use this version.

VIII.3.1.8. **Remark:** The definition applies equally well to functions between any finite-dimensional vector spaces V and W over $\mathbb{R}$ since it makes no reference to the standard coordinate system in $\mathbb{R}^n$: any norms on V and W can be used, since all such norms are equivalent (). Thus there is a well-defined notion of differentiability for functions between any finite-dimensional vector spaces over $\mathbb{R}$ (finite-dimensionality is essential here! But it does extend to functions between Banach spaces, cf. XXI.1.1.1..)

Uniqueness and a Formula

VIII.3.1.9. Suppose $\boldsymbol{f} : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\boldsymbol{a}$, with a linear transformation T which works in the definition. Fix j, $1 \leq j \leq n$. As a special case of the limit in the definition, we have

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{a}+t\mathbf{e}_j) - \boldsymbol{f}(\boldsymbol{a}) - T(t\mathbf{e}_j)}{|t|} = \boldsymbol{0} \ .$$

By considering left and right limits, we may eliminate the absolute value from the denominator:

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{a}+t\mathbf{e}_j) - \boldsymbol{f}(\boldsymbol{a}) - T(t\mathbf{e}_j)}{t} = \boldsymbol{0}$$

$$\lim_{t\to 0} \left[\frac{\boldsymbol{f}(\boldsymbol{a}+t\mathbf{e}_j) - \boldsymbol{f}(\boldsymbol{a})}{t} - T(\mathbf{e}_j) \right] = \boldsymbol{0}$$

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{a}+t\mathbf{e}_j) - \boldsymbol{f}(\boldsymbol{a})}{t} = T(\mathbf{e}_j) \ .$$

In particular, the limit on the left exists and equals the vector $T(\mathbf{e}_j)$, which is the j'th column of the matrix of T. But the left side is the definition of $\frac{\partial \boldsymbol{f}}{\partial x_j}(\boldsymbol{a})$. So we conclude:

VIII.3.1.10. PROPOSITION. Let $\boldsymbol{f} = (f_1, \ldots, f_m) : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at $\boldsymbol{a} \in U$. Then

(i) All partial derivatives $\frac{\partial f_i}{\partial x_j}(\boldsymbol{a})$ exist.

(ii) The linear transformation T in the definition of differentiability at $\boldsymbol{a}$ is unique, and the ij'th entry of $[T]$ is $\frac{\partial f_i}{\partial x_j}(\boldsymbol{a})$ for each i and j.

Thus the total differential $d\boldsymbol{f}(\boldsymbol{a})$ and total derivative $D\boldsymbol{f}(\boldsymbol{a})$ are well defined, and

$$D\boldsymbol{f}(\boldsymbol{a}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{a}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{a}) \end{bmatrix}.$$

VIII.3.1.11. The matrix

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{a}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{a}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{a}) \end{bmatrix}$$

is defined whenever all partial derivatives exist at $\boldsymbol{a}$, and existence of this matrix is a necessary condition for differentiability at $\boldsymbol{a}$. But existence of this matrix is not sufficient for differentiability (cf. VIII.3.7.1.–VIII.3.7.4.). This is a major difference between the one-variable case and the multidimensional case. The underlying reason is that one takes a limit as $\boldsymbol{h} \to \boldsymbol{0}$: in the one-dimensional case h can only approach 0 from the right or left, i.e. along the coordinate axis, but in the multidimensional case the approach can be very complicated, and what happens along the coordinate axes does not tell the whole story. Thus it is qualitatively harder for a multivariable function to be differentiable (i.e. for the limit to exist) than it is in the one-dimensional case.

We will not call this matrix the total derivative matrix or use the notation $D\boldsymbol{f}(\boldsymbol{a})$ unless $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$.

VIII.3.1.12. EXAMPLE. Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be linear. If $\boldsymbol{a}$ and $\boldsymbol{h}$ are any vectors in $\mathbb{R}^n$, we have

$$T(\boldsymbol{a} + \boldsymbol{h}) - T(\boldsymbol{a}) - T(\boldsymbol{h}) = \boldsymbol{0}$$

so, fixing $\boldsymbol{a}$, we certainly have

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{T(\boldsymbol{a} + \boldsymbol{h}) - T(\boldsymbol{a}) - T(\boldsymbol{h})}{\|\boldsymbol{h}\|} = \boldsymbol{0}$$

and thus T is differentiable at $\boldsymbol{a}$ with $dT(\boldsymbol{a}) = T$ and $DT(\boldsymbol{a}) = [T]$. So linear functions are differentiable everywhere; the differential of a linear function at any point is the function itself, and the total derivative matrix of T at any point is $[T]$ (i.e. "the derivative of a linear function is constant").

Coordinatewise Differentiability

We can examine differentiability coordinatewise in the *range* space:

VIII.3.1.13. PROPOSITION. Let $U \subseteq \mathbb{R}^n$ be open, $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$, $\boldsymbol{a} \in U$. Then $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$ if and only if each f_i is differentiable at $\boldsymbol{a}$. In this case, the total differential $d\boldsymbol{f}(\boldsymbol{a})$ has coordinate functions $df_i(\boldsymbol{a})$.

PROOF: This is an immediate consequence of the Coordinatewise Criterion for limits (): if $T = (T_1, \ldots, T_m)$ is linear, then

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - T(\boldsymbol{h})}{\|\boldsymbol{h}\|} = \boldsymbol{0}$$

if and only if

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{f_i(\boldsymbol{a} + \boldsymbol{h}) - f_i(\boldsymbol{a}) - T_i(\boldsymbol{h})}{\|\boldsymbol{h}\|} = 0$$

for $1 \leq i \leq m$. 👍

Differentiability and Continuity

Here is a simple but important observation:

VIII.3.1.14. PROPOSITION. Let $U \subseteq \mathbb{R}^n$ be open, $\boldsymbol{f} : U \to \mathbb{R}^m$, $\boldsymbol{a} \in U$. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$, then $\boldsymbol{f}$ is continuous at $\boldsymbol{a}$.

PROOF: From the assumption

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - d\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{h})}{\|\boldsymbol{h}\|} = \boldsymbol{0}$$

it follows that

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} [\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - d\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{h})] = \boldsymbol{0}$$

and, since linear functions are continuous (), $\lim_{\boldsymbol{h} \to \boldsymbol{0}} d\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{h}) = \boldsymbol{0}$, so $\lim_{\boldsymbol{h} \to \boldsymbol{0}} [\boldsymbol{f}(\boldsymbol{a} + \boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a})] = \boldsymbol{0}$ also. 👍

VIII.3.1.15. Note that this result relies on the continuity of linear functions. This is automatic on finite-dimensional normed spaces, but not infinite-dimensional ones; so continuity is an issue if differentiability of functions between Banach spaces is considered (XXI.1.1.1.).

VIII.3.2. Differentiability of Algebraic Combinations

The sum rule and product rule for ordinary differentiation have exact analogs in the vector case.

VIII.3.2.1. PROPOSITION. [SUM RULE] Let $U \subseteq \mathbb{R}^n$ be open, $\boldsymbol{f}, \boldsymbol{g} : U \to \mathbb{R}^m$, $\boldsymbol{a} \in U$. If $\boldsymbol{f}$ and $\boldsymbol{g}$ are differentiable at $\boldsymbol{a}$, then $\boldsymbol{f} + \boldsymbol{g}$ is differentiable at $\boldsymbol{a}$, and

$$D(\boldsymbol{f} + \boldsymbol{g})(\boldsymbol{a}) = D\boldsymbol{f}(\boldsymbol{a}) + D\boldsymbol{g}(\boldsymbol{a}) \ .$$

PROOF: We have

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{[\boldsymbol{f}+\boldsymbol{g}](\boldsymbol{a}+\boldsymbol{h}) - [\boldsymbol{f}+\boldsymbol{g}](\boldsymbol{a}) - [D\boldsymbol{f}(\boldsymbol{a}) + D\boldsymbol{g}(\boldsymbol{a})](\boldsymbol{h})}{\|\boldsymbol{h}\|}$$

$$= \lim_{\boldsymbol{h}\to\boldsymbol{0}} \left[\frac{\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} + \frac{\boldsymbol{g}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{g}(\boldsymbol{a}) - D\boldsymbol{g}(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} \right]$$

$$= \lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} + \lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{\boldsymbol{g}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{g}(\boldsymbol{a}) - D\boldsymbol{g}(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} = \boldsymbol{0} + \boldsymbol{0} = \boldsymbol{0} \ .$$



Since products of vectors are not defined, the product rule only makes sense for real-valued functions:

VIII.3.2.2. THEOREM. [PRODUCT RULE] Let $U \subseteq \mathbb{R}^n$ be open, $f, g : U \to \mathbb{R}$, $\boldsymbol{a} \in U$. If f and g are differentiable at $\boldsymbol{a}$, then fg is differentiable at $\boldsymbol{a}$, and

$$D(fg)(\boldsymbol{a}) = f(\boldsymbol{a})D\boldsymbol{g}(\boldsymbol{a}) + g(\boldsymbol{a})D\boldsymbol{f}(\boldsymbol{a}) \ .$$

PROOF: We have

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{f(\boldsymbol{a}+\boldsymbol{h})g(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})g(\boldsymbol{a}) - [f(\boldsymbol{a})Dg(\boldsymbol{a}) + g(\boldsymbol{a})Df(\boldsymbol{a})]\boldsymbol{h}}{\|\boldsymbol{h}\|}$$

$$= \lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{[f(\boldsymbol{a}+\boldsymbol{h})g(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}+\boldsymbol{h})g(\boldsymbol{a})] + [f(\boldsymbol{a}+\boldsymbol{h})g(\boldsymbol{a}) - f(\boldsymbol{a})g(\boldsymbol{a})] - [f(\boldsymbol{a})Dg(\boldsymbol{a}) + g(\boldsymbol{a})Df(\boldsymbol{a})]\boldsymbol{h}}{\|\boldsymbol{h}\|}$$

$$= \lim_{\boldsymbol{h}\to\boldsymbol{0}} \left[f(\boldsymbol{a}+\boldsymbol{h}) \frac{g(\boldsymbol{a}+\boldsymbol{h}) - g(\boldsymbol{a}) - Dg(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} + [f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})] \frac{Dg(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} + g(\boldsymbol{a}) \frac{f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}) - Df(\boldsymbol{a})]\boldsymbol{h}}{\|\boldsymbol{h}\|} \right]$$

$$= \left[\lim_{\boldsymbol{h}\to\boldsymbol{0}} f(\boldsymbol{a}+\boldsymbol{h}) \right] \left[\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{g(\boldsymbol{a}+\boldsymbol{h}) - g(\boldsymbol{a}) - Dg(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} \right] + \lim_{\boldsymbol{h}\to\boldsymbol{0}} \left[[f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})] \frac{Dg(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} \right]$$

$$+ g(\boldsymbol{a}) \left[\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a}) - Df(\boldsymbol{a})]\boldsymbol{h}}{\|\boldsymbol{h}\|} \right]$$

by () and () since the four limits exist: the second and fourth exist and are 0 by hypothesis, and since f is differentiable, hence continuous, at $\boldsymbol{a}$, the first limit is $f(\boldsymbol{a})$. For the third limit, note that, for any $\boldsymbol{h} \neq \boldsymbol{0}$,

$$\left| [f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})] \frac{Dg(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} \right| \leq \|Dg(\boldsymbol{a})\| |f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})|$$

and $\lim_{\boldsymbol{h}\to\boldsymbol{0}} [f(\boldsymbol{a}+\boldsymbol{h}) - f(\boldsymbol{a})] = 0$ since f is continuous at $\boldsymbol{a}$. Thus the third limit exists and is 0. So the overall limit is $f(\boldsymbol{a}) \cdot 0 + 0 + g(\boldsymbol{a}) \cdot 0 = 0$.



VIII.3.2.3. There is a version of the product rule for vector-valued functions using the dot product. The proof is essentially immediate from the two previous results; the only trick is to find the formula with suitable notation.

If $\boldsymbol{f}, \boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, then we may form $h = \boldsymbol{f} \cdot \boldsymbol{g}$; then $h : \mathbb{R}^n \to \mathbb{R}$. For $\boldsymbol{a} \in \mathbb{R}^n$, we have that $D\boldsymbol{f}(\boldsymbol{a})$ and $D\boldsymbol{g}(\boldsymbol{a})$ are $m \times n$ and $Dh(\boldsymbol{a})$ is $1 \times n$. To get the useful formula, we can regard $\boldsymbol{f}(\boldsymbol{a})$ and $\boldsymbol{g}(\boldsymbol{a})$, which are vectors in $\mathbb{R}^m$, as row vectors, i.e. $1 \times m$ matrices. Then $\boldsymbol{f}(\boldsymbol{a})D\boldsymbol{g}(\boldsymbol{a})$ and $\boldsymbol{g}(\boldsymbol{a})D\boldsymbol{f}(\boldsymbol{a})$ are defined and are $1 \times n$, the right size for $Dh(\boldsymbol{a})$. We have:

VIII.3.2.4. Proposition. [Dot Product Rule] Let $U \subseteq \mathbb{R}^n$ be open, $\boldsymbol{f}, \boldsymbol{g} : U \to \mathbb{R}^m$, $\boldsymbol{a} \in U$, $h = \boldsymbol{f} \cdot \boldsymbol{g} : U \to \mathbb{R}$. If $\boldsymbol{f}$ and $\boldsymbol{g}$ are differentiable at $\boldsymbol{a}$, then h is differentiable at $\boldsymbol{a}$, and

$$Dh(\boldsymbol{a}) = \boldsymbol{f}(\boldsymbol{a})D\boldsymbol{g}(\boldsymbol{a}) + \boldsymbol{g}(\boldsymbol{a})D\boldsymbol{f}(\boldsymbol{a}) \ .$$

Proof: If $\boldsymbol{f} = (f_1, \ldots, f_m)$ and $\boldsymbol{g} = (g_1, \ldots, g_m)$, then $h = f_1 g_1 + \cdots + f_m g_m$, so h is differentiable at $\boldsymbol{a}$ and

$$Dh(\boldsymbol{a}) = \sum_{k=1}^{m} [f_k(\boldsymbol{a})Dg_k(\boldsymbol{a}) + g_k(\boldsymbol{a})Df_k(\boldsymbol{a})] \ .$$

A simple calculation shows that this equals $\boldsymbol{f}(\boldsymbol{a})D\boldsymbol{g}(\boldsymbol{a}) + \boldsymbol{g}(\boldsymbol{a})D\boldsymbol{f}(\boldsymbol{a})$. 👍

VIII.3.3. A Criterion for Differentiability

For a function of one variable, differentiability and existence of the derivative are the same thing by definition. But for a function of more than one variable, existence of all first-order partials is not sufficient to insure differentiability, or even continuity; cf. VIII.3.7.1.–VIII.3.7.4.. However, it is rather remarkable that continuity of first-order partials is sufficient:

VIII.3.3.1. Theorem. Let U be an open subset of $\mathbb{R}^n$, and $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$. If all partial derivatives $\frac{\partial f_i}{\partial x_j}$ exist on U and are continuous at $\boldsymbol{a} \in U$, then $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$.

Proof: First note that by VIII.3.1.13. it suffices to assume $m = 1$, i.e. $f : U \to \mathbb{R}$ and $\frac{\partial f}{\partial x_j}$ exists on U and is continuous at $\boldsymbol{a}$ for each j.

We prove the result by induction on n. If $n = 1$, the statement is a tautology (and the hypotheses are even unnecessarily strong!) Now assume the statement is true for n, and let U be an open set in $\mathbb{R}^{n+1}$ and $f : U \to \mathbb{R}$ with $\frac{\partial f}{\partial x_j}$ defined on U and continuous at $\boldsymbol{c} = (\boldsymbol{a}, b) \in U$ for $1 \leq j \leq n+1$. We need to show that f is differentiable at $\boldsymbol{c}$. Set

$$V = \{\boldsymbol{x} \in \mathbb{R}^n : (\boldsymbol{x}, b) \in U\} \ .$$

Then V is an open set in $\mathbb{R}^n$ and $\boldsymbol{a} \in V$. Define $g : V \to \mathbb{R}$ by $g(\boldsymbol{x}) = f(\boldsymbol{x}, b)$. For $1 \leq j \leq n$, we have $\frac{\partial g}{\partial x_j}(\boldsymbol{x}) = \frac{\partial f}{\partial x_j}(\boldsymbol{x}, b)$, so for $1 \leq j \leq n$ $\frac{\partial g}{\partial x_k}$ is defined on V and continuous at $\boldsymbol{a}$. Thus g is differentiable at $\boldsymbol{a}$ by the inductive hypothesis, i.e.

$$\lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{g(\boldsymbol{a} + \boldsymbol{h}) - g(\boldsymbol{a}) - Dg(\boldsymbol{h})}{\|\boldsymbol{h}\|} = 0$$

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{f(\boldsymbol{a}+\boldsymbol{h},b)-f(\boldsymbol{a},b)-\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\boldsymbol{a},b)h_j}{\|\boldsymbol{h}\|}=0\ .$$

If $\boldsymbol{h}\in\mathbb{R}^n$, $k\in\mathbb{R}$, write $(\boldsymbol{h},k)\in\mathbb{R}^{n+1}$. Since $\|\boldsymbol{h}\|\le\|(\boldsymbol{h},k)\|$ for any k, we also have

$$\lim_{(\boldsymbol{h},k)\to(\boldsymbol{0},0)} \frac{f(\boldsymbol{a}+\boldsymbol{h},b)-f(\boldsymbol{a},b)-\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\boldsymbol{a},b)h_j}{\|(\boldsymbol{h},k)\|}=0\ . \tag{VIII.1}$$

If $\boldsymbol{h}$ and k are small, by the Mean Value Theorem there is a $d_{\boldsymbol{h},k}$ between b and $b+k$ such that

$$f(\boldsymbol{a}+\boldsymbol{h},b+k)-f(\boldsymbol{a}+\boldsymbol{h},b)=\frac{\partial f}{\partial x_{n+1}}(\boldsymbol{a}+\boldsymbol{h},d_{\boldsymbol{h},k})k\ .$$

Now

$$\begin{aligned}
&\lim_{(\boldsymbol{h},k)\to(\boldsymbol{0},0)} \frac{f(\boldsymbol{a}+\boldsymbol{h},b+k)-f(\boldsymbol{a},b)-\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\boldsymbol{a},b)h_j-\frac{\partial f}{\partial x_{n+1}}(\boldsymbol{a},b)k}{\|(\boldsymbol{h},k)\|}\\
&=\lim_{(\boldsymbol{h},k)\to(\boldsymbol{0},0)} \frac{f(\boldsymbol{a}+\boldsymbol{h},b)-f(\boldsymbol{a},b)+f(\boldsymbol{a}+\boldsymbol{h},b+k)-f(\boldsymbol{a}+\boldsymbol{h},b)-\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\boldsymbol{a},b)h_j-\frac{\partial f}{\partial x_{n+1}}(\boldsymbol{a},b)k}{\|(\boldsymbol{h},k)\|}\\
&=\lim_{(\boldsymbol{h},k)\to(\boldsymbol{0},0)} \left[\frac{f(\boldsymbol{a}+\boldsymbol{h},b)-f(\boldsymbol{a},b)-\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\boldsymbol{a},b)h_j}{\|(\boldsymbol{h},k)\|}+\frac{f(\boldsymbol{a}+\boldsymbol{h},b+k)-f(\boldsymbol{a}+\boldsymbol{h},b)-\frac{\partial f}{\partial x_{n+1}}(\boldsymbol{a},b)k}{\|(\boldsymbol{h},k)\|}\right]\\
&=\lim_{(\boldsymbol{h},k)\to(\boldsymbol{0},0)} \left[\frac{f(\boldsymbol{a}+\boldsymbol{h},b)-f(\boldsymbol{a},b)-\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}(\boldsymbol{a},b)h_j}{\|(\boldsymbol{h},k)\|}+\frac{\left(\frac{\partial f}{\partial x_{n+1}}(\boldsymbol{a}+\boldsymbol{h},d_{\boldsymbol{h},k})-\frac{\partial f}{\partial x_{n+1}}(\boldsymbol{a},b)\right)k}{\|(\boldsymbol{h},k)\|}\right]\ .
\end{aligned}$$

The limit of the first term is 0 by VIII.1, and the limit of the second term is also 0 by continuity of $\frac{\partial f}{\partial x_{n+1}}$ at $\boldsymbol{c}$ (using $|k|\le\|(\boldsymbol{h},k)\|$). Thus f is differentiable at $\boldsymbol{c}$.

VIII.3.3.2. Remark. Examining the proof carefully, we do not need all the partial derivatives to be continuous at $\boldsymbol{a}$: the partial with respect to one variable can be discontinuous. In fact, the proof shows that if, for some r, the restriction of $\boldsymbol{f}$ to an r-dimensional coordinate subspace is differentiable at $\boldsymbol{a}$, and the additional partial derivatives are continuous at $\boldsymbol{a}$, then $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$; cf. Exercise VIII.3.7.9..

VIII.3.3.3. The hypotheses of the theorem give *sufficient* conditions for $\boldsymbol{f}$ to be differentiable at $\boldsymbol{a}$. But these conditions are not necessary: cf. V.3.1.15., V.12.1., VIII.3.7.6.–VIII.3.7.7.. The main significance of the theorem is that the continuity conditions are often satisfied and easy to check in cases of interest, and the theorem is frequently by far the easiest way to establish differentiability.

The next special case of VIII.3.3.1. is the most important and commonly used one. We first make a definition:

VIII.3.3.4. DEFINITION. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$ a function. If $r \in \mathbb{N}$, then $\boldsymbol{f}$ is a $\mathcal{C}^r$ function if $\boldsymbol{f}$ and all its partial derivatives of order $\leq r$ exist and are continuous on U.

If $n = 1$, this definition agrees with the definition in (); and the agreement goes deeper (VIII.4.1.5.). We extend the notation to say that $\boldsymbol{f}$ is $\mathcal{C}^0$ on U if it is continuous on U, and $\mathcal{C}^\infty$ on U if it is $\mathcal{C}^r$ on U for all r, i.e. all partials of $\boldsymbol{f}$ of all orders exist and are continuous on U.

The continuity condition in the definition of $\mathcal{C}^\infty$ is necessary and not automatic (VIII.3.7.8.). ✋

VIII.3.3.5. COROLLARY. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$ a $\mathcal{C}^1$ function. Then $\boldsymbol{f}$ is differentiable at all points of U.

VIII.3.3.6. COROLLARY. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^m$. If $r \in \mathbb{N}$, suppose all partial derivatives of $\boldsymbol{f}$ of order r are continuous on U. Then $\boldsymbol{f}$ itself and all partial derivatives of $\boldsymbol{f}$ of order $< r$ are differentiable on U, hence *a fortiori* continuous on U. In particular, $\boldsymbol{f}$ is $\mathcal{C}^r$ on U.

PROOF: Working backwards, any partial derivative of $\boldsymbol{f}$ of order $r-1$ has continuous first-order partials on U, hence is differentiable (thus continuous) on U. The same is then true of the partials of $\boldsymbol{f}$ of order $r-2$, etc. 👍

VIII.3.4. The Chain Rule

VIII.3.4.1. The Chain Rule concerns differentiability and the derivative of a composition of functions. Two previous results suggest that the derivative of a composition of two functions should be the product of the derivatives:

(i) If $f, g : \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}$, then $(g \circ f)'(a) = g'(f(a))f'(a)$ (cf. V.3.4.3.).

(ii) If $S : \mathbb{R}^n \to \mathbb{R}^m$ and $T : \mathbb{R}^m \to \mathbb{R}^p$ are linear, then $T \circ S : \mathbb{R}^n \to \mathbb{R}^p$ is linear and $[T \circ S] = [T][S]$ ().

The Chain Rule of vector calculus generalizes these two facts. Note that, as always, the Chain Rule is a theorem, not a formula: it has hypotheses (that $\boldsymbol{f}$ and $\boldsymbol{g}$ are differentiable) and a conclusion with a theoretical part (that $\boldsymbol{g} \circ \boldsymbol{f}$ is differentiable) as well as a formula for the total derivative of $\boldsymbol{g} \circ \boldsymbol{f}$.

The Chain Rule is a prime example of the utility of vector calculus notation. The corresponding statement written in coordinates is unwieldy, although this form is often needed in applications.

VIII.3.4.2. THEOREM. [CHAIN RULE] Let U and V be open sets in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, $\boldsymbol{f} : U \to \mathbb{R}^m$ and $\boldsymbol{g} : V \to \mathbb{R}^p$ with $\boldsymbol{f}(U) \subseteq V$, $\boldsymbol{a} \in U$, $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a}) \in V$. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$ and $\boldsymbol{g}$ is differentiable at $\boldsymbol{b}$, then $\boldsymbol{g} \circ \boldsymbol{f}$ is differentiable at $\boldsymbol{a}$ and

$$D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{a}) = D\boldsymbol{g}(\boldsymbol{b})D\boldsymbol{f}(\boldsymbol{a}) = D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}))D\boldsymbol{f}(\boldsymbol{a}) \ .$$

Note that under the hypotheses, $D\boldsymbol{g}(\boldsymbol{b})$ is $p \times m$, $D\boldsymbol{f}(\boldsymbol{a})$ is $m \times n$, and $D(\boldsymbol{g} \circ \boldsymbol{f})(\boldsymbol{a})$ is $p \times n$, so the product of the matrices on the right is defined and is the same size as the matrix on the left (this is of course not a proof of the theorem, only an indication that the statement makes sense!)

It is easily seen that the two examples in VIII.3.4.1. are special cases of this result (cf. VIII.3.1.12.).

PROOF: The proof is largely a matter of choosing good notation, which can be a little complicated to keep straight.

For $\boldsymbol{h}$ in a neighborhood of $\mathbf{0}$ in $\mathbb{R}^n$, set

$$\phi(\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}$$

so we have

$$\lim_{\boldsymbol{h}\to\mathbf{0}} \frac{\phi(\boldsymbol{h})}{\|\boldsymbol{h}\|} = \lim_{\boldsymbol{h}\to\mathbf{0}} \frac{\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) - D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{h})}{\|\boldsymbol{h}\|} = \mathbf{0} \ .$$

Similarly, for $\boldsymbol{k}$ in a neighborhood of $\mathbf{0}$ in $\mathbb{R}^m$, set

$$\psi(\boldsymbol{k}) = \boldsymbol{g}(\boldsymbol{b}+\boldsymbol{k}) - \boldsymbol{g}(\boldsymbol{b}) - D\boldsymbol{g}(\boldsymbol{b})\boldsymbol{k}$$

so we have

$$\lim_{\boldsymbol{k}\to\mathbf{0}} \frac{\psi(\boldsymbol{k})}{\|\boldsymbol{k}\|} = \lim_{\boldsymbol{k}\to\mathbf{0}} \frac{\boldsymbol{g}(\boldsymbol{b}+\boldsymbol{k}) - \boldsymbol{g}(\boldsymbol{b}) - D\boldsymbol{g}(\boldsymbol{b})(\boldsymbol{k})}{\|\boldsymbol{h}\|} = \mathbf{0} \ .$$

Set $\omega(\boldsymbol{k}) = \frac{\psi(\boldsymbol{k})}{\|\boldsymbol{k}\|}$ for $\boldsymbol{k} \neq \mathbf{0}$ and $\omega(\mathbf{0}) = \mathbf{0}$. Then ω is continuous at $\mathbf{0}$.

If $\boldsymbol{h}$ is sufficiently close to $\mathbf{0}$, then $\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})$ is defined; and if we set

$$\boldsymbol{k} = \boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a}) = D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h} + \phi(\boldsymbol{h})$$

we have that $\boldsymbol{k} \to \mathbf{0}$ as $\boldsymbol{h} \to \mathbf{0}$ since $\boldsymbol{f}$ is continuous at $\boldsymbol{a}$, and

$$\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})) - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a})) = \boldsymbol{g}(\boldsymbol{b}+\boldsymbol{k}) - \boldsymbol{g}(\boldsymbol{b}) = D\boldsymbol{g}(\boldsymbol{b})\boldsymbol{k} + \psi(\boldsymbol{k})$$

$$= D\boldsymbol{g}(\boldsymbol{b})[D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h} + \phi(\boldsymbol{h})] + \psi(\boldsymbol{k}) = D\boldsymbol{g}(\boldsymbol{b})D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h} + D\boldsymbol{g}(\boldsymbol{b})\phi(\boldsymbol{h}) + \psi(\boldsymbol{k})$$

so we have

$$\frac{\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})) - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a})) - D\boldsymbol{g}(\boldsymbol{b})D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} = \frac{D\boldsymbol{g}(\boldsymbol{b})\phi(\boldsymbol{h})}{\|\boldsymbol{h}\|} + \frac{\psi(\boldsymbol{k})}{\|\boldsymbol{h}\|} \ .$$

Since $\|D\boldsymbol{g}(\boldsymbol{b})\phi(\boldsymbol{h})\| \leq \|D\boldsymbol{g}(\boldsymbol{b})\|\|\phi(\boldsymbol{h})\|$ and $\lim_{\boldsymbol{h}\to\mathbf{0}} \frac{\phi(\boldsymbol{h})}{\|\boldsymbol{h}\|} = \mathbf{0}$, the first term has limit $\mathbf{0}$. We show the second term also has limit $\mathbf{0}$. We have

$$\|\boldsymbol{k}\| = \|\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{a})\| = \|D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h} + \phi(\boldsymbol{h})\| \leq \|D\boldsymbol{f}(\boldsymbol{a})\|\|\boldsymbol{h}\| + \|\phi(\boldsymbol{h})\|$$

$$\frac{\|\boldsymbol{k}\|}{\|\boldsymbol{h}\|} \leq \|D\boldsymbol{f}(\boldsymbol{a})\| + \frac{\|\phi(\boldsymbol{h})\|}{\|\boldsymbol{h}\|}$$

and since the second term goes to 0 as $\boldsymbol{h} \to \mathbf{0}$, it is bounded in a neighborhood of $\mathbf{0}$. Thus

$$\lim_{\boldsymbol{h}\to\mathbf{0}} \frac{\psi(\boldsymbol{k})}{\|\boldsymbol{h}\|} = \lim_{\boldsymbol{h}\to\mathbf{0}} \frac{\|\boldsymbol{k}\|}{\|\boldsymbol{h}\|}\omega(\boldsymbol{k}) = \mathbf{0}$$

since the first factor is bounded and the second factor tends to $\mathbf{0}$. Thus

$$\lim_{\boldsymbol{h}\to\mathbf{0}} \frac{\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a}+\boldsymbol{h})) - \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{a})) - D\boldsymbol{g}(\boldsymbol{b})D\boldsymbol{f}(\boldsymbol{a})\boldsymbol{h}}{\|\boldsymbol{h}\|} = \mathbf{0} \ .$$



The Chain Rule in Coordinates

Writing the Chain Rule in coordinate form is more complicated and less clean than the vector-matrix form, but is usually necessary in practice. Rather than writing the general version, we will write it out in the typical notation and careful, correct form in the case $n = 3$, $m = 4$, $p = 2$.

VIII.3.4.3. We consider the case where u and v are functions of x, y, z, w, which are themselves functions of r, s, t. Specifically,

$$u = g_1(x, y, z, w)$$
$$v = g_2(x, y, z, w)$$
$$x = f_1(r, s, t)$$
$$y = f_2(r, s, t)$$
$$z = f_3(r, s, t)$$
$$w = f_4(r, s, t)$$

So, in vector notation, we have

$$(u, v) = \boldsymbol{g}(x, y, z, w)$$
$$(x, y, z, w) = \boldsymbol{f}(r, s, t)$$

where $\boldsymbol{f} = (f_1, f_2, f_3, f_4) : \mathbb{R}^3 \to \mathbb{R}^4$ and $\boldsymbol{g} = (g_1, g_2) : \mathbb{R}^4 \to \mathbb{R}^2$. So indirectly we have

$$u = h_1(r, s, t) = g_1(f_1(r, s, t), f_2(r, s, t), f_3(r, s, t), f_4(r, s, t))$$
$$v = h_2(r, s, t) = g_2(f_1(r, s, t), f_2(r, s, t), f_3(r, s, t), f_4(r, s, t))$$

where $\boldsymbol{h} = (h_1, h_2) = \boldsymbol{g} \circ \boldsymbol{f} : \mathbb{R}^3 \to \mathbb{R}^2$.

VIII.3.4.4. We suppose $\boldsymbol{f}$ is defined in a neighborhood of $\boldsymbol{a} = (r_0, s_0, t_0)$ and differentiable at (r_0, s_0, t_0), and that $\boldsymbol{g}$ is defined in a neighborhood of $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a}) = (x_0, y_0, z_0, w_0)$ and differentiable at (x_0, y_0, z_0, w_0), where

$$x_0 = f_1(r_0, s_0, t_0)$$
$$y_0 = f_2(r_0, s_0, t_0)$$
$$z_0 = f_3(r_0, s_0, t_0)$$
$$w_0 = f_4(r_0, s_0, t_0)$$

(note that the subscript 0 denotes a particular point, while the subscripts 1, 2, 3, 4 denote coordinate functions; we use the subscript 0 instead of distinct letters so we don't run out of letters). To avoid impossibly complicated formulas which will not fit on the page, we write as abbreviations

$$\frac{\partial u}{\partial x} := \frac{\partial g_1}{\partial x}(x_0, y_0, z_0, w_0) = \frac{\partial g_1}{\partial x}(f_1(r_0, s_0, t_0), f_2(r_0, s_0, t_0), f_3(r_0, s_0, t_0), f_4(r_0, s_0, t_0))$$
$$\frac{\partial v}{\partial z} := \frac{\partial g_2}{\partial z}(x_0, y_0, z_0, w_0) = \frac{\partial g_2}{\partial z}(f_1(r_0, s_0, t_0), f_2(r_0, s_0, t_0), f_3(r_0, s_0, t_0), f_4(r_0, s_0, t_0))$$

and similarly for $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, etc. Thus $\frac{\partial u}{\partial x}$, etc., denote particular numbers.

Similarly, write

$$\frac{\partial x}{\partial r} := \frac{\partial f_1}{\partial r}(r_0, s_0, t_0)$$
$$\frac{\partial w}{\partial s} := \frac{\partial f_4}{\partial s}(r_0, s_0, t_0)$$

and similarly for $\frac{\partial z}{\partial r}$, etc.

Finally, write

$$\frac{\partial u}{\partial r} := \frac{\partial h_1}{\partial r}(r_0, s_0, t_0)$$
$$\frac{\partial v}{\partial t} := \frac{\partial h_2}{\partial t}(r_0, s_0, t_0)$$

and similarly for $\frac{\partial u}{\partial s}$, etc.

VIII.3.4.5. The Chain Rule then says that $\boldsymbol{h}$ is differentiable at (r_0, s_0, t_0), and

$$D\boldsymbol{h}(r_0, s_0, t_0) = D\boldsymbol{g}(x_0, y_0, z_0, w_0) D\boldsymbol{f}(r_0, s_0, t_0)$$

$$= \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} & \frac{\partial u}{\partial z} & \frac{\partial u}{\partial w} \\ \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} & \frac{\partial v}{\partial z} & \frac{\partial v}{\partial w} \end{bmatrix} \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial s} & \frac{\partial x}{\partial t} \\ \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial s} & \frac{\partial y}{\partial t} \\ \\ \frac{\partial z}{\partial r} & \frac{\partial z}{\partial s} & \frac{\partial z}{\partial t} \\ \\ \frac{\partial w}{\partial r} & \frac{\partial w}{\partial s} & \frac{\partial w}{\partial t} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial u}{\partial x}\frac{\partial x}{\partial r} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial r} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial r} + \frac{\partial u}{\partial w}\frac{\partial w}{\partial r} & \cdots \\ \vdots & \ddots \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial u}{\partial r} & \frac{\partial u}{\partial s} & \frac{\partial u}{\partial t} \\ \\ \frac{\partial v}{\partial r} & \frac{\partial v}{\partial s} & \frac{\partial v}{\partial t} \end{bmatrix}$$

so we obtain formulas like

$$\frac{\partial v}{\partial s} = \frac{\partial v}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial v}{\partial y}\frac{\partial y}{\partial s} + \frac{\partial v}{\partial z}\frac{\partial z}{\partial s} + \frac{\partial v}{\partial w}\frac{\partial w}{\partial s}$$

and similarly for the other five entries. The masochistic reader should write out the careful version of this formula replacing the symbols $\frac{\partial u}{\partial x}$, etc., by their proper definitions, and all readers should note that there is a fundamental difference between expressions like $\frac{\partial u}{\partial r}$ and expressions like $\frac{\partial u}{\partial x}$ which is obscured by the abbreviated notation (cf. VIII.2.1.4.).

VIII.3.5. The Gradient and Tangent Planes

For a real-valued differentiable function f on a region in $\mathbb{R}^n$, there is an alternate representation or interpretation of the total derivative Df which has a nice geometric interpretation.

VIII.3.5.1. If $U \subseteq \mathbb{R}^n$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ is differentiable at $\boldsymbol{a}$, then $Df(\boldsymbol{a})$ is a $1 \times n$ matrix:

$$Df(\boldsymbol{a}) = \left[\begin{array}{cccc} \frac{\partial f}{\partial x_1}(\boldsymbol{a}) & \frac{\partial f}{\partial x_2}(\boldsymbol{a}) & \cdots & \frac{\partial f}{\partial x_n}(\boldsymbol{a}) \end{array} \right] .$$

A $1 \times n$ matrix can be interpreted as a row vector in $\mathbb{R}^n$, and thus $Df(\boldsymbol{a})$ can be alternately interpreted as a vector in $\mathbb{R}^n$. When it is, it is given a different name to emphasize the change of point of view:

VIII.3.5.2. Definition. Let $U \subseteq \mathbb{R}^n$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ a real-valued function. If f is differentiable at $\boldsymbol{a}$, the *gradient* of f at $\boldsymbol{a}$ is the vector $\nabla f(\boldsymbol{a})$ in $\mathbb{R}^n$ defined by

$$\nabla f(\boldsymbol{a}) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{a}), \frac{\partial f}{\partial x_2}(\boldsymbol{a}), \cdots, \frac{\partial f}{\partial x_n}(\boldsymbol{a}) \right) .$$

Note that, as usual, ∇f is regarded as a single symbol. Some references use the alternate notation $Grad(f)$.

VIII.3.5.3. Just as the total derivative matrix of a function at $\boldsymbol{a}$ is technically defined when all first-order partials exist at $\boldsymbol{a}$, even if the function is not differentiable at $\boldsymbol{a}$, the formula for the gradient makes sense if all first-order partials are just defined at $\boldsymbol{a}$. But just as we only use the notation $D\boldsymbol{f}(\boldsymbol{a})$ if $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$, we only use the notation $\nabla f(\boldsymbol{a})$ and say the gradient exists at $\boldsymbol{a}$ if f is actually differentiable at $\boldsymbol{a}$ (e.g. if all first-order partials of f are continuous at $\boldsymbol{a}$).



Directional Derivatives Revisited

If $f : U \to \mathbb{R}$ is differentiable at $\boldsymbol{a} \in U \subseteq \mathbb{R}^n$, then the directional derivatives can be easily calculated using the gradient. Let $\boldsymbol{u}$ be a nonzero vector in $\mathbb{R}^n$, and define $\gamma : \mathbb{R} \to \mathbb{R}^n$ by

$$\gamma(t) = \boldsymbol{a} + t\boldsymbol{u} .$$

Then $\gamma'(0) = \boldsymbol{u}$, or in matrix terms $D\gamma(0)$ is the $n \times 1$ matrix $\boldsymbol{u}$, regarded as a column vector. Set $\phi(t) = f(\gamma(t))$. If U is open, then ϕ is defined in an interval around 0, and by definition we have

$$\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a}) = \phi'(0)$$

and by the Chain Rule we have

$$\phi'(0) = D\phi(0) = Df(\gamma(0))\gamma'(0) = Df(\boldsymbol{a})\boldsymbol{u}$$

where $\boldsymbol{u}$ is regarded as an $n \times 1$ matrix (column vector). However, observe that the scalar $Df(\boldsymbol{a})\boldsymbol{u}$ is exactly the dot product $\nabla f(\boldsymbol{a}) \cdot \boldsymbol{u}$ (). Thus we conclude:

VIII.3.5.4. PROPOSITION. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ a real-valued function. If f is differentiable at $\boldsymbol{a}$, then for any nonzero vector $\boldsymbol{u} \in \mathbb{R}^n$, the directional derivative of f at $\boldsymbol{a}$ with respect to $\boldsymbol{u}$ is defined and

$$\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a}) = \nabla f(\boldsymbol{a}) \cdot \boldsymbol{u} \ .$$

VIII.3.5.5. This simple relationship does not necessarily hold if f is not differentiable at $\boldsymbol{a}$, even if all directional derivatives (and in particular all first-order partials) of f exist at $\boldsymbol{a}$. For example, if $f : \mathbb{R}^2 \to \mathbb{R}$ is the function of VIII.3.7.5., then $\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{0})$ exists for all $\boldsymbol{u}$. But $\frac{\partial f}{\partial x}(\boldsymbol{0}) = \frac{\partial f}{\partial y}(\boldsymbol{0}) = 0$, so "$\nabla f(\boldsymbol{0})$" is $\boldsymbol{0}$, although $\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{0}) \neq 0$ for some $\boldsymbol{u}$'s. This shows again that f cannot be differentiable at $\boldsymbol{0}$ (and, in particular, $\nabla f(\boldsymbol{0})$ is actually undefined).

Geometric Interpretation of the Gradient

Returning to the case where f is differentiable at $\boldsymbol{a}$, we can interpret VIII.3.5.4. geometrically:

VIII.3.5.6. COROLLARY. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ a real-valued function. If f is differentiable at $\boldsymbol{a}$ and $\nabla f(\boldsymbol{a}) \neq \boldsymbol{0}$, then for any unit vector $\boldsymbol{u} \in \mathbb{R}^n$, the directional derivative of f at $\boldsymbol{a}$ in the direction $\boldsymbol{u}$ is defined and

$$\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a}) = \|\nabla f(\boldsymbol{a})\| \cos\theta$$

where θ is the angle between $\nabla f(\boldsymbol{a})$ and $\boldsymbol{u}$.

The conclusion also holds if $\nabla f(\boldsymbol{a}) = \boldsymbol{0}$ (although θ is not defined): then $\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a}) = 0$ for any $\boldsymbol{u}$.

VIII.3.5.7. Thus $\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a})$ is maximum when $\theta = 0$, i.e. if and only if $\boldsymbol{u}$ is a unit vector in the direction of $\nabla f(\boldsymbol{a})$ (in the case $\nabla f(\boldsymbol{a}) \neq \boldsymbol{0}$). So if $\nabla f(\boldsymbol{a}) \neq \boldsymbol{0}$, it points in the direction of greatest increase of f at $\boldsymbol{a}$, and the length of the gradient gives the maximum rate of increase. In the case $n = 2$, the graph of f can be regarded as a surface in $\mathbb{R}^3$, and at any point of the surface where the function is differentiable with nonzero gradient the gradient points "straight uphill" (see Figure ()).

The Tangent Hyperplane

VIII.3.5.8. If $\nabla f(\boldsymbol{a}) \neq 0$, then for a unit vector $\boldsymbol{u} \in \mathbb{R}^n$ we have $\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a}) = 0$ if and only if $\boldsymbol{u} \perp \nabla f(\boldsymbol{a})$. The set of all such vectors spans an $(n-1)$-dimensional subspace (hyperplane) of $\mathbb{R}^n$. The hyperplane Through $\boldsymbol{a}$ parallel to this is the obvious candidate for the tangent hyperplane to the hypersurface

$$S = \{\boldsymbol{x} : f(\boldsymbol{x}) = c\}$$

where $c = f(\boldsymbol{a})$. (It may not be obvious that this set is a hypersurface in the usual sense, but it follows from the Implicit Function Theorem.)

VIII.3.6. Functional Independence

VIII.3.6.1. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ a $\mathcal{C}^1$ function. We say $\{f_1, \ldots, f_m\}$ is *functionally dependent* on U if there is a $\mathcal{C}^1$ function g on an open set V in $\mathbb{R}^{n-1}$ such that, for some k, $1 \leq k \leq m$, we have

$$(f_1(\boldsymbol{x}), \ldots, f_{k-1}(\boldsymbol{x}), f_{k+1}(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})) \in V$$

$$f_k(\boldsymbol{x}) = g(f_1(\boldsymbol{x}), \ldots, f_{k-1}(\boldsymbol{x}), f_{k+1}(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$$

for all $\boldsymbol{x} \in U$. Otherwise $\{f_1, \ldots, f_m\}$ is *functionally independent* on U. (The $\mathcal{C}^1$ requirement can be relaxed in various ways, but this is the most useful form of the definition.)

Geometrically, functional dependence says that $\boldsymbol{f}(\boldsymbol{x})$ lies on the graph of g in $\mathbb{R}^m \cong \mathbb{R}^{m-1} \times \mathbb{R}$ for all $\boldsymbol{x} \in U$. Note that functional independence implies, but is much stronger than, linear independence (Exercise ()).

VIII.3.6.2. An apparently more general functional dependence is to just require that there is a function F, $\mathcal{C}^1$ on an open set in $\mathbb{R}^m$ containing $\boldsymbol{f}(U)$ with nonvanishing derivative, such that $F(\boldsymbol{f}(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in U$. But this is actually not really more general, at least locally, by the Implicit Function Theorem.

VIII.3.6.3. In general, it is a difficult problem to determine whether a set of functions is functionally dependent. We will obtain from the Chain Rule a simple necessary condition in the case $n \geq m$, although this condition is by no means sufficient.

VIII.3.6.4. Proposition. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ a $\mathcal{C}^1$ function, with $n \geq m$. If $\{f_1, \ldots, f_m\}$ is functionally dependent, then for every $\{j_1, \ldots, j_m\} \subseteq \{1, \ldots, n\}$ and every $\boldsymbol{x} \in U$ we have

$$det \begin{bmatrix} \frac{\partial f_1}{\partial x_{j_1}}(\boldsymbol{x}) & \cdots & \frac{\partial f_1}{\partial x_{j_m}}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_{j_1}}(\boldsymbol{x}) & \cdots & \frac{\partial f_m}{\partial x_{j_m}}(\boldsymbol{x}) \end{bmatrix} = 0 \ .$$

Proof: By reordering $f_1, \ldots, f_m$, we may assume the k in the definition of functional dependence is m, i.e. there is a $\mathcal{C}^1$ function g such that $f_m(\boldsymbol{x}) = g(f_1(\boldsymbol{x}), \ldots, f_{m-1}(\boldsymbol{x}))$ for all $\boldsymbol{x} \in U$, to simplify notation. Fix $\boldsymbol{x} \in U$, and write $\boldsymbol{y} = (f_1(\boldsymbol{x}), \ldots, f_{m-1}(\boldsymbol{x})) \in \mathbb{R}^{m-1}$. Then by the Chain Rule, for each j we have

$$\frac{\partial f_m}{\partial x_j} = \sum_{k=1}^{m-1} \frac{\partial g}{\partial y_k}(\boldsymbol{y}) \frac{\partial f_k}{\partial x_j}(\boldsymbol{x})$$

so the last row in the matrix is a linear combination of the first $m-1$ rows (the k'th row is multiplied by $\frac{\partial g}{\partial y_k}(\boldsymbol{y})$), hence the matrix has zero determinant. 👍

VIII.3.6.5. Corollary. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$ a $\mathcal{C}^1$ function, with $n \geq m$. If there is an $\boldsymbol{a} \in U$ and $\{j_1, \ldots, j_m\} \subseteq \{1, \ldots, n\}$ such that

$$det \begin{bmatrix} \frac{\partial f_1}{\partial x_{j_1}}(\boldsymbol{a}) & \cdots & \frac{\partial f_1}{\partial x_{j_m}}(\boldsymbol{a}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_{j_1}}(\boldsymbol{a}) & \cdots & \frac{\partial f_m}{\partial x_{j_m}}(\boldsymbol{a}) \end{bmatrix} \neq 0$$

then $\{f_1, \ldots, f_m\}$ is functionally independent.

VIII.3.6.6. In fact, under the hypotheses of VIII.3.6.5., a stronger consequence holds: the Inverse Function Theorem (applied to a suitable m-dimensional affine subspace of $\mathbb{R}^n$) implies that the range of $\boldsymbol{f}$ has nonempty interior (which clearly implies functional independence).

VIII.3.6.7. The converse of VIII.3.6.4. does not hold, and there are no reasonable sufficient conditions for functional dependence along these lines. If the conclusion of VIII.3.6.4. holds, or if $n < m$, then the Straightening Theorem says that generically $\{f_1, \ldots, f_m\}$ is *locally* functionally dependent.

VIII.3.7. Exercises

VIII.3.7.1. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases} .$$

(a) Show that f is not continuous at $\mathbf{0}$.
(b) Show that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist everywhere.
(c) Show that $D_{\boldsymbol{u}} f(0,0)$ does not exist unless $\boldsymbol{u} = \boldsymbol{e}_1$ or $\boldsymbol{u} = \boldsymbol{e}_2$.

VIII.3.7.2. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} \frac{xy}{\sqrt{x^2+y^2}} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases} .$$

(a) Show that f is continuous at $\mathbf{0}$.
(b) Show that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist everywhere.
(c) Show that $D_{\boldsymbol{u}} f(0,0)$ does not exist unless $\boldsymbol{u} = \boldsymbol{e}_1$ or $\boldsymbol{u} = \boldsymbol{e}_2$.
(d) Show that f is not differentiable at $\mathbf{0}$. [Use VIII.3.5.4..]

VIII.3.7.3. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^4+y^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases} .$$

(a) Show that f is not continuous at $\mathbf{0}$. [Consider $\lim_{t\to 0} f(t,t^2)$.] Then f is *a fortiori* not differentiable at $\mathbf{0}$.
(b) Show that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist everywhere.
(c) Show that $D_{\boldsymbol{u}} f(x,y)$ exists for all $\boldsymbol{u}$ and for all (x,y) (and in particular for $(x,y) = (0,0)$).

VIII.3.7.4. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = (x^{1/3} + y^{1/3})^3 .$$

(a) Show that f is continuous everywhere, and in particular at $\mathbf{0}$.
(b) Show that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist everywhere.
(c) Show that $D_{\boldsymbol{u}} f(0,0)$ exists and equals $f(\boldsymbol{u})$ for all $\boldsymbol{u}$.
(d) Show that f is not differentiable at $\mathbf{0}$. [Use VIII.3.5.4..]

VIII.3.7.5. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} \frac{x^2 y}{x^2+y^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases} .$$

(a) Show that f is continuous everywhere, and in particular at $\mathbf{0}$.

(b) Show that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ exist everywhere.

(c) Show that $\frac{\partial f}{\partial \boldsymbol{u}}(0,0)$ exists for all $\boldsymbol{u}$.

(d) Show that f is not differentiable at $\mathbf{0}$.

VIII.3.7.6. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} x^2+y^2 & \text{if} \quad x^2+y^2 \text{ is rational} \\ 0 & \text{if} \quad x^2+y^2 \text{ is irrational} \end{cases} .$$

(a) Show that f is continuous only at $\mathbf{0}$.

(b) Show that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ only exist at $\mathbf{0}$.

(c) Show that f is differentiable at $\mathbf{0}$.

VIII.3.7.7. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} (x^2+y^2)\sin\frac{1}{x^2+y^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases} .$$

(a) Show that f is differentiable everywhere, and in particular at $\mathbf{0}$.

(b) Show that $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are not continuous at $\mathbf{0}$.

VIII.3.7.8. Let g be a $\mathcal{C}^\infty$ bump function on $[0, 2\pi]$ (V.17.2.4.) with $g(\pi/4) = 1$ and g identically 0 outside the interval $[\pi/8, 3\pi/8]$. Define $f : \mathbb{R}^2 \setminus \{\mathbf{0}\} \to \mathbb{R}$ using polar coordinates by

$$f(r,\theta) = g(\theta) .$$

Then f is constant on open rays from the origin. Extend f to $\mathbb{R}^2$ by setting $f(\mathbf{0}) = 0$.

(a) Show that f is $\mathcal{C}^\infty$ on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$.

(b) If $x \neq 0$, then f is identically 0 in a neighborhood of $(x, 0)$, and thus all partials of f of all orders are also identically 0 on this neighborhood. The same holds for points of the form $(0, y)$ for $y \neq 0$. In particular, all partials of f of all orders are identically zero on the nonzero axes.

(c) Show that all partials of f of all orders exist and are 0 at $\mathbf{0}$.

(d) Show that f is not continuous, let alone differentiable, at $\mathbf{0}$.

VIII.3.7.9. Modify the proof of Theorem VIII.3.3.1. to give the following precise version of Remark VIII.3.3.2.:

THEOREM. Let U be an open subset of $\mathbb{R}^n$, and $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{R}^m$. Suppose that, for some r, $1 \leq r < n$, and $\boldsymbol{c} = (\boldsymbol{a}, \boldsymbol{b}) \in U$, where $\boldsymbol{a} \in \mathbb{R}^r$ and $\boldsymbol{b} \in \mathbb{R}^{n-r}$, we have that

(1) The function $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{b})$ is differentiable at $\boldsymbol{a}$.

(2) All partial derivatives $\frac{\partial f_i}{\partial x_j}$, $1 \leq i \leq n$, $r+1 \leq j \leq n$, exist on U and are continuous at $\boldsymbol{c}$.

Then $\boldsymbol{f}$ is differentiable at $\boldsymbol{c}$.

VIII.3.7.10. Derive the Product Rule (V.3.3.3.) for functions of one variable from the multivariable Chain Rule as follows. Let f and g be defined on an open interval I in $\mathbb{R}$ and differentiable at $a \in I$, and set $h = fg$.

(a) Let $\gamma : I \to \mathbb{R}^2$ be defined by $\gamma(t) = (f(t), g(t))$. Then γ is differentiable at a and

$$D\gamma(a) = \begin{bmatrix} f'(a) \\ g'(a) \end{bmatrix} \text{ (or } \gamma'(a) = (f'(a), g'(a)) \text{).}$$

(b) Define $\phi : \mathbb{R}^2 \to \mathbb{R}$ by $\phi(x, y) = xy$. Then ϕ is differentiable everywhere and

$$D\phi(x, y) = \begin{bmatrix} y & x \end{bmatrix} \text{ (or } \nabla\phi(x, y) = (y, x) \text{).}$$

(c) Apply the Chain Rule to $h = \phi \circ \gamma$ to obtain that h is differentiable at a and the Product Rule formula.

This is a rather perverse and roundabout way to obtain the Product Rule, although it is logically sound since nowhere in the derivation or proof of the multivariable Chain Rule is the Product Rule used. However, it is instructive to see that from the right point of view, the Product Rule is just a consequence of matrix multiplication or the dot product. This helps demonstrate the unity and consistency of the operations of algebra and calculus.

VIII.4. Higher-Order Derivatives and Critical Point Analysis

VIII.4.1. Higher-Order Differentiability

VIII.4.1.1. If U is an open set in $\mathbb{R}^n$ and $\boldsymbol{f} : U \to \mathbb{R}^m$ is differentiable on U (i.e. at every point of U), the total derivative $d\boldsymbol{f}$ is a function from U to the set $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$. $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ is a vector space over $\mathbb{R}$ of dimension nm and can thus be identified with $\mathbb{R}^{nm}$ in a natural way (Exercise VIII.4.6.1.). Alternatively, $D\boldsymbol{f}$ is a function from U to the $m \times n$ matrices $\mathbb{M}_{m\times n}$, which can evidently be identified with $\mathbb{R}^{nm}$. Thus $d\boldsymbol{f}$ or, equivalently, $D\boldsymbol{f}$, may be regarded as a function from U to $\mathbb{R}^{nm}$.

VIII.4.1.2. Definition. If U is an open set in $\mathbb{R}^n$ and $\boldsymbol{f} : U \to \mathbb{R}^m$ is differentiable on U, then $\boldsymbol{f}$ is *second-order differentiable* at $\boldsymbol{a} \in U$ if $d\boldsymbol{f}$ (or, equivalently, $D\boldsymbol{f}$) is differentiable at $\boldsymbol{a}$. Write $d^2\boldsymbol{f}(\boldsymbol{a})$ for $d(d\boldsymbol{f})(\boldsymbol{a})$ and $D^2\boldsymbol{f}(\boldsymbol{a})$ for $D(D\boldsymbol{f})(\boldsymbol{a})$.

VIII.4.1.3. We can go on. If $d\boldsymbol{f}$ ($D\boldsymbol{f}$) is differentiable at $\boldsymbol{a}$, then $D^2\boldsymbol{f}(\boldsymbol{a})$ is an $nm\times n$ matrix, identifiable with $\mathbb{R}^{n^2m}$. Thus if $\boldsymbol{f}$ is second-order differentiable on U and $D^2\boldsymbol{f}$ is differentiable at $\boldsymbol{a} \in U$, then $\boldsymbol{f}$ is third-order differentiable at $\boldsymbol{a}$. Continuing, we can define what it means for $\boldsymbol{f}$ to be r'th order differentiable at a point or on the set U, for any $r \in \mathbb{N}$. We also say $\boldsymbol{f}$ is *infinitely differentiable* on U if it is r'th order differentiable on U for every r. We can also write $d^r\boldsymbol{f}(\boldsymbol{a})$ and $D^r\boldsymbol{f}(\boldsymbol{a})$ if $\boldsymbol{f}$ is differentiable to order r at $\boldsymbol{a}$; $D^r\boldsymbol{f}(\boldsymbol{a})$ is an $n^{r-1}m \times n$ matrix, which can be identified with a point in $\mathbb{R}^{n^rm}$.

The coordinate functions of $D^r\boldsymbol{f}$ are just the r'th order partials of the coordinate functions of $\boldsymbol{f}$ (Exercise VIII.4.6.1.). By VIII.3.1.13., we have:

VIII.4.1.4. Proposition. If U is an open set in $\mathbb{R}^n$ and $\boldsymbol{f} : U \to \mathbb{R}^m$, then $\boldsymbol{f}$ is r'th order differentiable at $\boldsymbol{a} \in U$ if and only if each partial derivative of $\boldsymbol{f}$ of order $r-1$ is defined in a neighborhood of $\boldsymbol{a}$ and is differentiable at $\boldsymbol{a}$.

We thus have the following immediate corollary of VIII.3.1.14. and VIII.3.3.1.:

VIII.4.1.5. Corollary. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$, and $r \in \mathbb{N}$. If $\boldsymbol{f}$ is r'th order differentiable on U, then $\boldsymbol{f}$ is $\mathcal{C}^{r-1}$ on U. If $\boldsymbol{f}$ is $\mathcal{C}^r$ on U, then it is r'th order differentiable on U. The function $\boldsymbol{f}$ is $\mathcal{C}^\infty$ on U if and only if it is infinitely differentiable on U.

Note that an r'th order differentiable function does not have to be $\mathcal{C}^r$ (VIII.3.3.3.).

VIII.4.1.6. There is a little more structure to higher derivatives: $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ has more structure than $\mathbb{R}^{nm}$. If $\boldsymbol{f}$ is second-order differentiable at $\boldsymbol{a}$, then $d^2\boldsymbol{f}(\boldsymbol{a})$ is actually a linear transformation from $\mathbb{R}^n$ to $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$, which can be thought of as a *bilinear* () function from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^m$. Similarly, for any r, $d^r\boldsymbol{f}(\boldsymbol{a})$ can be regarded as an r-multilinear function from $\mathbb{R}^n \times \cdots \times \mathbb{R}^n$ to $\mathbb{R}^m$. This point of view is important in formulating the multivariable version of Taylor's Theorem (VIII.4.3.5.), for example. (The two points of view are actually equivalent, since bilinear functions from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}^m$ correspond naturally to linear functions from $\mathbb{R}^n \otimes \mathbb{R}^n \cong \mathbb{R}^{n^2}$ to $\mathbb{R}^m$, etc.)

Equality of Mixed Partials Again

There is a variation of Theorem VIII.2.2.2. involving differentiability:

VIII.4.1.7. THEOREM. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^m$, $\boldsymbol{a} \in U$. If $\boldsymbol{f}$ is second-order differentiable at $\boldsymbol{a}$, then, for $1 \leq i, j \leq n$,

$$\frac{\partial^2 \boldsymbol{f}}{\partial x_j \, \partial x_i}(\boldsymbol{a}) = \frac{\partial^2 \boldsymbol{f}}{\partial x_i \, \partial x_j}(\boldsymbol{a}) \ .$$

For the proof, which is similar to the proof of VIII.2.2.2. and of comparable difficulty, see [Pug15].

VIII.4.1.8. Theorems VIII.2.2.2. and VIII.4.1.7. are complementary: neither directly implies the other. (Clairaut's Theorem, the version of the equality of mixed partials used most frequently, is a special case of both VIII.2.2.2. and VIII.4.1.7., so the distinction between these results is mostly academic.) In VIII.4.1.7. there is no assumption of continuity for any second-order partial, so it might seem to be more general than VIII.2.2.2.. However, the hypotheses of VIII.4.1.7. imply that *all* second-order partials of $\boldsymbol{f}$ exist at $\boldsymbol{a}$ (and even more), which is not necessary for VIII.2.2.2., as the next example shows.

VIII.4.1.9. EXAMPLE. Let $g : \mathbb{R} \to \mathbb{R}$ be a function which is $\mathcal{C}^1$, but does not have a second derivative anywhere, i.e. g is an antiderivative of a continuous but nondifferentiable function. Define $f : \mathbb{R}^2 \to \mathbb{R}$ by $f(x, y) = g(x)$. Then $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial^2 f}{\partial x \partial y}$, $\frac{\partial^2 f}{\partial y \partial x}$, and $\frac{\partial^2 f}{\partial y^2}$ all exist and are continuous everywhere (the last four are identically 0), so the hypotheses (and conclusion!) of VIII.2.2.2. are satisfied everywhere. But f is not second-order differentiable anywhere: in fact, $\frac{\partial^2 f}{\partial x^2}$ does not exist anywhere.

Theorem VIII.4.1.7. has an extension which is a true generalization of VIII.2.2.6.:

VIII.4.1.10. COROLLARY. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ an r'th order differentiable function from U to $\mathbb{R}^m$ ($r > 1$). Then any two mixed partials of $\boldsymbol{f}$ of total order $\leq r$ in which each coordinate direction appears the same number of times are equal on U.

VIII.4.1.11. From the point of view of VIII.4.1.6. (cf. VIII.4.6.1.), VIII.4.1.7. has a nice rephrasing: the bilinear form $d^2\boldsymbol{f}(\boldsymbol{a})$, when defined, is *symmetric*: $d\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x}, \boldsymbol{y}) = d\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Similarly, the r-multilinear map $d^r\boldsymbol{f}(\boldsymbol{a})$, when defined, is symmetric in the sense that is invariant under any permutation of the arguments.

VIII.4.2. Smooth Partitions of Unity and the Smooth Urysohn Lemma

We give multidimensional analogs of the results of ().

VIII.4.2.1. PROPOSITION. Let $\boldsymbol{x} \in \mathbb{R}^n$, and $0 < s < r$. Then there is a $\mathcal{C}^\infty$ function from $\mathbb{R}^n$ to $[0,1]$ which is identically 1 on $B_s(\boldsymbol{x})$, strictly positive on $B_r(\boldsymbol{x})$, and identically 0 outside $B_r(\boldsymbol{x})$. In particular, there is a $\mathcal{C}^\infty$ function which is 1 at $\boldsymbol{x}$, strictly positive on $B_r(\boldsymbol{x})$, and 0 outside $B_r(\boldsymbol{x})$.

PROOF: Let g be a $\mathcal{C}^\infty$ function from $\mathbb{R}$ to $[0,1]$ which is 1 precisely on $(-\infty, s]$ and 0 precisely on $[r, +\infty)$ (), and set $f(\boldsymbol{y}) = g(\|\boldsymbol{y} - \boldsymbol{x}\|)$. 👍

Convergence of Series of Smooth Functions

VIII.4.2.2. THEOREM. Let U be an open set in $\mathbb{R}^n$ and V a bounded open set with $\bar{V} \subseteq U$. Let $0 \leq r \leq \infty$, and (f_k) a sequence of $\mathcal{C}^r$ functions from U to $\mathbb{R}$. Then there is a sequence (ϵ_k) of positive numbers such that if (c_k) is a sequence in $\mathbb{R}$ with $|c_k| \leq \epsilon_k$ for all k, then the series

$$\sum_{k=1}^{\infty} c_k f_k$$

converges uniformly on V to a $\mathcal{C}^r$ function f on V, and for any multi-index α of total degree $\leq r$, the series

$$\sum_{k=1}^{\infty} c_k \frac{\partial^\alpha f_k}{\partial x^\alpha}$$

converges uniformly on V to $\frac{\partial^\alpha f}{\partial x^\alpha}$.

Smooth Functions With Prescribed Zeroes

VIII.4.2.3. THEOREM. Let A be a closed subset of $\mathbb{R}^n$. Then there is a $\mathcal{C}^\infty$ function f from $\mathbb{R}^n$ to $[0,1)$ such that $Z_f = A$ (i.e. f is identically zero on A and strictly positive on A^c).

PROOF: Let $E_1 = A \cup \{\boldsymbol{x} : \|\boldsymbol{x}\| \geq 2\}$, and for each $m > 1$ let $E_m = A \cup \{\boldsymbol{x} : \|\boldsymbol{x}\| \geq m+1\} \cup B_{m-1}(\mathbf{0})$. Then E_m is a closed set, and E_m^c is a bounded open set. Fix m. Then E_m^c is a countable union of open balls (B_{mk}). ByVIII.4.2.1., for each k there is a $\mathcal{C}^\infty$ function f_{mk} from $\mathbb{R}^n$ to $[0,1]$ which is positive precisely on B_{mk}. By VIII.4.2.2. there is a sequence of positive numbers (ϵ_{mk}) such that

$$\sum_{k=1}^{\infty} \epsilon_{mk} f_{mk}$$

converges uniformly on $B_{m+2}(\mathbf{0})$ to a $\mathcal{C}^\infty$ function f_m, which is zero precisely on $E_m \cap B_{m+2}(\mathbf{0})$; f_m extends to a $\mathcal{C}^\infty$ function on all of $\mathbb{R}$ by setting it equal to 0 outside $B_{m+2}(\mathbf{0})$. The function f_m is nonnegative, and nonzero exactly on the points of A^c whose norm is between $m-1$ and $m+1$ (and at $\mathbf{0}$ if $\mathbf{0} \notin A$). Then $\sum_m f_m$ converges pointwise to a $\mathcal{C}^\infty$ function f on $\mathbb{R}$ since, for each $\boldsymbol{x} \in \mathbb{R}^n$, on any open ball around $\boldsymbol{x}$ only finitely many f_m are nonzero. This f is nonnegative, and zero precisely on A. To make the values in $[0,1)$, compose with a $\mathcal{C}^\infty$ function on $\mathbb{R}$ such as $\frac{2}{\pi}\arctan$. 👍

The Smooth Urysohn Lemma

VIII.4.2.4. THEOREM. [SMOOTH URYSOHN LEMMA] Let A be a closed subset of $\mathbb{R}^n$ and U an open set in $\mathbb{R}^n$ containing A. Then there is a C^∞ function ϕ from $\mathbb{R}^n$ to $[0,1]$ such that $\phi \equiv 1$ on A and $\phi \equiv 0$ on $\mathbb{R} \setminus U$. We may choose ϕ so that $\phi^{-1}(\{1\}) = A$ and $\phi^{-1}(\{0\}) = \mathbb{R} \setminus U$.

Alternatively, if E and F are disjoint closed subsets of $\mathbb{R}^n$, there is a smooth function $\phi : \mathbb{R}^n \to [0,1]$ such that $\phi^{-1}(\{1\}) = E$ and $\phi^{-1}(\{0\}) = F$ (set $A = E$ and $U = \mathbb{R}^n \setminus F$).

PROOF: We prove the alternate form. First suppose E is compact. For $\boldsymbol{x} \in E$, let $g(\boldsymbol{x}) = d(\boldsymbol{x}, A)$. Then g is continuous and strictly positive on E, so has a minimum $\epsilon > 0$. The open balls of radius ϵ centered at points of E cover E, so have a finite subcover, centered at points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in E$. Let D_j $(1 \le j \le k)$ be the open ball of radius ϵ centered at $\boldsymbol{x}_j$, and f_j a $\mathcal{C}^\infty$ function which is strictly positive on D_j and 0 outside D_j. Set $f = f_1 + \cdots + f_k$; then f is $\mathcal{C}^\infty$, identically 0 on F, and strictly positive on E. If $c > 0$ is the minimum of f on E, let h be a $\mathcal{C}^\infty$ function on $\mathbb{R}$ with $h(0) = 0$ and h identically 1 on $[c, +\infty)$, and set $\phi = h \circ f$. Then ϕ is $\mathcal{C}^\infty$ on $\mathbb{R}^n$, identically 0 on F, and identically 1 on E.

To make ϕ have the property that $\phi^{-1}(\{1\}) = E$ and $\phi^{-1}(\{0\}) = F$, let θ be a C^∞ function from $\mathbb{R}^n$ to $[0,1)$ such that $\theta = 0$ precisely on $E \cup F$ (V.17.2.9.). Then $\psi_1 = \phi(1-\theta)$ is a C^∞ function from $\mathbb{R}^n$ to $[0,1]$ which is 0 on F and which is 1 precisely on E. Similarly, $\psi_2 = 1 - (1-\phi)(1-\theta)$ is a C^∞ function from $\mathbb{R}^n$ to $[0,1]$ which is 1 on E and 0 precisely on F. Then $\psi = \frac{1}{2}(\psi_1 + \psi_2)$ is a C^∞ function from $\mathbb{R}^n$ to $[0,1]$ which is 1 precisely on E and 0 precisely on F.

Now consider the case of general E and F. Let (h_m) $(m \ge 0)$ be a sequence of $\mathcal{C}^\infty$ functions from $\mathbb{R}$ to $[0,1]$ such that h_m is identically 0 outside $(m-1, m+1)$, h_0 is identically 1 in an interval around 0, and $\sum_m h_m$ is identically 1 (the sum makes sense since only finitely many $h_m(t)$ are nonzero at any particular t). For each m let $E_m = E \cap \{\boldsymbol{x} : m-1 \le \|\boldsymbol{x}\| \le m+1\}$ $(E_1 = E \cap \bar{B}_2(\boldsymbol{0}))$ and $F_m = F \cap \{\boldsymbol{x} : m-1 \le \|\boldsymbol{x}\| \le m+1\}$ $(F_1 = F \cap \bar{B}_2(\boldsymbol{0}))$. Let ϕ_m be a $\mathcal{C}^\infty$ function from $\mathbb{R}^n$ to $[0,1]$ with $E_m = f_m^{-1}(\{1\})$ and $F_m = f_m^{-1}(\{0\})$, and for $\boldsymbol{x} \in \mathbb{R}^n$ set

$$\phi(\boldsymbol{x}) = \sum_{m=1}^{\infty} h_m(\|\boldsymbol{x}\|)\phi_m(\boldsymbol{x})$$

(the sum is well defined and $\mathcal{C}^\infty$ since on any open ball only finitely many terms are not identically 0). Then ϕ has the desired properties. 👍

Smooth Partitions of Unity

VIII.4.2.5. THEOREM. Let $X \subseteq \mathbb{R}^n$, and $\mathcal{U}$ an open cover of X. Then there is a sequence (ϕ_k) of $\mathcal{C}^\infty$ functions from $\mathbb{R}^n$ to $[0,1]$ such that

(i) For each k there is a $U_k \in \mathcal{U}$ such that ϕ_k is identically 0 outside U_k.

(ii) For every compact subset E of X only finitely many ϕ_k are not identically 0 on E.

(iii) For every $\boldsymbol{x} \in X$, $\sum_k \phi_k(\boldsymbol{x}) = 1$ (this is a finite sum for each $\boldsymbol{x}$ by (ii)).

PROOF: We first refine the open cover $\mathcal{U}$. For each $\boldsymbol{x} \in X$ there is an open neighborhood (even ball) V of $\boldsymbol{x}$ such that there is a $U \in \mathcal{U}$ with $\bar{V} \subseteq U$. Thus the set of all nonempty open sets V in $\mathbb{R}^n$ such that there is

a $U \in \mathcal{U}$ with $\bar{V} \subseteq U$ is an open cover $\mathcal{V}$ of X. Since X is second countable in the relative topology, there is a countable subcover $\{V_k : k \in \mathbb{N}\}$ of $\mathcal{V}$ (). For each k choose a $U_k \in \mathcal{U}$ with $\bar{V}_k \subseteq U_k$.

For each k let f_k be a $\mathcal{C}^\infty$ function from $\mathbb{R}^n$ to $[0,1]$ which is 1 on $\bar{V}_k$ and 0 outside U_k (VIII.4.2.4.). Set $\phi_1 = f_1$ and for $k > 1$

$$\phi_k = f_k(1 - f_1) \cdots (1 - f_{k-1}) \ .$$

By induction we have

$$\sum_{k=1}^{m} \phi_k = 1 - (1 - f_1) \cdots (1 - f_m)$$

for any m. If E is any compact subset of X, then $E \subseteq V_1 \cup \cdots \cup V_m$ for some m, so if $\boldsymbol{x} \in E$, $f_k(\boldsymbol{x}) = 1$ for some k, $1 \leq k \leq m$; hence $\phi_k(\boldsymbol{x}) = 0$ for $k > m$ and $\sum_{k=1}^{m} \phi_k(\boldsymbol{x}) = 1$. If $\boldsymbol{x} \in X$, setting $E = \{x\}$ we obtain $\sum_{k=1}^{m} \phi_k(\boldsymbol{x}) = 1$. 👍

The Countable AC is used in this proof.

VIII.4.3. Taylor's Theorem in Several Variables

We can obtain a version of Taylor's Theorem () for real-valued functions of several variables, using the one-variable version. The notation becomes complicated, but the generalization is straightforward.

VIII.4.3.1. The setup throughout this subsection is as follows. We have a convex open set $U \subseteq \mathbb{R}^d$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ $(n+1)$'st order differentiable on U. This implies that all partial derivatives of f of order $\leq n+1$ are defined on U, and that the ones of order $\leq n$ are differentiable, hence continuous, on U.

We fix an $\boldsymbol{x} \in U$, $\boldsymbol{x} \neq \boldsymbol{a}$. Parametrize the line through $\boldsymbol{a}$ and $\boldsymbol{x}$ by

$$\gamma(t) = \boldsymbol{a} + t(\boldsymbol{x} - \boldsymbol{a}) \ .$$

Then γ is an affine function from $\mathbb{R}$ to $\mathbb{R}^d$, and $\gamma(0) = \boldsymbol{a}$, $\gamma(1) = \boldsymbol{x}$. Set $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{a}$, and $I = \{t : \boldsymbol{a} + t\boldsymbol{h} \in U\}$. Then I is an open interval in $\mathbb{R}$ containing $[0, 1]$. Define $g : I \to \mathbb{R}$ by

$$g(t) = f(\gamma(t)) = f(\boldsymbol{a} + t\boldsymbol{h}) \ .$$

By the Chain Rule, g is $(n+1)$-times differentiable on I, so by the one-variable Taylor's Theorem () we have

$$g(t) = \sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!} t^k + r_n(t)$$

for $t \in I$, where

$$r_n(t) = \frac{g^{(n+1)}(s)}{(n+1)!} t^{n+1}$$

for some s, $0 < s < t$ (depending on t). In particular, taking $t = 1$, we have

$$f(\boldsymbol{x}) = \sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!} + \frac{g^{(n+1)}(s)}{(n+1)!}$$

for some s, $0 < s < 1$.

We just need to convert this formula into a formula involving f, its partial derivatives, and $\boldsymbol{x}$ (or, equivalently, $\boldsymbol{h}$).

VIII.4.3.2. The first couple of terms are no problem:

$$g(0) = f(\boldsymbol{a})$$

and, by the Chain Rule, if $\boldsymbol{h} = (h_1, \ldots, h_d)$, we have $\gamma'(t) = \boldsymbol{h}$ for all t, so

$$g'(0) = df(\gamma(0))\boldsymbol{h} = Df(\boldsymbol{a})\boldsymbol{h} = \nabla f(\boldsymbol{a}) \cdot \boldsymbol{h}$$

in our various points of view (the reader should interpret each of these expressions appropriately). In coordinates, we have

$$g'(0) = \sum_{k=1}^{d} \frac{\partial f}{\partial x_k}(\boldsymbol{a}) h_k \ .$$

VIII.4.3.3. More generally, for $t \in I$ we have

$$g'(t) = \sum_{k=1}^{d} \frac{\partial f}{\partial x_k}(\gamma(t))(\boldsymbol{a})h_k$$

so we can compute $g''(0)$ by applying the Chain Rule again:

$$g''(0) = \sum_{k=1}^{d} h_k D\left(\frac{\partial f}{\partial x_k}\right)(\boldsymbol{a})\boldsymbol{h} = \sum_{k=1}^{d} h_k \nabla\left(\frac{\partial f}{\partial x_k}\right)(\boldsymbol{a}) \cdot \boldsymbol{h} \ .$$

We have

$$\nabla\left(\frac{\partial f}{\partial x_k}\right)(\boldsymbol{a}) = \left(\frac{\partial^2 f}{\partial x_1\, \partial x_k}(\boldsymbol{a}), \ldots, \frac{\partial^2 f}{\partial x_d\, \partial x_k}(\boldsymbol{a})\right)$$

so, expanding the dot product in coordinates, we get

$$g''(0) = \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^2 f}{\partial x_j\, \partial x_k}(\boldsymbol{a})h_j h_k \ .$$

This can be written as

$$\boldsymbol{h} \cdot \left[D^2 f(\boldsymbol{a})\boldsymbol{h}\right]$$

where

$$D^2 f(\boldsymbol{a}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1\, \partial x_1} & \frac{\partial^2 f}{\partial x_1\, \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\, \partial x_d} \\ \frac{\partial^2 f}{\partial x_2\, \partial x_1} & \frac{\partial^2 f}{\partial x_2\, \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2\, \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d\, \partial x_1} & \frac{\partial^2 f}{\partial x_d\, \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_d\, \partial x_d} \end{bmatrix} .$$

This matrix is called the *Hessian matrix* of f at $\boldsymbol{a}$, especially if $\nabla f(\boldsymbol{a}) = \boldsymbol{0}$. This matrix actually gives a tensor of rank 2 (). Note that if $n \geq 2$, the Hessian matrix is symmetric (VIII.2.2.2.). We can use the symbolic notation (cf. VIII.4.1.6.)

$$d^2 f(\boldsymbol{a})\boldsymbol{h}^2 = \boldsymbol{h} \cdot \left[D^2 f(\boldsymbol{a})\boldsymbol{h}\right] = \sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^2 f}{\partial x_j\, \partial x_k}(\boldsymbol{a})h_j h_k \ .$$

VIII.4.3.4. In the same way, we can compute the higher derivatives of g:

$$g'''(0) = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} \frac{\partial^3 f}{\partial x_i\, \partial x_j\, \partial x_k}(\boldsymbol{a})h_i h_j h_k$$

which we may write symbolically as $d^3 f(\boldsymbol{a})\boldsymbol{h}^3$, corresponding to the three-dimensional array $D^3 f(\boldsymbol{a})$ whose (i,j,k) entry is $\frac{\partial^3 f}{\partial x_i\,\partial x_j\,\partial x_k}(\boldsymbol{a})$. (This is a tensor of rank 3.) Continuing, we get for arbitrary $r \leq n+1$:

$$g^{(r)}(0) = d^m f(\boldsymbol{a})\boldsymbol{h}^r = \sum_{j_1=1}^{d}\sum_{j_2=1}^{d}\cdots\sum_{j_r=1}^{d}\frac{\partial^r f}{\partial x_{j_1}\,\partial x_{j_2}\cdots\partial x_{j_r}}(\boldsymbol{a})h_{j_1}h_{j_2}\cdots h_{j_r}\ .$$

We also write $d^0 f(\boldsymbol{a}) = f(\boldsymbol{a})$.

Thus we obtain the final result:

VIII.4.3.5. Theorem. [Taylor's Theorem With Remainder, Multivariable Version] Let U be a convex open set in $\mathbb{R}^d$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ a function which is $(n+1)$'st order differentiable ($n \geq 0$) on U. Then, for $\boldsymbol{x} \in U$, we have

$$f(\boldsymbol{x}) = \sum_{k=0}^{n}\frac{1}{k!}d^k f(\boldsymbol{a})(\boldsymbol{x}-\boldsymbol{a})^k + r_n(\boldsymbol{x})$$

where

$$r_n(\boldsymbol{x}) = \frac{1}{(n+1)!}d^{n+1}f(\boldsymbol{c})(\boldsymbol{x}-\boldsymbol{a})^{n+1}$$

for some $\boldsymbol{c}$ on the line segment between $\boldsymbol{a}$ and $\boldsymbol{x}$.

We can relax the hypothesis that U be convex to just requiring that U be star-shaped with respect to $\boldsymbol{a}$, i.e. for every $\boldsymbol{x} \in U$, the line segment between $\boldsymbol{a}$ and $\boldsymbol{x}$ is contained in U.

The Two-Dimensional Case

VIII.4.3.6. Let us write out the case $d = 2$. We write (x,y) for the coordinates, and (a,b) for the point we expand around. We combine together terms which are equal due to the equality of mixed partials ().

VIII.4.3.7. Corollary. [Taylor's Theorem With Remainder, Two-Dimensional Version] Let U be a convex open set in $\mathbb{R}^2$, $(a,b) \in U$, $f : U \to \mathbb{R}$ an $(n+1)$'st order differentiable function. Then, for any $(x,y) \in U$,

$$f(x,y) = f(a,b) + \left[\frac{\partial f}{\partial x}(a,b)(x-a) + \frac{\partial f}{\partial y}(a,b)(y-b)\right]$$
$$+\frac{1}{2}\left[\frac{\partial^2 f}{\partial x^2}(a,b)(x-a)^2 + 2\frac{\partial^2 f}{\partial x\,\partial y}(a,b)(x-a)(y-b) + \frac{\partial^2 f}{\partial y^2}(a,b)(y-b)^2\right]$$
$$+\frac{1}{6}\left[\frac{\partial^3 f}{\partial x^3}(a,b)(x-a)^3 + 3\frac{\partial^3 f}{\partial x^2\,\partial y}(a,b)(x-a)^2(y-b) + 3\frac{\partial^3 f}{\partial x\,\partial y^2}(a,b)(x-a)(y-b)^2 + \frac{\partial^3 f}{\partial y^3}(a,b)(y-b)^3\right]$$
$$+\cdots+\frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}\frac{\partial^n f}{\partial x^k\,\partial y^{n-k}}(a,b)(x-a)^k(y-b)^{n-k} + r_n(x,y)$$

where

$$r_n(x,y) = \frac{1}{(n+1)!} d^{n+1} f(\boldsymbol{c})((x-a, y-b)^{n+1}$$

for some $\boldsymbol{c}$ on the line segment between (a,b) and (x,y).

Taylor Polynomials in the Multivariable Case

VIII.4.3.8. The multivariable Taylor's Theorem VIII.4.3.5. can also be developed from scratch in the same way as the one-variable case, although the *proof* via the one-variable case is much simpler. We begin by defining the Taylor polynomials:

VIII.4.3.9. DEFINITION. Let U be a convex open set in $\mathbb{R}^d$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ a function which is n'th order differentiable ($n \geq 0$) on U. Then, for $\boldsymbol{x} \in U$,

$$p_n(\boldsymbol{x}) = \sum_{k=0}^{n} \frac{1}{k!} d^k f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})^k$$

is the *n'th order Taylor polynomial of f around $\boldsymbol{a}$*.

The next result is a straightforward but somewhat ugly computation:

VIII.4.3.10. PROPOSITION. Let U be a convex open set in $\mathbb{R}^d$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ a function which is n'th order differentiable ($n \geq 0$) on U. Then the n'th order Taylor polynomial of f around $\boldsymbol{a}$ is a polynomial of degree $\leq n$ in the coordinates, and is the unique polynomial of degree $\leq n$ whose partial derivatives of all orders $\leq n$ at $\boldsymbol{a}$ agree with those of f at $\boldsymbol{a}$.

Taylor's Theorem then becomes:

VIII.4.3.11. THEOREM. [TAYLOR'S THEOREM, MULTIVARIABLE VERSION] Let U be a convex open set in $\mathbb{R}^d$, $\boldsymbol{a} \in U$, and $f : U \to \mathbb{R}$ a function which is $(n+1)$'st order differentiable ($n \geq 0$) on U. Let p_n be the n'th order Taylor polynomial of f around $\boldsymbol{a}$. Then, for $\boldsymbol{x} \in U$, we have

$$f(\boldsymbol{x}) = p_n(\boldsymbol{x}) + r_n(\boldsymbol{x})$$

where

$$r_n(\boldsymbol{x}) = \frac{1}{(n+1)!} d^{n+1} f(\boldsymbol{c})(\boldsymbol{x} - \boldsymbol{a})^{n+1}$$

for some $\boldsymbol{c}$ on the line segment between $\boldsymbol{a}$ and $\boldsymbol{x}$. In particular,

$$\lim_{\boldsymbol{x} \to \boldsymbol{a}} \frac{f(\boldsymbol{x}) - p_n(\boldsymbol{x})}{\|\boldsymbol{x} - \boldsymbol{a}\|^n} = 0$$

and p_n is the unique polynomial of order $\leq n$ with this property.

PROOF: Only the last statement needs proof. We only give the proof in the case where f is $\mathcal{C}^{n+1}$ on U; the statement is true if f is just $(n+1)$'st order differentiable on U, and even if it is only n'th order differentiable at $\boldsymbol{a}$, but the proof is more delicate (). If f is $\mathcal{C}^{n+1}$, and V is a bounded neighborhood of $\boldsymbol{a}$ with $\bar{V} \subseteq U$, then the $(n+1)$'st order partials of f are uniformly bounded on V, and it follows that there is a constant M such that $|r_n(\boldsymbol{x})| \leq M\|\boldsymbol{x}-\boldsymbol{a}\|^{n+1}$ for $\boldsymbol{x} \in V$, so

$$\lim_{\boldsymbol{x}\to\boldsymbol{a}} \frac{r_n(\boldsymbol{x})}{\|\boldsymbol{x}-\boldsymbol{a}\|^n} = 0 \ .$$

VIII.4.4. Critical Point Analysis

We can give a second-derivative test to determine local and maxima and minima of functions of several variables in analogy with the one-dimensional case, although, as might be expected, the criterion is a little more complicated. Note that all the definitions and results we give here exactly reduce to the usual ones in the one-dimensional case.

We first note the definition of local extrema and the analog of ():

VIII.4.4.1. DEFINITION. Let U be an open set in $\mathbb{R}^d$, and f a function from U to $\mathbb{R}$. Then f has a *local maximum* at $\boldsymbol{a} \in U$ if there is an open ball B around $\boldsymbol{a}$, contained in U, such that $f(\boldsymbol{x}) \leq f(\boldsymbol{a})$ for all $\boldsymbol{x} \in B$; the local maximum is *strict* if there is a B such that $f(\boldsymbol{x}) < f(\boldsymbol{a})$ for all $\boldsymbol{x} \in B$, $\boldsymbol{x} \neq \boldsymbol{a}$. The function f has a *local minimum* at $\boldsymbol{a} \in U$ if there is an open ball B around $\boldsymbol{a}$, contained in U, such that $f(\boldsymbol{x}) \geq f(\boldsymbol{a})$ for all $\boldsymbol{x} \in B$. The function f has a *local extremum* at $\boldsymbol{a}$ if it has either a local maximum or a local minimum at $\boldsymbol{a}$. Strict local minima and strict local extrema are defined accordingly.

The next result is an immediate consequence of () by fixing all but one coordinate and noting that if f has a local extremum at $\boldsymbol{a}$, so does any such restricted function.

VIII.4.4.2. PROPOSITION. Let U be an open set in $\mathbb{R}^d$, $f : U \to \mathbb{R}$, and $\boldsymbol{a} \in U$. If f has a local extremum at $\boldsymbol{a}$, then any first-order partial derivative of f which exists at $\boldsymbol{a}$ is zero at $\boldsymbol{a}$.

VIII.4.4.3. COROLLARY. Let U be an open set in $\mathbb{R}^d$, $f : U \to \mathbb{R}$, and $\boldsymbol{a} \in U$. If f has a local extremum at $\boldsymbol{a}$, and f is differentiable at $\boldsymbol{a}$, then $Df(\boldsymbol{a}) = 0$ (i.e. $\nabla f(\boldsymbol{a}) = \boldsymbol{0}$).

VIII.4.4.4. DEFINITION. Let U be an open set in $\mathbb{R}^d$, $f : U \to \mathbb{R}$, and $\boldsymbol{a} \in U$. Then $\boldsymbol{a}$ is a *critical point* of f if $\nabla f(\boldsymbol{a}) = \boldsymbol{0}$ or if f is not differentiable at $\boldsymbol{a}$.

The Second Derivative or Hessian Matrix

VIII.4.4.5. Recall () that if f is second-order differentiable at $\boldsymbol{a}$, the matrix

$$D^2 f(\boldsymbol{a}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{a}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{a}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(\boldsymbol{a}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{a}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{a}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(\boldsymbol{a}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\boldsymbol{a}) & \frac{\partial^2 f}{\partial x_d \partial x_2}(\boldsymbol{a}) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(\boldsymbol{a}) \end{bmatrix}$$

is called the *Hessian matrix* of f at $\boldsymbol{a}$. It is a symmetric matrix which is the multidimensional version of the second derivative of f at $\boldsymbol{a}$. If $\boldsymbol{a}$ is a critical point of f, this matrix can determine whether f has a local extremum at $\boldsymbol{a}$.

Directional Second Derivatives

VIII.4.4.6. Let f be second-order differentiable on an open set $U \subseteq \mathbb{R}^d$, and $\boldsymbol{a} \in U$. If $\boldsymbol{u}$ is a unit vector in $\mathbb{R}^d$, let $\gamma(t) = \boldsymbol{a} + t\boldsymbol{u}$ and $g(t) = f(\gamma(t))$. Then γ and g are defined and second-order differentiable in an interval around 0, and $g'(0)$ is exactly the directional derivative $\frac{\partial f}{\partial \boldsymbol{u}}(\boldsymbol{a})$. It makes sense to interpret $g''(0)$ as the *directional second derivative* of f in the direction $\boldsymbol{u}$, and write it $\frac{\partial^2 f}{\partial \boldsymbol{u}^2}(\boldsymbol{a})$. By a computation identical to the one in VIII.4.3.3., we have:

VIII.4.4.7. Proposition. We have

$$\frac{\partial^2 f}{\partial \boldsymbol{u}^2}(\boldsymbol{a}) = d^2 f(\boldsymbol{a})\boldsymbol{u}^2 = (D^2 f(\boldsymbol{a})\boldsymbol{u}) \cdot \boldsymbol{u} \ .$$

Positive Definite Matrices

Recall the following facts from linear algebra ():

VIII.4.4.8. Proposition. Let A be a symmetric $d \times d$ matrix with real entries. Then all eigenvalues of A are real; A has d real eigenvalues, counted with multiplicity, and the eigenvectors of A span $\mathbb{R}^d$. Eigenvectors corresponding to different eigenvalues are orthogonal.

VIII.4.4.9. Proposition. Let A be a symmetric $d \times d$ matrix with real entries. The following are equivalent:

(i) $(A\boldsymbol{x}) \cdot \boldsymbol{x} > 0$ for every nonzero $\boldsymbol{x} \in \mathbb{R}^d$.

(ii) All eigenvalues of A are strictly positive.

A (symmetric) A satisfying these conditions is called *positive definite*. In fact, if A is positive definite and $\lambda > 0$ is the smallest eigenvalue of A, then $(A\boldsymbol{x}) \cdot \boldsymbol{x} \geq \lambda \|\boldsymbol{x}\|^2$ for every $\boldsymbol{x}$.

VIII.4.4.10. Similarly, $(A\boldsymbol{x}) \cdot \boldsymbol{x} < 0$ for every nonzero $\boldsymbol{x} \in \mathbb{R}^d$ if and only if all eigenvalues of A are strictly negative. Such a matrix is called *negative definite*. We also make the following definitions, which are perhaps not quite standard:

VIII.4.4.11. Definition. Let A be a symmetric $d \times d$ matrix.

A is *indefinite* if it has both positive and negative eigenvalues (it may or may not have 0 as an eigenvalue).

A is *degenerate* if 0 is an eigenvalue, and all the other eigenvalues (if any) have the same sign.

VIII.4.4.12. Thus, if A is a symmetric $d \times d$ matrix, there are four mutually exclusive possibilities:

A is positive definite.

A is negative definite.

A is indefinite.

A is degenerate.

If $d = 1$, the indefinite case cannot occur, and the other three mean the 1×1 matrix (number) is positive, negative, or zero respectively.

The last statement of VIII.4.4.9. has an important consequence:

VIII.4.4.13. Corollary. Let A be a positive definite [negative definite, indefinite] $d \times d$ matrix. Then there is an $\epsilon > 0$ such that every symmetric $n \times n$ matrix B with $\|A - B\| < \epsilon$ is positive definite [negative definite, indefinite]. In fact, we can take ϵ to be the minimum of the absolute values of the nonzero eigenvalues of A.

The main result is:

VIII.4.4.14. THEOREM. [SECOND DERIVATIVE TEST, MULTIVARIABLE VERSION] Let U be an open set in $\mathbb{R}^d$, and $f : U \to \mathbb{R}$ a $\mathcal{C}^2$ function. If $\boldsymbol{a} \in U$ and $\nabla f(\boldsymbol{a}) = \mathbf{0}$, then

(i) If $D^2 f(\boldsymbol{a})$ is positive definite, then f has a strict local minimum at $\boldsymbol{a}$.

(ii) If $D^2 f(\boldsymbol{a})$ is negative definite, then f has a strict local maximum at $\boldsymbol{a}$.

(iii) If $D^2 f(\boldsymbol{a})$ is indefinite, then f has neither a local maximum nor local minimum at $\boldsymbol{a}$.

(iv) If $D^2 f(\boldsymbol{a})$ is degenerate, then f may or may not have a local extremum at $\boldsymbol{a}$ and a local extremum may or may not be strict, and almost no conclusion may be drawn, i.e. the test fails. The only conclusion is that if $D^2 f(\boldsymbol{a})$ has a positive eigenvalue, f cannot have a local maximum at $\boldsymbol{a}$, and if it has a negative eigenvalue, then f cannot have a local minimum at $\boldsymbol{a}$.

PROOF: We first prove (iii) and part of (iv). This argument only requires that f be second-order differentiable at $\boldsymbol{a}$, not necessarily $\mathcal{C}^2$ on a neighborhood. Suppose $D^2 f(\boldsymbol{a})$ has a positive eigenvalue λ, and let $\boldsymbol{u}$ be a unit eigenvector with eigenvalue λ. If g denotes the restriction of f to the line L through $\boldsymbol{a}$ in the direction $\boldsymbol{u}$, then

$$g''(0) = \frac{\partial^2 f}{\partial \boldsymbol{u}^2}(\boldsymbol{a}) = (D^2 f(\boldsymbol{a})\boldsymbol{u}) \cdot \boldsymbol{u} = (\lambda \boldsymbol{u}) \cdot \boldsymbol{u} = \lambda > 0$$

and thus f restricted to L has a strict minimum at $\boldsymbol{a}$ (). So f cannot have a local maximum, strict or not, at $\boldsymbol{a}$. Similarly, if $D^2 f(\boldsymbol{a})$ has a negative eigenvalue, then f cannot have a local minimum at $\boldsymbol{a}$.

This argument does not obviously suffice to prove (i) and (ii): radial behavior does not obviously tell the whole story without some uniformity. But we can easily obtain (i) and (ii) from Taylor's Theorem with $n = 1$. Suppose $D^2 f(\boldsymbol{a})$ is positive definite. Then there is a ball B centered at $\boldsymbol{a}$ such that $B \subseteq U$ and $D^2 f(\boldsymbol{b})$ is positive definite for all $\boldsymbol{b} \in B$ (). Since $\nabla f(\boldsymbol{a}) = \mathbf{0}$, Taylor's Theorem with $n = 1$ says that, for any $\boldsymbol{x} \in B$, $\boldsymbol{x} \neq \boldsymbol{a}$,

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \frac{1}{2} d^2 f(\boldsymbol{c})(\boldsymbol{x} - \boldsymbol{a})^2 = f(\boldsymbol{a}) + [D^2 f(\boldsymbol{c})(\boldsymbol{x} - \boldsymbol{a})] \cdot (\boldsymbol{x} - \boldsymbol{a})$$

for some $\boldsymbol{c}$ on the line segment between $\boldsymbol{a}$ and $\boldsymbol{x}$. Any such $\boldsymbol{c}$ is in B and hence $D^2 f(\boldsymbol{c})$ is positive definite, $[D^2 f(\boldsymbol{c})(\boldsymbol{x} - \boldsymbol{a})] \cdot (\boldsymbol{x} - \boldsymbol{a}) > 0$, $f(\boldsymbol{x}) > f(\boldsymbol{a})$. The case where $D^2 f(\boldsymbol{a})$ is negative definite is analogous.

There are 1-dimensional examples () giving the other statements of (iv). See VIII.4.4.18. for some two-dimensional examples. 👍

The Two-Dimensional Case

VIII.4.4.15. The case $d = 2$ is slightly simpler than the general case. In this case, $D^2 f(\boldsymbol{a})$ has two eigenvalues, which may coincide (exactly when $D^2 f(\boldsymbol{a})$ is a scalar multiple of I_2). It is positive definite if both eigenvalues are positive, negative definite if both are negative, indefinite if they are of opposite sign, and degenerate if one (or both) are zero. In fact, a symmetric (real) 2×2 matrix is

positive definite if and only if the diagonal entries and the determinant are positive.

negative definite if and only if the diagonal entries are negative and the determinant is positive.

indefinite if and only if the determinant is negative.

degenerate if and only if the determinant is zero.

[The determinant of a symmetric matrix is the product of the eigenvalues (counted with multiplicity).]

Two-Dimensional Examples

The following examples are generic:

VIII.4.4.16. EXAMPLES. (i) Let $f(x, y) = x^2 + y^2$. Then $\nabla f(x, y) = (2x, 2y)$, so $\nabla f(0, 0) = \mathbf{0}$.

$$D^2 f(0, 0) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

which is positive definite, so f has a strict local minimum (actually a strict global minimum) at $(0, 0)$. See Figure VIII.4.

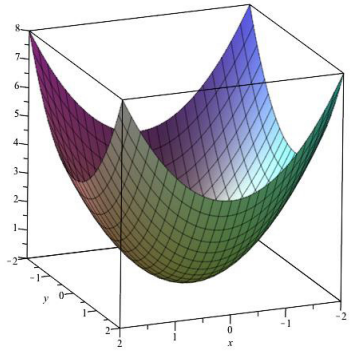


Figure VIII.4: Maple plot of graph of $f(x, y) = x^2 + y^2$

If the eigenvalues are different, the surface is elongated in one direction. For example, if $f(x, y) = x^2+2y^2$, we have

$$D^2 f(0, 0) = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix} .$$

See Figure VIII.5.

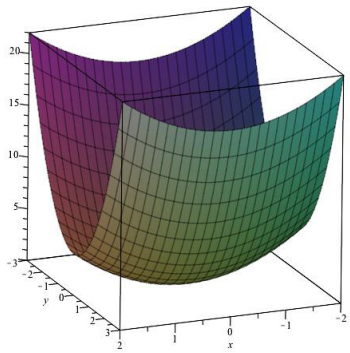


Figure VIII.5: Maple plot of graph of $f(x, y) = x^2 + 2y^2$

(ii) The negatives of these examples give strict local maxima (the Hessians are negative definite):

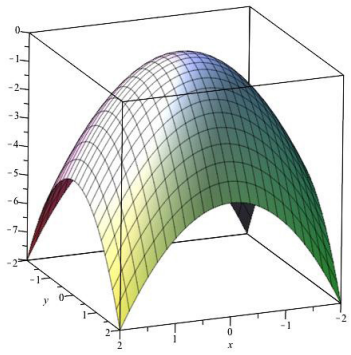


Figure VIII.6: Maple plot of graph of $f(x, y) = -x^2 - y^2$

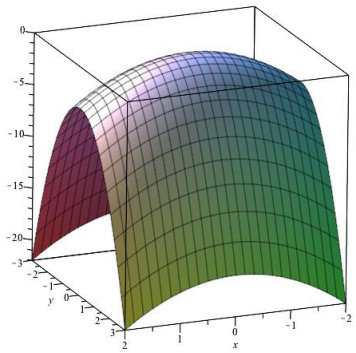


Figure VIII.7: Maple plot of graph of $f(x, y) = -x^2 - 2y^2$

(iii) Let $f(x,y) = x^2 - y^2$. Then $\nabla f(x,y) = (2x, -2y)$, so $\nabla f(0,0) = \mathbf{0}$.

$$D^2 f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

which is indefinite, so f does not have a local extremum at $(0,0)$; it has what is called a *saddle point* at $(0,0)$. See Figure VIII.8.

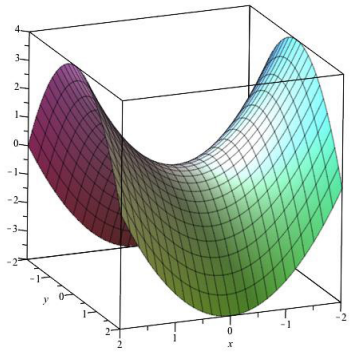


Figure VIII.8: Maple plot of graph of $f(x,y) = x^2 - y^2$

VIII.4.4.17. In the general two-dimensional case, since the corresponding eigenvectors are orthogonal, we can make an affine change of coordinates so that $\boldsymbol{a} = \mathbf{0}$, $f(\boldsymbol{a}) = 0$, and $D^2 f(\boldsymbol{a})$ is a diagonal matrix. Thus, if $D^2 f(\boldsymbol{a})$ is not degenerate, f locally approximately looks like one of the above cases (by the Morse Lemma (), there is a smooth change of coordinates so it is locally *exactly* one of these).

VIII.4.4.18. EXAMPLES. If $D^2f(\boldsymbol{a})$ is degenerate, almost any sort of behavior nearby is possible:
(i) Let $f(x,y) = x^2 + y^4$. Then

$$D^2f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

and f has a strict local minimum at $\mathbf{0}$. See Figure VIII.9.

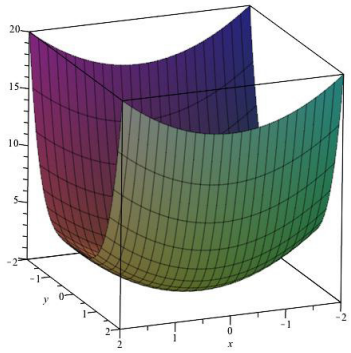


Figure VIII.9: Maple plot of graph of $f(x,y) = x^2 + y^4$

(ii) Let $f(x,y) = x^2$. Then

$$D^2f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

and f has a local minimum at $\mathbf{0}$, but it is not strict. See Figure VIII.10.

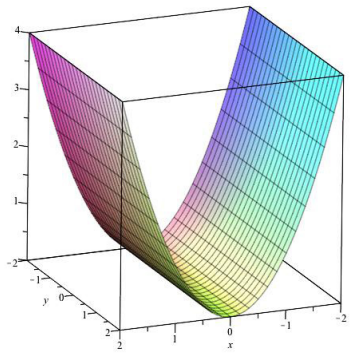


Figure VIII.10: Maple plot of graph of $f(x,y) = x^2$

(iii) Let $f(x,y) = x^2 - y^4$. Then

$$D^2 f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

and f has a saddle point at $\mathbf{0}$. See Figure VIII.11.

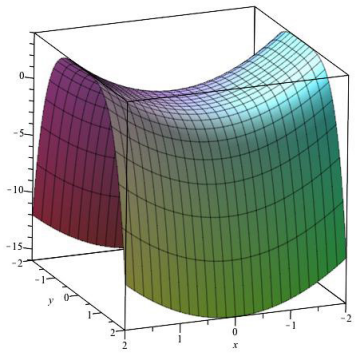


Figure VIII.11: Maple plot of graph of $f(x,y) = x^2 - y^4$

(iv) Let $f(x,y) = x^2 + y^3$. Then

$$D^2 f(0,0) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

and f has what may be called a chair point at $\mathbf{0}$. See Figure VIII.12.

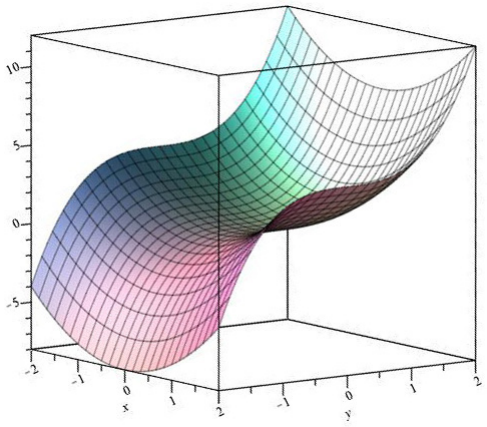


Figure VIII.12: Maple plot of graph of $f(x,y) = x^2 + y^3$

(v) In (i)–(iv), we can instead take $f(x,y)$ to be x^4+y^4, x^4, x^4-y^4, or x^4+y^3 respectively; then $D^2f(0,0)=0$ and the same conclusion holds. Also try $f(x,y)=x^3+y^3$. We can also take $f(x,y)$ to be $-x^4-y^4$ or $-x^4$ to get a local maximum, strict or not, with the same Hessian.
(vi) Let $f(x,y)=x^3-3xy^2$. Then $\nabla f(0,0)=\mathbf{0}$ and $D^2f(0,0)=0$. The graph is called a *monkey saddle* for obvious reasons. See Figure VIII.13.

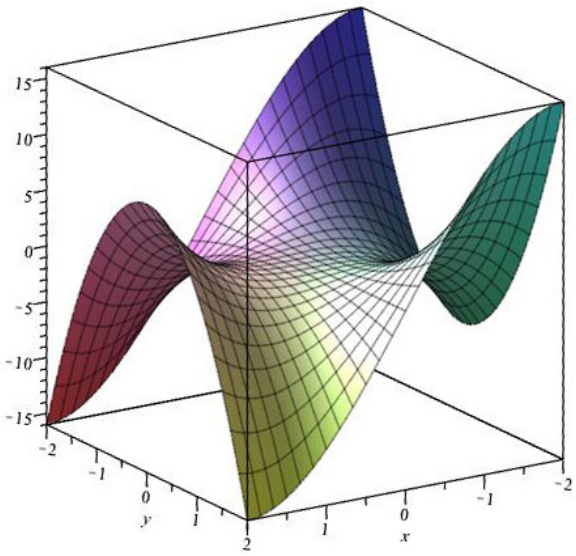


Figure VIII.13: Maple plot of graph of $f(x,y)=x^3-3xy^2$

(vii) There are much more complicated examples if $D^2f(\boldsymbol{a})$ is degenerate even if $d=1$, such as () and (). These can be made into two-dimensional examples by taking $g(x,y)=f(x)$, $g(x,y)=f(x)f(y)$, etc.

VIII.4.5. The Morse Lemma

VIII.4.6. Exercises

VIII.4.6.1. (a) Let $\boldsymbol{f}:U\subseteq\mathbb{R}^n\to\mathbb{R}^m$ be second-order differentiable at $\boldsymbol{a}\in U$. Show that

$$d^2\boldsymbol{f}(\boldsymbol{a})(\mathbf{e}_j)(\mathbf{e}_k)=\frac{\partial^2\boldsymbol{f}}{\partial x_j\partial x_k}(\boldsymbol{a})$$

for any j,k. Thus the second-order partials of $\boldsymbol{f}$ are the natural coordinates for the identification of $d^2\boldsymbol{f}$ regarded as a function from U to $\mathbb{R}^{nm}$.
(b) More generally, if $\boldsymbol{f}$ is r'th order differentiable at $\boldsymbol{a}$, and $d^r\boldsymbol{f}(\boldsymbol{a})$ is regarded as an r-multilinear map from $\mathbb{R}^n\times\cdots\times\mathbb{R}^n$ to $\mathbb{R}^m$, then for any $i_1,\dots,i_r$ we have

$$d^r\boldsymbol{f}(\boldsymbol{a})(\mathbf{e}_{i_1},\dots,\mathbf{e}_{i_r})=\frac{\partial^r\boldsymbol{f}}{\partial x_{i_1}\cdots\partial x_{i_r}}(\boldsymbol{a})\ .$$

(c) Give a similar description of $D^r \boldsymbol{f}(\boldsymbol{a})$.

VIII.5. The Implicit and Inverse Function Theorems

The Implicit and Inverse Function Theorems are the deepest results from differential calculus, and among the most important; they are the primary means of passing from the *infinitesimal* to the *local* in analysis and geometry. Although they are distinct results, each is a relatively easy corollary of the other, so one only needs to carry out the (rather difficult) proof of one of them to obtain both (and in fact the two theorems have a common generalization, cf. VIII.5.2.1.). They can both be regarded as aspects of the general problem of solving systems of functional equations; existence and uniqueness theorems for solutions to differential equations fall into the same class of results (see XI.1.10.12.). See III.5.3.7. and III.5.3.8. for purely algebraic versions.

These results are stated in less generality than they have to be in many references, with proofs more difficult than they have to be (and at that not even always entirely complete or correct). There are several standard proofs, and some of these proofs arguably give better insight into *why* the theorems are true than the simplest ones. We will give proofs of each theorem which illustrate two common approaches; each approach gives versions not obtainable from the others.

These theorems have a long and complicated history, going back at least to Descartes and used extensively by Newton, cf. VIII.5.11.7. (although for a long time it was not always appreciated that they had to be *proved*), and there are many versions. We will give only a few. The Smooth ($\mathcal{C}^1$) Implicit Function Theorem (VIII.5.1.7.) was first proved by Dini in the 1870's, although Dini's contributions were not properly recognized for a long time since he published them only in his lecture notes, which had a limited circulation. For continuous and higher-order smooth versions, see [Gou03] and [You09]. The book [KP02a] contains the most thorough treatment available of the various forms of the Implicit and Inverse Function Theorems.

VIII.5.1. Statements of the Theorems

VIII.5.1.1. The idea of the theorems is that if one has n equations in m unknowns, then there is "usually" a unique solution if $n = m$, many solutions if $n < m$, and no solutions if $n > m$; if $n < m$, it should be possible to solve for n of the variables as functions of the others (specifying values for $m - n$ of the variables yields a system of n equations in the remaining n variables). These statements are literally true if the equations are linear (cf. ()) unless there is degeneracy in the system. Since smooth functions are "approximately linear", it might also be possible (and in fact *is*) to do the same locally in the smooth case.

VIII.5.1.2. Vector notation is a convenient way to rephrase systems of equations. An equation in unknowns $x_1, \ldots, x_m$ can, by moving everything to one side of the equation, be written as $F(x_1, \ldots, x_m) = 0$, where $F : \mathbb{R}^m \to \mathbb{R}$ is a function. A set of n equations

$$F_1(x_1, \ldots, x_m) = 0$$

$$F_2(x_1, \ldots, x_m) = 0$$

$$\ldots$$

$$F_n(x_1, \ldots, x_m) = 0$$

can be written as $\boldsymbol{F}(x_1, \ldots, x_m) = \mathbf{0}$, or $\boldsymbol{F}(\boldsymbol{x}) = \mathbf{0}$, where $\boldsymbol{F}$ is the function from $\mathbb{R}^m$ (or a subset) to $\mathbb{R}^n$ with coordinate functions $(F_1, \ldots, F_n)$, and $\boldsymbol{x} = (x_1, \ldots, x_m)$.

VIII.5.1.3. In the case $m > n$, it is convenient to change notation slightly by setting $p = m - n$. We would like to be able to solve the system $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}$ for n of the variables in terms of the other p variables. (Here the relative sizes of n and p do not matter; we could have $n < p$, $n = p$, or $n > p$.) We can in general choose any n of the variables to solve for in terms of the other ones; we will notationally choose the last n and rename them $y_1, \ldots, y_n$, i.e. our equation to solve is of the form

$$\boldsymbol{F}(x_1, \ldots, x_p, y_1, \ldots, y_n) = \boldsymbol{0}$$

or $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$, where $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. A *solution* on a subset W of $\mathbb{R}^p$ is a function $\boldsymbol{f} : W \to \mathbb{R}^n$ such that the pair $(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ satisfies the equation for all $\boldsymbol{x} \in W$, i.e. $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{0}$ for all $\boldsymbol{x} \in W$. We say the function $\boldsymbol{f}$ is *implicitly defined* by the equation $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$. It is rarely possible to find an explicit formula for an implicitly defined function in the nonlinear case.

VIII.5.1.4. Examples with $n = p = 1$ from elementary calculus (e.g. the one in VIII.5.1.11.) show that one can only expect the equation $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$ to define $\boldsymbol{y}$ as a continuous function of $\boldsymbol{x}$ locally, i.e. in some neighborhood of a point $\boldsymbol{a}$ where $(\boldsymbol{a}, \boldsymbol{b})$ satisfies the equation, and then only around points where there is some nondegeneracy condition on $\boldsymbol{F}$. If $n = p = 1$ and F is smooth, the graph of the equation $F(x, y) = 0$ is a curve in $\mathbb{R}^2$, and this curve is in general only locally the graph of a smooth function, and not even necessarily around every point of the curve; points where the curve crosses itself or has a vertical tangent must be excluded (see Exercise VIII.5.11.2.).

VIII.5.1.5. In the statements of the theorems, we will use the following notation. Let $\boldsymbol{F}$ be a function from an open set U in $\mathbb{R}^{p+n}$ to $\mathbb{R}^n$. Write the coordinates of $\mathbb{R}^{p+n}$ as $(x_1, \ldots, x_p, y_1, \ldots, y_n)$, or $(\boldsymbol{x}, \boldsymbol{y})$ with $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ as above, and let $F_1, \ldots, F_n$ be the coordinate functions of $\boldsymbol{F}$. So, in coordinates,

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = (F_1(\boldsymbol{x}, \boldsymbol{y}), \ldots, F_n(\boldsymbol{x}, \boldsymbol{y})) = (F_1(x_1, \ldots, x_p, y_1, \ldots, y_n), \ldots, F_n(x_1, \ldots, x_p, y_1, \ldots, y_n))) \ .$$

Let $(\boldsymbol{x}, \boldsymbol{y}) \in U$, and suppose all partial derivatives of the form $\frac{\partial F_i}{\partial y_j}(\boldsymbol{x}, \boldsymbol{y})$ exist. Write

$$D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\partial(F_1, \ldots, F_n)}{\partial(y_1, \ldots, y_n)}(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} \frac{\partial F_1}{\partial y_1}(\boldsymbol{x}, \boldsymbol{y}) & \cdots & \frac{\partial F_1}{\partial y_n}(\boldsymbol{x}, \boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial y_1}(\boldsymbol{x}, \boldsymbol{y}) & \cdots & \frac{\partial F_n}{\partial y_n}(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$$

and, if all partial derivatives of the form $\frac{\partial F_i}{\partial x_j}(\boldsymbol{x}, \boldsymbol{y})$ exist, write

$$D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\partial(F_1, \ldots, F_n)}{\partial(x_1, \ldots, x_p)}(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(\boldsymbol{x}, \boldsymbol{y}) & \cdots & \frac{\partial F_1}{\partial x_p}(\boldsymbol{x}, \boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_1}(\boldsymbol{x}, \boldsymbol{y}) & \cdots & \frac{\partial F_n}{\partial x_p}(\boldsymbol{x}, \boldsymbol{y}) \end{bmatrix}$$

(note that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ is $n \times n$ and $D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ is $n \times p$ if they are defined). Then if $\boldsymbol{F}$ is differentiable at $(\boldsymbol{x}, \boldsymbol{y})$, we have that the $n \times (p + n)$ matrix $D\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ partitions as

$$D\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = \left[\begin{array}{ccc} & \vdots & \\ D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) & \vdots & D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) \\ & \vdots & \end{array}\right] .$$

VIII.5.1.6. THEOREM. [IMPLICIT FUNCTION THEOREM, CONTINUOUS VERSION] Let U be an open set in $\mathbb{R}^{p+n}$ and $(\boldsymbol{a},\boldsymbol{b}) \in U$. Let $\boldsymbol{F} : U \to \mathbb{R}^n$ with $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is continuous on U and that $D_{\boldsymbol{y}}\boldsymbol{F}$ exists and is continuous on U, i.e. that $\frac{\partial \boldsymbol{F}}{\partial y_j}$ exists and is continuous on U for $1 \le j \le n$. If $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible, i.e. $det(D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})) \neq 0$, then there are:

(i) An open set V in $\mathbb{R}^{p+n}$ containing $(\boldsymbol{a},\boldsymbol{b})$, with $V \subseteq U$, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$ is invertible for all $(\boldsymbol{x},\boldsymbol{y}) \in V$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $\boldsymbol{f} : W \to \mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{b}$ and, for all $\boldsymbol{x} \in W$, $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{0} \ .$$

Additionally, $\boldsymbol{f}$ is continuous on W.

VIII.5.1.7. THEOREM. [IMPLICIT FUNCTION THEOREM, SMOOTH VERSION] Let U be an open set in $\mathbb{R}^{p+n}$ and $(\boldsymbol{a},\boldsymbol{b}) \in U$. Let $\boldsymbol{F} : U \to \mathbb{R}^n$ with $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is $\mathcal{C}^r$ on U for some r, $1 \le r \le \infty$. If $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible, i.e. $det(D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})) \neq 0$, then there are:

(i) An open set V in $\mathbb{R}^{p+n}$ containing $(\boldsymbol{a},\boldsymbol{b})$, with $V \subseteq U$, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$ is invertible for all $(\boldsymbol{x},\boldsymbol{y}) \in V$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $\boldsymbol{f} : W \to \mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{b}$ and, for all $\boldsymbol{x} \in W$, $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{0} \ .$$

Additionally, $\boldsymbol{f}$ is $\mathcal{C}^r$ on W, and for any $\boldsymbol{x} \in W$, we have

$$D\boldsymbol{f}(\boldsymbol{x}) = -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$$

(in particular, $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ is invertible for all $\boldsymbol{x} \in W$).

Slightly informally, for $\boldsymbol{x}$ sufficiently near $\boldsymbol{a}$, $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ is the unique $\boldsymbol{y}$ near $\boldsymbol{b}$ such that $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}$, i.e. we have "solved" the equation $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}$ for $\boldsymbol{y}$ as a function of $\boldsymbol{x}$.

VIII.5.1.8. Note that the difference between the continuous version and the $\mathcal{C}^1$ version (smooth version with $r = 1$) is merely in the differentiability of $\boldsymbol{F}$ in the $\boldsymbol{x}$-coordinates; continuity of $\boldsymbol{F}$ and continuous differentiability in the $\boldsymbol{y}$-coordinates are necessary in both cases (there is a version with this weakened: see VIII.5.3.9. and VIII.5.7.7.). To obtain continuous differentiability of higher order for the implicitly defined function, higher-order continuous differentiability of $\boldsymbol{F}$ in both $\boldsymbol{x}$- and $\boldsymbol{y}$-coordinates is needed.

VIII.5.1.9. The uniqueness of the implicitly defined function $\boldsymbol{f}$ precisely means that there is an open neighborhood V' of $(\boldsymbol{a},\boldsymbol{b})$ (specifically, $V' = \{(\boldsymbol{x},\boldsymbol{y}) \in V : \boldsymbol{x} \in W\}$) such that the part of the level set $\Gamma = \{(\boldsymbol{x},\boldsymbol{y}) \in U : \boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}\}$ in V' is exactly the graph of $\boldsymbol{f}$, i.e. Γ is locally the graph of a (continuous or smooth) function (cf. VIII.5.1.17.).

The use of $\boldsymbol{0}$ in the statements is arbitrary: it could be replaced by any constant $\boldsymbol{c} \in \mathbb{R}^n$ by applying the theorems to $\tilde{\boldsymbol{F}} = \boldsymbol{F} - \boldsymbol{c}$.

VIII.5.1.10. EXAMPLE. [?] To see how this works in practice, consider the following set of three equations in five variables x, y, u, v, w:

$$u^5 + xv^2 - y + w = 0$$
$$v^5 + yu^2 - x + w = 0$$
$$w^4 + y^5 - x^4 = 1$$

We want to solve these equations for u, v, w as functions of x and y near $(1,1)$ with $u(1,1) = 1$, $v(1,1) = 1$, $w(1,1) = -1$ [note that $(x, y, u, v, w) = (1, 1, 1, 1, -1)$ satisfies the equations] and compute the partial derivatives of u, v, w at $(1,1)$.

We will apply the Implicit Function Theorem VIII.5.1.7. with $p = 2$, $n = 3$ (x, y play the role of x_1, x_2 and u, v, w play the role of y_1, y_2, y_3). Define $\boldsymbol{F} : \mathbb{R}^5 \to \mathbb{R}^3$ by

$$\boldsymbol{F}(x, y, u, v, w) = (u^5 + xv^2 - y - w, v^5 + yu^2 - x + w, w^4 + y^5 - x^4 - 1)$$

and note that $\boldsymbol{F}$ is $\mathcal{C}^\infty$ on all of $\mathbb{R}^5$, so the differentiability hypothesis is satisfied on $U = \mathbb{R}^5$. We set $\boldsymbol{a} = (1,1)$, $\boldsymbol{b} = (1,1,-1)$, so $\boldsymbol{c} := (\boldsymbol{a}, \boldsymbol{b}) = (1,1,1,1,-1)$. Set

$$\Gamma = \{(x, y, u, v, w) : \boldsymbol{F}(x, y, u, v, w) = \boldsymbol{0}\}$$

and note that $\boldsymbol{c} \in \Gamma$.

We want to show there is a neighborhood V of $\boldsymbol{c}$ and a unique differentiable ($\mathcal{C}^\infty$) function $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^3$, defined on a neighborhood W of $(1,1)$, $\boldsymbol{f}(x, y) = (u(x, y), v(x, y), w(x, y))$, with $\boldsymbol{f}(1,1) = (1,1,-1)$ and

$$(x, y, u(x, y), v(x, y), w(x, y)) \in \Gamma \cap V$$

for all $(x, y) \in W$, i.e. $\boldsymbol{F}(x, y, u(x, y), v(x, y), w(x, y)) = \boldsymbol{0}$ for all $(x, y) \in W$.

We have

$$D\boldsymbol{F}(x, y, u, v, w) = \begin{bmatrix} v^2 & -1 & 5u^4 & 2xv & 1 \\ -1 & u^2 & 2yu & 5v^4 & 1 \\ -4x^3 & 5y^4 & 0 & 0 & 4w^3 \end{bmatrix}.$$

Partition this matrix after the second column:

$$D\boldsymbol{F}(x, y, u, v, w) = \left[\begin{array}{cc:ccc} v^2 & -1 & 5u^4 & 2xv & 1 \\ -1 & u^2 & 2yu & 5v^4 & 1 \\ -4x^3 & 5y^4 & 0 & 0 & 4w^3 \end{array}\right]$$

to obtain

$$D_{(x,y)}\boldsymbol{F}(x, y, u, v, w) = \begin{bmatrix} v^2 & -1 \\ -1 & u^2 \\ -4x^3 & 5y^4 \end{bmatrix}$$

$$D_{(u,v,w)}\boldsymbol{F}(x, y, u, v, w) = \begin{bmatrix} 5u^4 & 2xv & 1 \\ 2yu & 5v^4 & 1 \\ 0 & 0 & 4w^3 \end{bmatrix}.$$

Thus

$$D\boldsymbol{F}(1,1,1,1,-1) = \begin{bmatrix} 1 & -1 & 5 & 2 & 1 \\ -1 & 1 & 2 & 5 & 1 \\ -4 & 5 & 0 & 0 & -4 \end{bmatrix}$$

$$D_{(x,y)}\boldsymbol{F}(1,1,1,1,-1) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ -4 & 5 \end{bmatrix}.$$

$$D_{(u,v,w)}\boldsymbol{F}(1,1,1,1,-1) = \begin{bmatrix} 5 & 2 & 1 \\ 2 & 5 & 1 \\ 0 & 0 & -4 \end{bmatrix}.$$

We have that $det(D_{(u,v,w)}\boldsymbol{F}(1,1,1,1,-1)) = -84 \neq 0$, so this matrix is invertible and the hypotheses of the Implicit Function Theorem are satisfied. Thus the desired function $\boldsymbol{f}$ exists (and is $\mathcal{C}^\infty$).

We have

$$D\boldsymbol{f}(1,1) = \begin{bmatrix} 5 & 2 & 1 \\ 2 & 5 & 1 \\ 0 & 0 & -4 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ -4 & 5 \end{bmatrix}$$

$$= \begin{bmatrix} 5/21 & -2/21 & 1/28 \\ -2/21 & 5/21 & 1/28 \\ 0 & 0 & -1/4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ -4 & 5 \end{bmatrix} = \begin{bmatrix} 4/21 & -13/84 \\ -10/21 & -11/12 \\ 1 & -5/4 \end{bmatrix}.$$

Thus, for example, $\frac{\partial v}{\partial y}(1,1) = -\frac{11}{12}$.

If (x_0, y_0) is near $(1,1)$, we could compute $D\boldsymbol{f}(x_0, y_0)$ the same way:

$$D\boldsymbol{f}(x_0,y_0) = \begin{bmatrix} 5u(x_0,y_0)^4 & 2x_0v(x_0,y_0) & 1 \\ 2y_0u(x_0,y_0) & 5v(x_0,y_0)^4 & 1 \\ 0 & 0 & 4w(x_0,y_0)^3 \end{bmatrix}^{-1} \begin{bmatrix} v(x_0,y_0)^2 & -1 \\ -1 & u(x_0,y_0)^2 \\ -4x_0^3 & 5y_0^4 \end{bmatrix}$$

but we have no effective way to compute $u(x_0, y_0)$, $v(x_0, y_0)$, $w(x_0, y_0)$, so this is of limited use. If we know $(x_0, y_0, u_0, v_0, w_0)$ is a point of Γ close to $\boldsymbol{c}$, we could put in these values u_0, v_0, w_0 for $u(x_0, y_0)$, $v(x_0, y_0)$, $w(x_0, y_0)$ and evaluate $D\boldsymbol{f}(x_0, y_0)$. (We have glossed over the difficult question just *how* close $(x_0, y_0, u_0, v_0, w_0)$ must be to $\boldsymbol{c}$ for this to work.)

We could have solved for other variables. For example, we could (in principle) solve for x, y, u as functions of v and w around $(1, -1)$. To verify this, we must show that

$$D_{(x,y,u)}\boldsymbol{F}(1,1,1,1,-1) = \begin{bmatrix} 1 & -1 & 5 \\ -1 & 1 & 2 \\ -4 & 5 & 0 \end{bmatrix}$$

is invertible, which it is (the determinant is -7), or for x, y, w as functions of u, v around $(1, 1)$, since

$$D_{(x,y,w)}\boldsymbol{F}(1,1,1,1,-1) = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & 1 \\ -4 & 5 & -4 \end{bmatrix}$$

is invertible (the determinant is -2). The reader is invited to check the other seven possibilities.

VIII.5.1.11. EXAMPLE. In the statement of the theorems, one would expect the neighborhood V to be contained in U (this of course does not preclude the possibility that the conclusion could hold on a larger V). To show the necessity in general of passing to a smaller neighborhood V of $(\boldsymbol{a}, \boldsymbol{b})$ as well as the (expected) passage to a small neighborhood W of $\boldsymbol{a}$, consider the basic example of implicit differentiation every calculus class begins with: the case $p = n = 1$ and

$$F(x, y) = x^2 + y^2 - 25 \ .$$

The set $\{(x, y) : F(x, y) = 0\}$ is, of course, the circle of radius 5 around the origin in $\mathbb{R}^2$. Let $(a, b) = (3, 4)$. In this case the set U could be taken to be all of $\mathbb{R}^2$, but no matter how small an open interval W around $a = 3$ in $\mathbb{R}$ is taken, there is not a unique function f on W giving a solution, since we could have $f(x) = \pm\sqrt{25 - x^2}$ with the sign chosen arbitrarily for each x (so long as $f(3) = 4$). There is a unique *continuous solution* for W small, but to obtain a unique *solution*, one must cut down to a smaller neighborhood V of $(3, 4)$. Actually V does not have to be all that small in this case: it only needs not to contain any points on the lower half-circle with x-coordinate close to 3.

In this case, if f is the unique solution on an interval W (which must be contained in the interval $(-5, 5)$), then by the theorem

$$f'(x) = -\frac{\frac{\partial F}{\partial x}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))} = -\frac{2x}{2f(x)} = -\frac{x}{f(x)}$$

for any $x \in W$. Here we can find an explicit formula for f: $f(x) = \sqrt{25 - x^2}$, and verify the formula for f' directly. In most examples it is impossible to find an explicit formula for the implicitly defined function.

VIII.5.1.12. EXAMPLE. For a simpler and more dramatic example showing that V must be chosen smaller than U in general, let $F(x, y) = \sin y$, with $U = \mathbb{R}^2$. The set

$$\Gamma = \{(x, y) : F(x, y) = 0\}$$

consists of the union of the horizontal lines $y = n\pi$ for $n \in \mathbb{Z}$. $(0, 0)$ is a point of Γ, and within the horizontal strip

$$V = \{(x, y) : -\pi < y < \pi\}$$

the equation $F(x, y) = 0$ can be solved for y as a (constant) function of x, but not in all of U. (To satisfy conclusion (i) of the theorem, we must further reduce V to

$$\left\{(x, y) : -\frac{\pi}{2} < y < \frac{\pi}{2}\right\}$$

to have $\frac{\partial F}{\partial y} \neq 0$ on V.) In fact, the equation $F(x, y) = 0$ can be locally solved for y as a function of x around every point of Γ, but not globally on Γ.

VIII.5.1.13. EXAMPLE. In Example VIII.5.1.12., the zero set Γ is naturally a disjoint union of graphs of functions. But the situation can be more complicated in higher dimensions. Let

$$U = \{(x, y, z) : x^2 + y^2 > 0\}$$

be $\mathbb{R}^3$ with the z-axis removed, and define ϕ and F on U by

$$\phi(x,y,z) = \begin{cases} z - \tan^{-1}\frac{y}{x} & \text{if} \quad x > 0 \\ z - \frac{\pi}{2} & \text{if} \quad x = 0,\ y > 0 \\ z - \tan^{-1}\frac{y}{x} - \pi & \text{if} \quad x < 0 \\ z + \frac{\pi}{2} & \text{if} \quad x = 0,\ y < 0 \end{cases}$$

and $F(x,y,z) = \sin\phi(x,y,z)$. Then ϕ is discontinuous where $x = 0$ and $y < 0$, but F is $\mathcal{C}^\infty$ on U. If $\Gamma = \{(x,y,z) \in U : F(x,y,z) = 0\}$, then $\frac{\partial F}{\partial z} \neq 0$ at all points of Γ, so the equation $F(x,y,z) = 0$ can be (even explicitly) locally solved for z at every point of Γ, but not globally (this phenomenon is called *holonomy* ()). The surface Γ consists of two intertwined helicoids, the surfaces with equations $z = \theta$ and $z = \theta + \pi$ in cylindrical coordinates, intersected with U. Neither helicoid is globally the graph of a function.

VIII.5.1.14. To obtain a reasonable Implicit Function Theorem, the function $\boldsymbol{F}$ must be continuous. This in itself is not enough to guarantee an implicitly defined function, even a nonunique one, even if $p = n = 1$ (Exercise VIII.5.11.4.). But if there is a (unique) implicitly defined function, it is automatically continuous:

VIII.5.1.15. Proposition. Let F be a continuous function on a subset U of $\mathbb{R}^{p+n}$, with values in $\mathbb{R}^n$, and K a compact subset of U. Suppose there is an open set W in $\mathbb{R}^p$ on which there is a unique function $\boldsymbol{f} : W \to \mathbb{R}^n$ implicitly defined by the equation $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}$ with respect to K, i.e. for each $\boldsymbol{x} \in W$ there is a unique $\boldsymbol{y} \in \mathbb{R}^n$ such that $(\boldsymbol{x},\boldsymbol{y}) \in K$ and $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}$. Then $\boldsymbol{f}$ is continuous.

Proof: Let $(\boldsymbol{x}_k)$ be a sequence in W with $\boldsymbol{x}_k \to \boldsymbol{a} \in W$. Suppose $\boldsymbol{f}(\boldsymbol{x}_k) \not\to \boldsymbol{f}(\boldsymbol{a})$. Then by compactness of K there is a subsequence $(\boldsymbol{x}_{k_j})$ such that $(\boldsymbol{f}(\boldsymbol{x}_{k_j}))$ converges to a point $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{y} \neq \boldsymbol{f}(\boldsymbol{a})$, with $(\boldsymbol{a},\boldsymbol{y}) \in K$. But then $\boldsymbol{0} = \boldsymbol{F}(\boldsymbol{x}_{k_j}, \boldsymbol{f}(\boldsymbol{x}_{k_j})) \to \boldsymbol{F}(\boldsymbol{a},\boldsymbol{y})$ by continuity of $\boldsymbol{F}$, so $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{y}) = \boldsymbol{F}(\boldsymbol{a},\boldsymbol{f}(\boldsymbol{a})) = \boldsymbol{0}$, contradicting the uniqueness assumption on $\boldsymbol{F}$ and K. 👍

VIII.5.1.16. The Implicit Function Theorem is strictly a local result, i.e. the conclusion just holds on *some* neighborhood W of $\boldsymbol{a}$. If $D\boldsymbol{F}$ is locally Lipschitz, in particular if $\boldsymbol{F}$ is $\mathcal{C}^2$, it is possible to determine an explicit W on which the conclusion holds (cf. XI.1.10.12.).

Geometric Interpretation of the Implicit Function Theorem

We can interpret the result of the Implicit Function Theorem geometrically as showing that the level set of $\boldsymbol{F}$ through $(\boldsymbol{a},\boldsymbol{b})$ is locally a "smooth hypersurface":

VIII.5.1.17. Corollary. Let U be an open set in $\mathbb{R}^{p+n}$ and $(\boldsymbol{a},\boldsymbol{b}) \in U$. Let $\boldsymbol{F} : U \to \mathbb{R}^n$ with $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is $\mathcal{C}^r$ on U for some r, $1 \leq r \leq \infty$. Set

$$\Gamma = \{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^{p+n} : \boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}\}\ .$$

If $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible, i.e. $det(D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})) \neq 0$, then there is an open neighborhood V of $(\boldsymbol{a},\boldsymbol{b})$ in $\mathbb{R}^{p+n} \cong \mathbb{R}^p \times \mathbb{R}^n$ such that $\Gamma \cap V$ is the graph of a $\mathcal{C}^r$ function from an open set W in $\mathbb{R}^p$ to $\mathbb{R}^n$. Thus $\Gamma \cap V$ is a p-dimensional $\mathcal{C}^r$ submanifold of $\mathbb{R}^{p+n}$ (XIV.1.1.12.).

The result can be rephrased more symmetrically if it is not specified in advance which n coordinates are being locally solved for:

VIII.5.1.18. Corollary. Let U be an open set in $\mathbb{R}^{p+n}$ and $\boldsymbol{c} \in U$. Let $\boldsymbol{F} : U \to \mathbb{R}^n$ with $\boldsymbol{F}(\boldsymbol{c}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is $\mathcal{C}^r$ on U for some r, $1 \leq r \leq \infty$. Set

$$\Gamma = \{\boldsymbol{z} \in \mathbb{R}^{p+n} : \boldsymbol{F}(\boldsymbol{z}) = \boldsymbol{0}\} .$$

If $D\boldsymbol{F}(\boldsymbol{c})$ has rank n, then there is an open neighborhood V of $\boldsymbol{c}$ in $\mathbb{R}^{p+n}$ such that $\Gamma \cap V$ is a p-dimensional $\mathcal{C}^r$ submanifold of $\mathbb{R}^{p+n}$. If $D\boldsymbol{F}(\boldsymbol{z})$ has rank n at every $\boldsymbol{z} \in \Gamma$, then Γ is a p-dimensional $\mathcal{C}^r$ submanifold of $\mathbb{R}^{p+n}$.

Note that since $D\boldsymbol{F}(\boldsymbol{c})$ is $n \times (p+n)$, its rank must be $\leq n$; thus the hypothesis is that $D\boldsymbol{F}(\boldsymbol{c})$ has maximal rank possible. If $n = 1$, the condition is just that $DF(\boldsymbol{c}) \neq 0$.

Proof: The hypothesis implies (and is equivalent to) that there is a set $\{x_{j_1}, \ldots, x_{j_n}\}$ of n coordinates such that the corresponding columns of $D\boldsymbol{F}(\boldsymbol{c})$ are linearly independent. By reordering the coordinates we may make these coordinates be the last n and rename them $y_1, \ldots, y_n$. Then with the renamed coordinates, $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{c})$ is invertible, so VIII.5.1.17. applies to give the result. 👍

VIII.5.1.19. There is often more than one choice for the n coordinates to be solved for, although not every set of n coordinates can be solved for in general. For example, in the situation of VIII.5.1.11., if $\boldsymbol{c}$ is any point on the level curve $F(x,y) = 0$, i.e.

$$\Gamma = \{(x,y) : x^2 + y^2 = 25\}$$

except for one of the four points on the coordinate axes, one can implicitly (and, in this case, even explicitly) locally solve for either of x and y as a function of the other. In this case, $DF(x,y)$ has rank 1 (i.e. is nonzero) at every point of Γ, so the conclusion is that Γ is a smooth 1-manifold (an obvious fact in this case).

VIII.5.1.20. If $\boldsymbol{c}$ is a point of Γ where $D\boldsymbol{F}(\boldsymbol{c})$ has rank less than n, then Γ is not a smooth manifold around $\boldsymbol{c}$ in general, and the local structure can be (but is not necessarily) quite pathological (VIII.5.1.22., VIII.5.1.23., VIII.5.11.23., VIII.5.11.1.). Indeed, if $F(x,y) = x^2 + y^2$, the set $\{(x,y) : F(x,y) = 0\}$ is a single point!

VIII.5.1.21. In fact, under the hypotheses of VIII.5.1.17. or VIII.5.1.18., one can be more specific about the local structure of Γ (VIII.5.8.8.).

VIII.5.1.22. EXAMPLE. Define $F : \mathbb{R}^2 \to \mathbb{R}$ (i.e. $p = n = 1$) by

$$F(x, y) = x^3 - 2xy + y^3$$

and set $\Gamma = \{(x, y) : F(x, y) = 0\}$. Then Γ is a "curve" in $\mathbb{R}^2$ (a *folium of Descartes*). We have

$$DF(x, y) = \begin{bmatrix} 3x^2 - 2y & 3y^2 - 2x \end{bmatrix}$$

which is nonzero at all points of Γ except $\mathbf{0}$; thus Γ is indeed a smooth curve except at the origin. See Figure VIII.14.

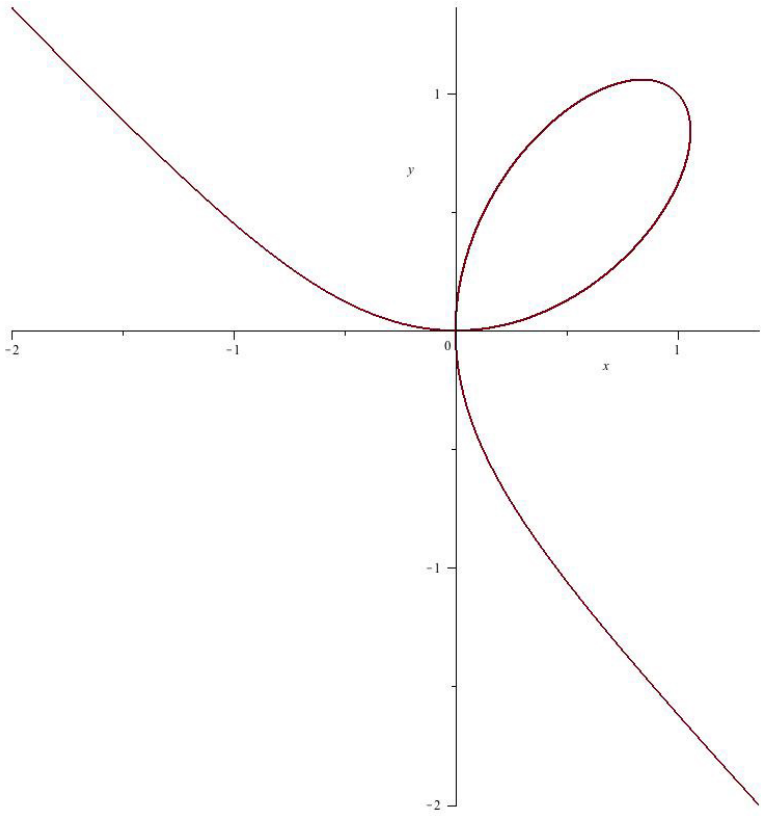


Figure VIII.14: Maple plot of $x^3 - 2xy + y^3 = 0$

The curve can be locally solved for y as a function of x around any point (a, b) on Γ where $3b^2 - 2a \neq 0$, i.e. everywhere except at the origin and the point in the first quadrant where the tangent line is vertical, $\left(\frac{2^{5/3}}{3}, \frac{2^{4/3}}{3}\right) \approx (1.058, 0.84)$. The slope of the tangent line at (a, b) is $-\frac{3a^2-2b}{3b^2-2a}$. In particular, at $(1, 1)$ the tangent line has slope -1.

VIII.5.1.23. EXAMPLE. Define $F : \mathbb{R}^3 \to \mathbb{R}$ (i.e. $p = 2$, $n = 1$) by

$$F(x, y, z) = x^2 + y^2 - z^2$$

and set $\Gamma = \{(x, y, z) : F(x, y, z) = 0\}$. Then Γ is a "surface" in $\mathbb{R}^2$; in fact, Γ is a cone with vertex at the origin (Figure VIII.15). We have

$$DF(x, y, z) = \begin{bmatrix} 2x & 2y & -2z \end{bmatrix}$$

which is rank 1 (nonzero) everywhere except at $\mathbf{0}$. Thus the Implicit Function Theorem implies that the cone is a smooth surface except possibly at the origin (which is geometrically obvious). It is in fact not a surface in any neighborhood of $\mathbf{0}$. If u stands for x, y, or z, then the equation $F(x, y, z) = 0$ can be locally solved for u as a function of the other two variables in a neighborhood of any point where $u \neq 0$.

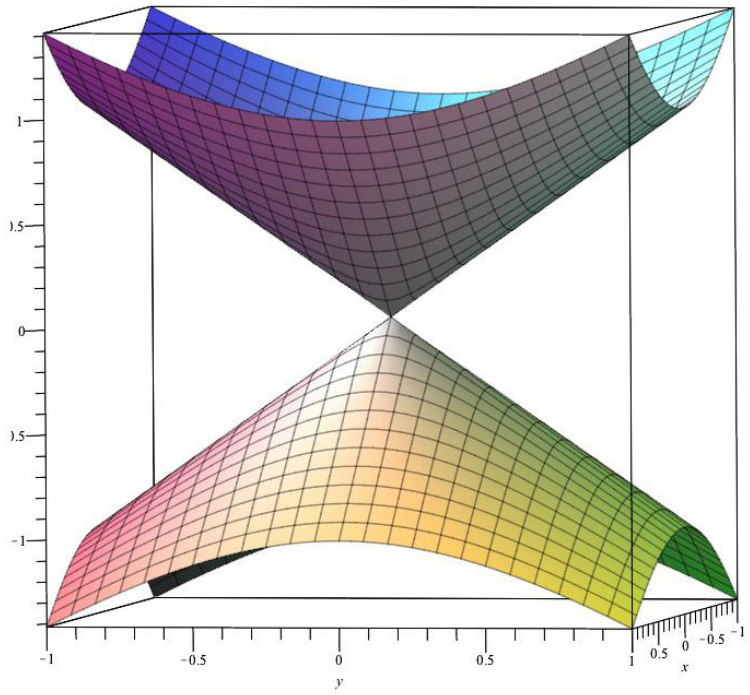


Figure VIII.15: Maple plot of $x^2 + y^2 = z^2$

VIII.5.1.24. The result VIII.5.1.18. can be used also to analyze intersections of surfaces or hypersurfaces. We illustrate with the case of two surfaces in $\mathbb{R}^3$, but the same principles apply to the case of k hypersurfaces in $\mathbb{R}^{p+n}$ for any $k, n, p \in \mathbb{N}$, $k \leq n$ (Exercise VIII.5.11.24.).

Suppose F_1, F_2 are smooth ($\mathcal{C}^r$) functions from $\mathbb{R}^3$ to $\mathbb{R}$, and set $\Gamma_1 = \{(x, y, z) : F_1(x, y, z) = 0\}$, $\Gamma_2 = \{(x, y, z) : F_2(x, y, z) = 0\}$. Then Γ_1 and Γ_2 are smooth surfaces, at least where DF_1 and DF_2 are nonzero. We would expect that $\Gamma = \Gamma_1 \cap \Gamma_2$ should be a smooth curve, at least where the surfaces intersect transversally, i.e. where DF_1 and DF_2 satisfy some nondegeneracy condition.

The proper condition is most easily described by considering DF_1 and DF_2 as ∇F_1 and ∇F_2, interpreting them as vectors in $\mathbb{R}^3$. The result is:

VIII.5.1.25. Corollary. Let F_1 and F_2 be $\mathcal{C}^r$ functions ($1 \leq r \leq \infty$) from $\mathbb{R}^3$ to $\mathbb{R}$, $\Gamma_1 = \{(x, y, z) : F_1(x, y, z) = 0\}$, $\Gamma_2 = \{(x, y, z) : F_2(x, y, z) = 0\}$, $\Gamma = \Gamma_1 \cap \Gamma_2$. Let $\boldsymbol{c} \in \Gamma$. If $\nabla F_1(\boldsymbol{c})$ and $\nabla F_2(\boldsymbol{c})$ are linearly independent, then there is an open neighborhood V of $\boldsymbol{c}$ in $\mathbb{R}^3$ such that $\Gamma \cap V = (\Gamma_1 \cap V) \cap (\Gamma_2 \cap V)$ is a one-dimensional $\mathcal{C}^r$ submanifold of $\mathbb{R}^3$ (i.e. a smooth curve through $\boldsymbol{c}$). If $\nabla F_1(\boldsymbol{x})$ and $\nabla F_2(\boldsymbol{x})$ are linearly independent for all $\boldsymbol{x} \in \Gamma$, then Γ is a one-dimensional $\mathcal{C}^r$ submanifold of $\mathbb{R}^3$.

Note that the hypotheses imply that Γ_1 and Γ_2 are smooth surfaces in a neighborhood of $\boldsymbol{c}$. The condition is then that the surfaces are not tangent at $\boldsymbol{c}$ (i.e. have distinct tangent planes).

Proof: Define $\boldsymbol{F} : \mathbb{R}^3 \to \mathbb{R}^2$ by $\boldsymbol{F} = (F_1, F_2)$. The condition that $\nabla F_1(\boldsymbol{c})$ and $\nabla F_2(\boldsymbol{c})$ are linearly independent is equivalent to the statement that $D\boldsymbol{F}(\boldsymbol{c})$ has rank two. Thus VIII.5.1.18. applies. 👍

VIII.5.1.26. Under the hypotheses, Γ_1 and Γ_2 intersect transversally at $\boldsymbol{c}$ in the sense of ().

The Derivative Formula

VIII.5.1.27. There is wide variation in the literature about the importance of the formula for the derivative in the Implicit Function Theorem (there is more uniform respect for it in the Inverse Function Theorem). Some texts like [**?**] do not mention it at all, whereas in [HH99] we read, "What would we do with an implicit function if we didn't know how to differentiate it?" The formula is more central to the "one-variable" case, the only case done in some references, in the guise of justifying implicit differentiation as done in first-semester calculus. We take a somewhat middle attitude: while the derivative formula is unquestionably important, the essence of the Implicit/Inverse Function Theorem is the *existence* of the implicit/inverse function and the fact that it is unique and $\mathcal{C}^r$. Once we know that, the derivative formula is just a straightforward application of the Chain Rule.

The Inverse Function Theorem

VIII.5.1.28. If $\boldsymbol{f}$ is a function defined on an open subset U of $\mathbb{R}^n$, with values in $\mathbb{R}^m$, we would like to know when $\boldsymbol{f}$ has an inverse function with the same continuity or differentiability properties as $\boldsymbol{f}$. An obvious necessary condition is that $\boldsymbol{f}$ be one-to-one (injective) on U, and to discuss differentiability we also need that $\boldsymbol{f}(U)$ is open in $\mathbb{R}^m$.

The Invariance of Domain Theorem () gives a complete answer to the continuous version of the question: if $\boldsymbol{f}$ is continuous and one-to one, then $\boldsymbol{f}(U)$ is open if and only if $n = m$, and in this case $\boldsymbol{f}^{-1}$ is automatically continuous on $\boldsymbol{f}(U)$.

The answer to the differentiability question is more elementary. The next result does not depend on the Invariance of Domain Theorem:

VIII.5.1.29. PROPOSITION. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in U$, V an open set in $\mathbb{R}^m$, and $\boldsymbol{f}$ a one-to-one function from U onto V. Set $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$. Suppose $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$. Then $\boldsymbol{f}^{-1} : V \to U$ is differentiable at $\boldsymbol{b}$ if and only if $m = n$ and $D\boldsymbol{f}(\boldsymbol{a})$ is invertible. If these conditions hold, then $D(\boldsymbol{f}^{-1})(\boldsymbol{b}) = [D\boldsymbol{f}(\boldsymbol{a})]^{-1}$.

PROOF: One direction is almost immediate. If $\boldsymbol{f}^{-1}$ is differentiable at $\boldsymbol{b}$, then by the Chain Rule we have

$$I_n = D(\boldsymbol{f}^{-1} \circ \boldsymbol{f})(\boldsymbol{a}) = D(\boldsymbol{f}^{-1})(\boldsymbol{b}) D\boldsymbol{f}(\boldsymbol{a})$$

$$I_m = D(\boldsymbol{f} \circ \boldsymbol{f}^{-1})(\boldsymbol{b}) = D\boldsymbol{f}(\boldsymbol{a}) D(\boldsymbol{f}^{-1})(\boldsymbol{b})$$

and so $m = n$, $D\boldsymbol{f}(\boldsymbol{a})$ is invertible, and $D(\boldsymbol{f}^{-1})(\boldsymbol{b}) = D\boldsymbol{f}(\boldsymbol{a})^{-1}$ by (). 👍

For the converse direction, we need a lemma:

VIII.5.1.30. LEMMA. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in U$, and $\boldsymbol{f}$ a one-to-one function from U to $\mathbb{R}^n$. Suppose $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$ and $D\boldsymbol{f}(\boldsymbol{a})$ is invertible. Then there are strictly positive constants c, d and a neighborhood W of $\boldsymbol{a}$ contained in U, such that

$$c\|\boldsymbol{x} - \boldsymbol{a}\| < \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\| < d\|\boldsymbol{x} - \boldsymbol{a}\|$$

for all $\boldsymbol{x} \in W \setminus \{\boldsymbol{a}\}$. In fact, we may choose c to be any positive number less than $\|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|^{-1}$ and d any number greater than $\|D\boldsymbol{f}(\boldsymbol{a})\|$.

PROOF: Fix c, d with $0 < c < \|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|^{-1}$ and $d > \|D\boldsymbol{f}(\boldsymbol{a})\|$, and set

$$\epsilon = \min\left(\|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|^{-1} - c, d - \|D\boldsymbol{f}(\boldsymbol{a})\|\right) > 0 \ .$$

Since $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$, there is a $\delta > 0$ such that, whenever $0 < \|\boldsymbol{x} - \boldsymbol{a}\| < \delta$, $\boldsymbol{x} \in U$ and

$$\left| \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} - \frac{\|D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} \right| \leq \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a}) - D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} < \epsilon$$

$$-\epsilon < \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} - \frac{\|D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} < \epsilon \ .$$

But we have $\|D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\| \leq \|D\boldsymbol{f}(\boldsymbol{a})\|\|\boldsymbol{x} - \boldsymbol{a}\|$ by definition of the operator norm, so

$$\frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} < \frac{\|D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} + \epsilon \leq \|D\boldsymbol{f}(\boldsymbol{a})\| + \epsilon \leq d$$

and similarly

$$\|\boldsymbol{x} - \boldsymbol{a}\| = \|D\boldsymbol{f}(\boldsymbol{a})^{-1}(D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}))\| \leq \|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|\|D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|$$

$$c \leq \|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|^{-1} - \epsilon \leq \frac{\|D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} - \epsilon < \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|}$$

so we may take W to be the open ball of radius δ around $\boldsymbol{a}$. 👍

VIII.5.1.31. We now prove the other direction of VIII.5.1.29.. Suppose $m = n$ and $D\boldsymbol{f}(\boldsymbol{a})$ is invertible. Since f is a bijection between U and V, for each $\boldsymbol{x} \in U$ there is a unique $\boldsymbol{y} \in V$ with $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, and conversely for each $\boldsymbol{y} \in V$ there is a unique $\boldsymbol{x} \in U$ with $\boldsymbol{x} = \boldsymbol{f}^{-1}(\boldsymbol{y})$. We will suppress the function names and just use $\boldsymbol{x}$ and $\boldsymbol{y}$ with the understanding that they are related in this manner.

We must show that for any $\epsilon > 0$, there is a $\delta > 0$ such that, whenever $0 < \|\boldsymbol{y} - \boldsymbol{b}\| < \delta$, $\boldsymbol{y} \in V$ and

$$\frac{\|\boldsymbol{x} - \boldsymbol{a} - D\boldsymbol{f}(\boldsymbol{a})^{-1}(\boldsymbol{y} - \boldsymbol{b})\|}{\|\boldsymbol{y} - \boldsymbol{b}\|} < \epsilon \ .$$

By VIII.5.1.30., reducing U if necessary we may assume there are positive constants c and d such that

$$c\|\boldsymbol{x} - \boldsymbol{a}\| < \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a})\| < d\|\boldsymbol{x} - \boldsymbol{a}\|$$

$$\frac{1}{d} < \frac{\|\boldsymbol{x} - \boldsymbol{a}\|}{\|\boldsymbol{y} - \boldsymbol{b}\|} < \frac{1}{c}$$

for all $\boldsymbol{x} \in U$, $\boldsymbol{x} \neq \boldsymbol{a}$. Since $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$, there is an $\eta > 0$ such that, whenever $0 < \|\boldsymbol{x} - \boldsymbol{a}\| < \eta$, $\boldsymbol{x} \in U$ and

$$\frac{\|\boldsymbol{y} - \boldsymbol{b} - D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} < \frac{\epsilon c}{\|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|} \ .$$

Since

$$\boldsymbol{x} - \boldsymbol{a} - D\boldsymbol{f}(\boldsymbol{a})^{-1}(\boldsymbol{y} - \boldsymbol{b}) = D\boldsymbol{f}(\boldsymbol{a})^{-1}[D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) - (\boldsymbol{y} - \boldsymbol{b})]$$

we have, for $\|\boldsymbol{y} - \boldsymbol{b}\| < c\eta$, $\|\boldsymbol{x} - \boldsymbol{a}\| < \eta$ and

$$\frac{\|\boldsymbol{x} - \boldsymbol{a} - D\boldsymbol{f}(\boldsymbol{a})^{-1}(\boldsymbol{y} - \boldsymbol{b})\|}{\|\boldsymbol{y} - \boldsymbol{b}\|} \leq \frac{\|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|\|D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) - (\boldsymbol{y} - \boldsymbol{b})\|}{\|\boldsymbol{y} - \boldsymbol{b}\|}$$

$$= \frac{\|\boldsymbol{x} - \boldsymbol{a}\|}{\|\boldsymbol{y} - \boldsymbol{b}\|}\|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|\frac{\|\boldsymbol{y} - \boldsymbol{b} - D\boldsymbol{f}(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a})\|}{\|\boldsymbol{x} - \boldsymbol{a}\|} < \frac{1}{c}\|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|\frac{\epsilon c}{\|D\boldsymbol{f}(\boldsymbol{a})^{-1}\|} = \epsilon$$

so we may take $\delta = c\eta$. 👍

VIII.5.1.32. Corollary. Let U be an open set in $\mathbb{R}^n$, V an open set in $\mathbb{R}^m$, and $\boldsymbol{f}$ a one-to-one function from U onto V. If $\boldsymbol{f}$ is differentiable at every point of U, then $\boldsymbol{f}^{-1} : V \to U$ is differentiable at every point of V if and only if $m = n$ and, for every $\boldsymbol{x} \in U$, $D\boldsymbol{f}(\boldsymbol{x})$ is invertible. If these conditions hold, then $D(\boldsymbol{f}^{-1})(\boldsymbol{y})$ is invertible for every $\boldsymbol{y} \in V$ and $D(\boldsymbol{f}^{-1})(\boldsymbol{y}) = D\boldsymbol{f}(\boldsymbol{f}^{-1}(\boldsymbol{y}))^{-1}$.

Actually, if $\boldsymbol{f} : U \to V$ is one-to-one, onto, and differentiable, then m must equal n even without the assumption that $D\boldsymbol{f}(\boldsymbol{x})$ is invertible for every $\boldsymbol{x} \in U$, by Invariance of Domain; but we do not need to use this fact for the Corollary.

VIII.5.1.33. There are two principal drawbacks to these results, however. First, the Invariance of Domain is a deep theorem requiring a considerable excursion into topology to prove. And second, even for the differentiable result which is more elementary, it must be verified that the function is locally one-to-one, with open range, which is often not easy to do directly. The most useful result would be one in which (local) injectivity can be deduced simply from nondegeneracy of the derivative. This is the principal content of the Inverse Function Theorem:

VIII.5.1.34. THEOREM. [INVERSE FUNCTION THEOREM] Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in U$, and $\boldsymbol{f} : U \to \mathbb{R}^n$ a $\mathcal{C}^r$ function for some r, $1 \leq r \leq \infty$. If $D\boldsymbol{f}(\boldsymbol{a})$ is invertible, then there is a neighborhood V of $\boldsymbol{a}$, $V \subseteq U$, such that

(i) $\boldsymbol{f}$ is one-to-one on V.

(ii) $\boldsymbol{f}(V)$ is open in $\mathbb{R}^n$.

(iii) $D\boldsymbol{f}(\boldsymbol{x})$ is invertible for all $\boldsymbol{x} \in V$.

(iv) $\boldsymbol{f}^{-1}$ is $\mathcal{C}^r$ on $\boldsymbol{f}(V)$ and, for all $\boldsymbol{y} \in \boldsymbol{f}(V)$,

$$D(\boldsymbol{f}^{-1})(\boldsymbol{y}) = [D\boldsymbol{f}(\boldsymbol{f}^{-1}(\boldsymbol{y}))]^{-1} \ .$$

VIII.5.1.35. If $D\boldsymbol{f}(\boldsymbol{x})$ is not invertible everywhere on U, it will obviously be necessary to restrict to a smaller neighborhood V of $\boldsymbol{a}$ to get (iii) and (iv). But this may be necessary even if $D\boldsymbol{f}(\boldsymbol{x})$ is invertible for all $\boldsymbol{x} \in U$ to make $\boldsymbol{f}$ injective.

If $n = 1$ and I is an interval in $\mathbb{R}$, and $f : I \to R$ is differentiable on I with $f'(x) \neq 0$ for all $x \in I$, then f is strictly monotone on I () and thus one-to-one; so if $U = I$ we may take $V = U$. But if $n > 1$ invertibility of $D\boldsymbol{f}$ on U does not guarantee that $\boldsymbol{f}$ is one-to-one on U, even if U is connected. A simple and important example is the following function, which is the transformation from rectangular coordinates to polar coordinates in $\mathbb{R}^2$ (hence we use (r, θ) for the coordinates). Set $U = \{(r, \theta) : r > 0\}$, and define $\boldsymbol{f} : U \to \mathbb{R}^2$ by

$$\boldsymbol{f}(r, \theta) = (r \cos\theta, r \sin\theta) \ .$$

Then $D\boldsymbol{f}(r, \theta)$ is invertible everywhere on U; but $\boldsymbol{f}$ is not one-to-one on all of U, only locally one-to-one.

VIII.5.2. Twin Theorems and a Generalization

The Inverse Function Theorem and Implicit Function Theorem (smooth version) are twin theorems in the sense that each is easily derivable from the other; the arguments are much easier than proving either theorem from scratch. In fact, either argument can be used to prove a common generalization:

VIII.5.2.1. THEOREM. [GENERAL IMPLICIT/INVERSE FUNCTION THEOREM] Let $n \in \mathbb{N}$, $p \in \mathbb{N} \cup \{0\}$, $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$, U an open neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$ in $\mathbb{R}^{p+n}$, $\boldsymbol{F} : U \to \mathbb{R}^n$ a $\mathcal{C}^r$ function ($1 \leq r \leq \infty$) with $\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{c}$ and $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b})$ invertible. Then there are:

(i) An open neighborhood V of $(\boldsymbol{a}, \boldsymbol{b})$ in $\mathbb{R}^{p+n}$, contained in U, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ is invertible for all $(\boldsymbol{x}, \boldsymbol{y}) \in V$;

(ii) An open neighborhood W of $(\boldsymbol{a}, \boldsymbol{c})$ in $\mathbb{R}^{p+n}$;

(iii) A unique function $\boldsymbol{G} : W \to \mathbb{R}^n$ satisfying $\boldsymbol{G}(\boldsymbol{a}, \boldsymbol{c}) = \boldsymbol{b}$ and, for all $(\boldsymbol{x}, \boldsymbol{z}) \in W$, $(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) = \boldsymbol{z} \ .$$

Additionally, $\boldsymbol{G}$ is $\mathcal{C}^r$ on W, and for all $(\boldsymbol{x},\boldsymbol{z})\in W$,

$$D\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})=\begin{bmatrix} & \vdots & \\ -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z}))^{-1}D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})) & \vdots & D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z}))^{-1} \\ & \vdots & \end{bmatrix}.$$

Slightly informally, for $(\boldsymbol{x},\boldsymbol{z})$ sufficiently near $(\boldsymbol{a},\boldsymbol{c})$, $\boldsymbol{y}=\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})$ is the unique $\boldsymbol{y}$ near $\boldsymbol{b}$ such that $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{z}$, i.e. we have "solved" the equation $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{z}$ for $\boldsymbol{y}$ as a function of $\boldsymbol{x}$ and $\boldsymbol{z}$.

VIII.5.2.2. Note that the Inverse Function Theorem is the case $p=0$ (so the $\boldsymbol{a}$ is absent). The Smooth Implicit Function Theorem is a corollary by taking $\boldsymbol{c}=\boldsymbol{0}$ and setting $\boldsymbol{f}(\boldsymbol{x})=\boldsymbol{G}(\boldsymbol{x},\boldsymbol{0})$. In fact, the Smooth Implicit Function Theorem says that for certain fixed $\boldsymbol{z}$ near $\boldsymbol{c}$, there is a function $\boldsymbol{f}_{\boldsymbol{z}}$ implicitly defined near $\boldsymbol{a}$ by $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{z}$, which is $\mathcal{C}^r$ in $\boldsymbol{x}$; we have $\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})=\boldsymbol{f}_{\boldsymbol{z}}(\boldsymbol{x})$, and the additional content of the above theorem is that the $\boldsymbol{f}_{\boldsymbol{z}}$ are defined for all $\boldsymbol{z}$ near $\boldsymbol{c}$ and "vary smoothly" in $\boldsymbol{z}$. Thus Theorem VIII.5.2.1. is a "parametrized version" of both the Implicit and Inverse Function Theorems: holding $\boldsymbol{x}$ fixed gives the Inverse Function Theorem and holding $\boldsymbol{z}$ fixed gives the (Smooth) Implicit Function Theorem.

The situation is symmetric in $\boldsymbol{F}$ and $\boldsymbol{G}$. $\boldsymbol{G}$ satisfies the hypotheses of the theorem, so there is a $\mathcal{C}^r$ function $\boldsymbol{H}$ defined near $(\boldsymbol{a},\boldsymbol{b})$ for which $\boldsymbol{H}(\boldsymbol{a},\boldsymbol{b})=\boldsymbol{c}$ and $\boldsymbol{G}(\boldsymbol{x},\boldsymbol{H}(\boldsymbol{x},\boldsymbol{y}))=\boldsymbol{y}$ for $(\boldsymbol{x},\boldsymbol{y})$ near $(\boldsymbol{a},\boldsymbol{b})$. By the formula, $D\boldsymbol{H}(\boldsymbol{x},\boldsymbol{y})=D\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$ for all $(\boldsymbol{x},\boldsymbol{y})$ near $(\boldsymbol{a},\boldsymbol{b})$, and since $\boldsymbol{H}(\boldsymbol{a},\boldsymbol{b})=\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$, we have $\boldsymbol{H}=\boldsymbol{F}$ on a neighborhood of $(\boldsymbol{a},\boldsymbol{b})$, i.e. $\boldsymbol{H}=\boldsymbol{F}$ (perhaps restricted to a smaller neighborhood of $(\boldsymbol{a},\boldsymbol{b})$). In particular, we have $\boldsymbol{G}(\boldsymbol{x},\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}))=\boldsymbol{y}$ for $(\boldsymbol{x},\boldsymbol{y})$ near $(\boldsymbol{a},\boldsymbol{b})$ (this can be verified directly).

We will give simple proofs of VIII.5.2.1. from both the Inverse Function Theorem and the Smooth Implicit Function Theorem.

Proof From Inverse Function Theorem

VIII.5.2.3. Assume the Inverse Function Theorem VIII.5.1.34.. We will deduce VIII.5.2.1..

Suppose $U\subseteq\mathbb{R}^{p+n}$ is open, $(\boldsymbol{a},\boldsymbol{b})\in U$, and $\boldsymbol{F}:U\to\mathbb{R}^n$ is a $\mathcal{C}^r$ function with $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})=\boldsymbol{c}$ and $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ invertible. Define

$$\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x},\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}))$$

for $(\boldsymbol{x},\boldsymbol{y})\in U$. Then $\tilde{\boldsymbol{F}}$ is a $\mathcal{C}^r$ function from U to $\mathbb{R}^{p+n}$. We have

$$D\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{y})=\begin{bmatrix} I_p & 0 \\ D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) & D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) \end{bmatrix}$$

for $(\boldsymbol{x},\boldsymbol{y})\in U$; thus $D\tilde{\boldsymbol{F}}(\boldsymbol{a},\boldsymbol{b})$ is invertible (). So by the Inverse Function Theorem there is a neighborhood V of $(\boldsymbol{a},\boldsymbol{b})$ on which $\tilde{\boldsymbol{F}}$ is one-to-one, $W=\tilde{\boldsymbol{F}}(V)$ is open, and the inverse function $\Phi:W\to V$ is $\mathcal{C}^r$. Since $\tilde{\boldsymbol{F}}(\boldsymbol{a},\boldsymbol{b})=(\boldsymbol{a},\boldsymbol{c})$, we have $(\boldsymbol{a},\boldsymbol{c})\in W$ and $\Phi(\boldsymbol{a},\boldsymbol{c})=(\boldsymbol{a},\boldsymbol{b})$.

Write $\Phi(\boldsymbol{x},\boldsymbol{z})=(\boldsymbol{H}(\boldsymbol{x},\boldsymbol{z}),\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z}))$, where $\boldsymbol{H}:W\to\mathbb{R}^p$ and $\boldsymbol{G}:W\to\mathbb{R}^n$. Then $\boldsymbol{H}$ and $\boldsymbol{G}$ are $\mathcal{C}^r$, and $\boldsymbol{G}(\boldsymbol{a},\boldsymbol{c})=\boldsymbol{b}$. Since for $(\boldsymbol{x},\boldsymbol{z})\in W$ we have

$$(\boldsymbol{x},\boldsymbol{z})=\tilde{\boldsymbol{F}}(\Phi(\boldsymbol{x},\boldsymbol{z}))=\tilde{\boldsymbol{F}}(\boldsymbol{H}(\boldsymbol{x},\boldsymbol{z}),\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z}))=(\boldsymbol{H}(\boldsymbol{x},\boldsymbol{z}),\boldsymbol{F}(\boldsymbol{H}(\boldsymbol{x},\boldsymbol{z}),\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})))$$

we obtain that $\boldsymbol{H}(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}$ and

$$\boldsymbol{F}(\boldsymbol{H}(\boldsymbol{x},\boldsymbol{z}),\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})) = \boldsymbol{F}(\boldsymbol{x},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})) = \boldsymbol{z}$$

for $(\boldsymbol{x},\boldsymbol{z}) \in W$.

Uniqueness of $\boldsymbol{G}$ follows from the fact that $\tilde{\boldsymbol{F}}$ is one-to-one: if $\tilde{\boldsymbol{G}}$ is a function from W to $\mathbb{R}^n$ with $(\boldsymbol{x},\tilde{\boldsymbol{G}}(\boldsymbol{x},\boldsymbol{z})) \in V$ and $\boldsymbol{F}(\boldsymbol{x},\tilde{\boldsymbol{G}}(\boldsymbol{x},\boldsymbol{z})) = \boldsymbol{z}$ for all $(\boldsymbol{x},\boldsymbol{z}) \in W$, then for $(\boldsymbol{x},\boldsymbol{z}) \in W$ we have

$$\tilde{\boldsymbol{F}}(\boldsymbol{x},\tilde{\boldsymbol{G}}(\boldsymbol{x},\boldsymbol{z})) = (\boldsymbol{x},\boldsymbol{z}) = \tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z}))$$

so $\tilde{\boldsymbol{G}}(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})$.

Thus VIII.5.2.1. is proved (see Exercise VIII.5.11.5. for the derivative formula). 👍

Proof From Implicit Function Theorem

VIII.5.2.4. Now assume the Smooth Implicit Function Theorem VIII.5.1.7. holds. We will deduce VIII.5.2.1..

Suppose $U \subseteq \mathbb{R}^{p+n}$ is open, $(\boldsymbol{a},\boldsymbol{b}) \in U$, and $\boldsymbol{F} : U \to \mathbb{R}^n$ is as in the statement of VIII.5.2.1.. Set $\boldsymbol{c} = \boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$. Write $\mathbb{R}^{p+2n} \cong \mathbb{R}^p \times \mathbb{R}^n \times \mathbb{R}^n$, and set

$$U' = \{(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y}) \in \mathbb{R}^{p+2n} : (\boldsymbol{x},\boldsymbol{y}) \in U\} .$$

Then U' is open in $\mathbb{R}^{p+2n}$, and $(\boldsymbol{a},\boldsymbol{c},\boldsymbol{b}) \in U'$. Define $\tilde{\boldsymbol{F}} : U' \to \mathbb{R}^n$ by

$$\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y}) = \boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{z} .$$

Then $\tilde{\boldsymbol{F}}$ is $\mathcal{C}^r$ on U', $\tilde{\boldsymbol{F}}(\boldsymbol{a},\boldsymbol{c},\boldsymbol{b}) = \boldsymbol{0}$, and

$$D\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y}) = \begin{bmatrix} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) & -I_n & D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) \end{bmatrix}$$

for $(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y}) \in U'$, and in particular $D_{\boldsymbol{y}}\tilde{\boldsymbol{F}}(\boldsymbol{a},\boldsymbol{c},\boldsymbol{b}) = D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible.

We can thus solve the equation $\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{y}) = \boldsymbol{0}$ implicitly for $\boldsymbol{y}$ as a function of $(\boldsymbol{x},\boldsymbol{z})$, i.e. there are open neighborhoods W of $(\boldsymbol{a},\boldsymbol{c})$ and V of $(\boldsymbol{a},\boldsymbol{b})$ and a unique function $\boldsymbol{G} : W \to \mathbb{R}^n$ such that $\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z}) \in V$ and $\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})) = \boldsymbol{0}$ for all $(\boldsymbol{x},\boldsymbol{z}) \in W$, and $\boldsymbol{G}$ is $\mathcal{C}^r$. But $\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})) = \boldsymbol{0}$ is equivalent to $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{G}(\boldsymbol{x},\boldsymbol{z})) = \boldsymbol{z}$.

Thus VIII.5.2.1. is proved (cf. Exercise VIII.5.11.5.). 👍

VIII.5.2.5. There are versions of the Continuous and Mixed Implicit Function Theorems similar to VIII.5.2.1. (Exercise VIII.5.11.16.).

VIII.5.3. Proof of the Implicit Function Theorem, Case $n = 1$

It is much easier to prove the Implicit Function Theorem in the case where we are only solving for one variable in terms of the others. We also obtain the desired results under somewhat weaker hypotheses. It does not make any difference in the proof how many other variables there are, i.e. we will prove the case where $n = 1$ and p is arbitrary.

VIII.5.3.1. The general setup for this case is a function $F : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Write $(\boldsymbol{x}, y) = (x_1, \ldots, x_p, y)$ for the coordinates of a general point of U. We will need to assume at a minimum that F is continuous on U and that $\frac{\partial F}{\partial y}(\boldsymbol{a}, b)$ exists and is nonzero. To obtain most of the conclusions we will need a little more. Of course, the conditions need only hold on *some* neighborhood of $(\boldsymbol{a}, b)$, since for the conclusion we will have to restrict to a smaller neighborhood of $(\boldsymbol{a}, b)$ anyway in general, i.e. the conclusion will be only a local one.

The Continuous Version

VIII.5.3.2. Theorem. [One-Dimensional Weak Implicit Function Theorem, Continuous Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Suppose F is continuous on U and $\frac{\partial F}{\partial y}$ exists and is nonzero at $(\boldsymbol{a}, b)$. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) Functions $\overline{f}, \underline{f} : W \to \mathbb{R}$ such that $\overline{f}(\boldsymbol{a}) = \underline{f}(\boldsymbol{a}) = b$, $(\boldsymbol{x}, \overline{f}(\boldsymbol{x}))$ and $(\boldsymbol{x}, \underline{f}(\boldsymbol{x}))$ are in V and

$$F(\boldsymbol{x}, \overline{f}(\boldsymbol{x})) = F(\boldsymbol{x}, \underline{f}(\boldsymbol{x})) = 0$$

for all $\boldsymbol{x} \in W$, and such that if f is any function from W to $\mathbb{R}$ with $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$, then $\underline{f} \leq f \leq \overline{f}$.

Additionally, $\overline{f}$ is upper semicontinuous and $\underline{f}$ is lower semicontinuous on W, and $\overline{f}$ and $\underline{f}$ are continuous at $\boldsymbol{a}$.

Proof: By passing to a smaller neighborhood, we may assume U is convex. Suppose $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) > 0$. Then there are $y_1 > b$ and $y_2 < b$ such that $(\boldsymbol{a}, y_1)$ and $(\boldsymbol{a}, y_2)$ are in U and $\frac{F(\boldsymbol{a},z)-F(\boldsymbol{a},b)}{z-b} > 0$ for $y_2 \leq z \leq y_1$, $z \neq b$ (in fact, this will be true for all such y_1, y_2 sufficiently close to b). It follows that $F(\boldsymbol{a}, y_1) > 0$ and $F(\boldsymbol{a}, y_2) < 0$ (and in fact, $F(\boldsymbol{a}, z) > 0$ for $b < z \leq y_1$ and $F(\boldsymbol{a}, z) < 0$ for $y_2 \leq z < b$). Then there is an open neighborhood W of $\boldsymbol{a}$ in $\mathbb{R}^p$ such that, for all $\boldsymbol{x} \in W$, $(\boldsymbol{x}, y_1)$ and $(\boldsymbol{x}, y_2)$ are in U, $F(\boldsymbol{x}, y_1) > 0$, and $F(\boldsymbol{x}, y_2) < 0$. So for each fixed $\boldsymbol{x} \in W$, there is at least one y between y_2 and y_1 for which $F(\boldsymbol{x}, y) = 0$ by the Intermediate Value Theorem, since F is continuous on U. Among these, there is a largest $\overline{y}$ and a smallest $\underline{y}$. Set $\overline{f}(\boldsymbol{x}) = \overline{y}$ and $\underline{f}(\boldsymbol{x}) = \underline{y}$. This defines $\overline{f}$ and $\underline{f}$ on W. We obviously have $\overline{f}(\boldsymbol{a}) = \underline{f}(\boldsymbol{a}) = b$.

Set $V = W \times (y_2, y_1) \subseteq U$. If f is any real-valued function on W such that $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$, we obviously have $\underline{f}(\boldsymbol{x}) \leq f(\boldsymbol{x}) \leq \overline{f}(\boldsymbol{x})$ for all $\boldsymbol{x} \in W$.

It remains to show that $\overline{f}$ and $\underline{f}$ are semicontinuous on W. We give the argument for $\overline{f}$; the one for $\underline{f}$ is essentially identical. Let $\boldsymbol{x}_n \in W$, $\boldsymbol{x}_n \to \boldsymbol{x}_0$. Passing to a subsequence we may assume $\overline{f}(\boldsymbol{x}_n) \to z$ for some z between y_2 and y_1. But then $0 = F(\boldsymbol{x}_n, \overline{f}(\boldsymbol{x}_n)) \to F(\boldsymbol{x}_0, z)$ by continuity of F, so $F(\boldsymbol{x}_0, z) = 0$ and

$z \leq \overline{f}(\boldsymbol{x}_0)$. Thus $\limsup_{\boldsymbol{x}\to\boldsymbol{x}_0} \overline{f}(\boldsymbol{x}) \leq \overline{f}(\boldsymbol{x}_0)$ for any $\boldsymbol{x}_0 \in W$. Similarly $\liminf_{\boldsymbol{x}\to\boldsymbol{x}_0} \underline{f}(\boldsymbol{x}) \geq \underline{f}(\boldsymbol{x}_0)$ for any $\boldsymbol{x}_0 \in W$. Since

$$b = \underline{f}(\boldsymbol{a}) \leq \liminf_{\boldsymbol{x}\to\boldsymbol{a}} \underline{f}(\boldsymbol{x}) \leq \limsup_{\boldsymbol{x}\to\boldsymbol{a}} \overline{f}(\boldsymbol{x}) \leq \overline{f}(\boldsymbol{a}) = b$$

we have that $\overline{f}$ and $\underline{f}$ are continuous at $\boldsymbol{a}$.

The argument in the case $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) < 0$ is essentially identical. 👍

VIII.5.3.3. Under the hypotheses of VIII.5.3.2., even if $p = 1$, there is not necessarily a continuous function f on a neighborhood W of $\boldsymbol{a}$ with $f(\boldsymbol{a}) = b$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$(Exercise VIII.5.11.4.).

VIII.5.3.4. Theorem. [One-Dimensional Implicit Function Theorem, Continuous Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Suppose F is continuous on U and $\frac{\partial F}{\partial y}$ exists and is nonzero at all points of U. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(\boldsymbol{a}) = b$, and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$.

Additionally, f is continuous on W.

Proof: Everything follows immediately from VIII.5.3.2. except the uniqueness of f, i.e. that $\overline{f} = \underline{f}$. Define V and W as in the proof of VIII.5.3.2.. Fix $\boldsymbol{x} \in W$. If there were y and z, $y \neq z$, with $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}, z)$ in V and $F(\boldsymbol{x}, y) = F(\boldsymbol{x}, z) = 0$, then by Rolle's Theorem there would be a c between y and z with $\frac{\partial F}{\partial y}(\boldsymbol{x}, c) = 0$, contradicting the assumption that $\frac{\partial F}{\partial y}$ is never zero on U. Thus there is a unique y such that $(\boldsymbol{x}, y) \in V$ and $F(\boldsymbol{x}, y) = 0$, and we may set $f(\boldsymbol{x}) = y$. This f is the unique real-valued function defined on W with $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$. 👍

VIII.5.3.5. Corollary. Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Suppose F is continuous on U and $\frac{\partial F}{\partial y}$ exists at all points of U, and is continuous at $(\boldsymbol{a}, b)$, with $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) \neq 0$. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(\boldsymbol{a}) = b$, and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$.

Additionally, f is continuous on W.

Proof: By continuity of $\frac{\partial F}{\partial y}$ at $(\boldsymbol{a}, b)$, there is a neighborhood U' of $(\boldsymbol{a}, b)$ contained in U such that $\frac{\partial F}{\partial y}(\boldsymbol{x}, y) \neq 0$ for all $(\boldsymbol{x}, y) \in U'$. Apply VIII.5.3.4. with U' in place of U. 👍

A Pictorial Version

VIII.5.3.6. We illustrate the argument in the case $p = n = 1$ (the picture represents the case of general p if the horizontal axis is regarded as multidimensional). Suppose U is an open set in $\mathbb{R}^2$, and F is a continuous function from U to $\mathbb{R}$ such that $\frac{\partial F}{\partial y}$ exists and is nonzero on U, and let $(a, b) \in U$ with $F(a, b) = 0$. Fix an open rectangle R centered at (a, b) and contained in U, and do the following construction within R. We assume that $\frac{\partial F}{\partial y}(a, b) > 0$; the proof in the case $\frac{\partial F}{\partial y}(a, b) < 0$ is identical with red and blue reversed. Because of the shapes in the diagrams (the proportions may vary), this proof can be called the "T proof."

Since $F(a, b) = 0$ and $\frac{\partial F}{\partial y}(a, b) > 0$, $g(y) = F(a, y)$ is increasing at b (V.3.6.8.), i.e. there are $y_1 < b$ and $y_2 > b$ such that (a, y_1) and $(a, y_2) \in R$ and $F(a, y) > F(a, b) = 0$ for $b < y \leq y_2$, and $F(a, y) < 0$ for $y_1 \leq y < b$, i.e. $F(x, y) > 0$ for all points (x, y) in red in Figure VIII.16 and $F(x, y) < 0$ for all (x, y) in blue:

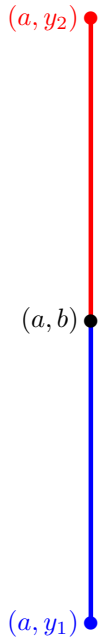


Figure VIII.16: Implicit Function Theorem I

We may assume $y_2 - b = b - y_1$ if desired by taking a minimum; although this is not necessary for the proof, it will be convenient later since this can be called ϵ.

Since $F(a, y_2) > 0$ and F is continuous, there are $x_1 < a$ and $x_2 > a$ such that (x_1, y_2) and (x_2, y_2) are in R and $F(x, y_2) > 0$ for $x_1 \leq x \leq x_2$ (). Similarly, since $F(a, y_1) < 0$, there are $x_3 < a$ and $x_4 > a$ with (x_3, y_1) and (x_4, y_1) in R and $F(x, y_1) < 0$ for $x_3 \leq x \leq x_4$. As before, we may assume $x_3 = x_1$ and $x_4 = x_2$ and that $x_2 - a = a - x_1$ by taking a minimum; this can be called δ. Thus all red points (x, y) in Figure VIII.17 have $F(x, y) > 0$, and all blue points have $F(x, y) < 0$:

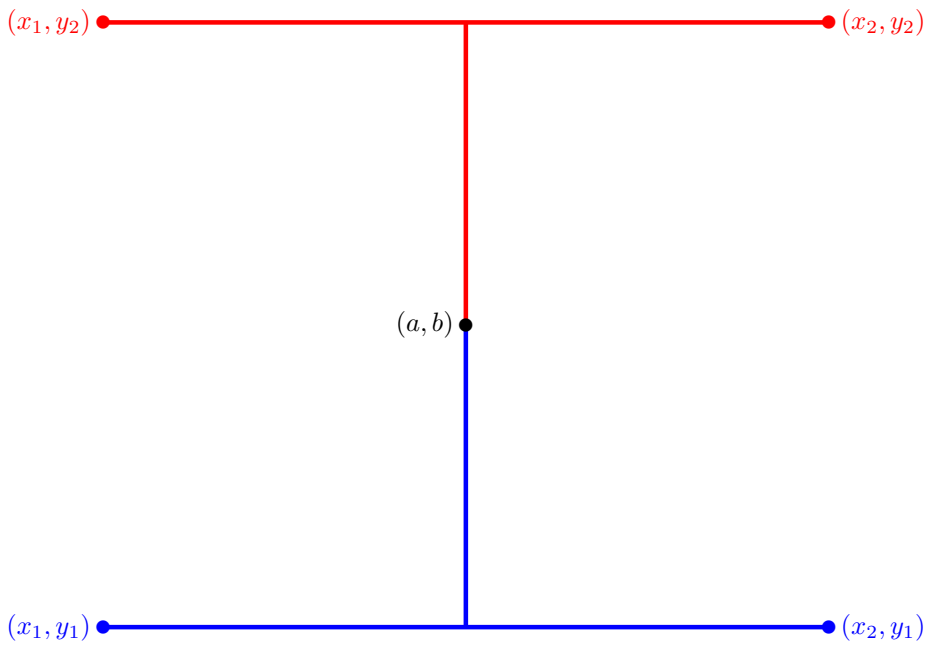


Figure VIII.17: Implicit Function Theorem II

In the case of general p, the top and bottom horizontal segments become balls in $\mathbb{R}^p$ centered at $\boldsymbol{a}$, which can be taken to have the same radius which we can call δ.

Now fix x_0, $x_1 \leq x_0 \leq x_2$. We have that $F(x_0, y_1) < 0$ and $F(x_0, y_2) > 0$. Since $F(x_0, y)$ is a continuous function of y, there must be a y, $y_1 < y < y_2$, with $F(x_0, y) = 0$. If there were more than one such y, by Rolle's Theorem there would be a y in between with $\frac{\partial F}{\partial y}(x_0, y) = 0$, contradicting that $\frac{\partial F}{\partial y}$ is nonzero on U. Thus there is a *unique* y between y_1 and y_2 with $F(x_0, y) = 0$, and we call this y $f(x_0)$. This procedure uniquely defines a function f on the interval $[x_1, x_2]$ with values in (y_1, y_2), which satisfies $F(x, f(x)) = 0$ for $x_1 \leq x \leq x_2$. We also have $F(x, y) > 0$ for $f(x) < y \leq y_2$ and $F(x, y) < 0$ if $y_1 \leq y < f(x)$, for any $x \in [x_1, x_2]$. See Figure VIII.18.

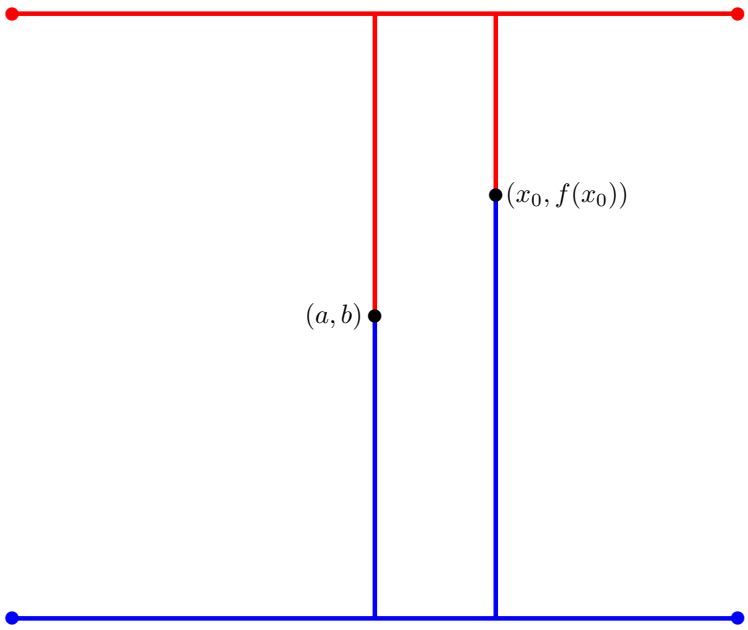


Figure VIII.18: Implicit Function Theorem III

To show f is continuous, fix x_0, $x_1 < x_0 < x_2$. Let $\epsilon > 0$. We may assume $\epsilon \leq \min(y_2 - f(x_0), f(x_0) - y_1)$ without loss of generality. We repeat the first part of the proof replacing a and b by x_0 and $f(x_0)$, and y_1 and y_2 by $f(x_0) \pm \epsilon$: $F(x_0, y) > 0$ for $f(x_0) < y \leq f(x_0) + \epsilon$ and $F(x_0, y) < 0$ for $f(x_0) - \epsilon \leq y < f(x_0)$, so there is a δ, $0 < \delta \leq \min(x_2 - x_0, x_0 - x_1)$, such that $F(x, y) > 0$ for all red points in Figure VIII.19 and $F(x, y) < 0$ for all blue points.

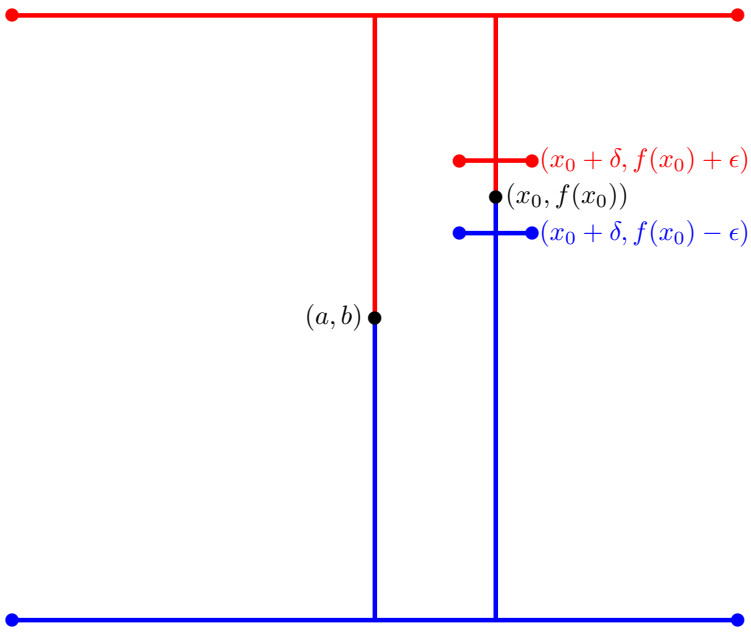


Figure VIII.19: Implicit Function Theorem IV

If $|x - x_0| < \delta$, there is thus a unique y, $f(x_0) - \epsilon < y < f(x_0) + \epsilon$, with $F(x, y) = 0$; this y must be $f(x)$, i.e. $|f(x) - f(x_0)| < \epsilon$. Thus f is continuous at x_0.

The Differentiable Version

The next lemma contains the most delicate part of the entire proof of the Implicit Function Theorem. It is needed for the differentiable and smooth versions of the theorem if $n = 1$, and even for the continuous version for general n. If the first-order partials of F are continuous at $(\boldsymbol{a}, b)$, the proof can be simplified (Exercise VIII.5.11.9.).

VIII.5.3.7. LEMMA. Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Suppose F is continuous on U and differentiable at $(\boldsymbol{a}, b)$, and that $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) \neq 0$. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

such that whenever f is a real-valued function on W with $f(\boldsymbol{a}) = b$ and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$ (there exist such functions by VIII.5.3.2.), then f is differentiable at $\boldsymbol{a}$ and, for $1 \leq j \leq p$,

$$\frac{\partial f}{\partial x_j}(\boldsymbol{a}) = -\frac{\frac{\partial F}{\partial x_j}(\boldsymbol{a}, b)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b)} \ .$$

PROOF: The little-o notation () is very helpful in this proof. As $\boldsymbol{h} = (\Delta\boldsymbol{x}, \Delta y) = (\Delta x_1, \ldots, \Delta x_p, \Delta y) \to 0$, we have

$$\Delta F = F(\boldsymbol{a}+\Delta\boldsymbol{x}, b+\Delta y) - F(\boldsymbol{a}, b) = \frac{\partial F}{\partial x_1}(\boldsymbol{a}, b)\Delta x_1 + \cdots + \frac{\partial F}{\partial x_p}(\boldsymbol{a}, b)\Delta x_p + o(1)\|\Delta x\| + \frac{\partial F}{\partial y}(\boldsymbol{a}, b)\Delta y + o(1)\Delta y$$

and if we set $y = f(\boldsymbol{x})$, so $\Delta y = f(\boldsymbol{a}+\Delta\boldsymbol{x}) - f(\boldsymbol{a})$, we have $\Delta F = 0$ since $F(\boldsymbol{a}+\Delta\boldsymbol{x}, b+\Delta y) = F(\boldsymbol{a}, b) = 0$, i.e.

$$\Delta y \left[\frac{\partial F}{\partial y}(\boldsymbol{a}, b) + o(1)\right] = -\frac{\partial F}{\partial x_1}(\boldsymbol{a}, b)\Delta x_1 - \cdots - \frac{\partial F}{\partial x_p}(\boldsymbol{a}, b)\Delta x_p + o(1)\|\Delta x\| \ .$$

By VIII.5.3.2., we have $\underline{f} \leq f \leq \overline{f}$, and hence by the Squeeze Theorem f is continuous at $\boldsymbol{a}$; thus $\Delta y = o(1)$ as $\Delta\boldsymbol{x} \to \boldsymbol{0}$, and hence $\overline{\boldsymbol{h}} = \boldsymbol{o}(1)$ as $\Delta\boldsymbol{x} \to \boldsymbol{0}$. Thus we have that the $o(1)$'s in the last expression are still $o(1)$ as $\Delta\boldsymbol{x} \to \boldsymbol{0}$. Also, if $\Delta\boldsymbol{x}$ is sufficiently small, we can divide by the expression in brackets on the left (using that $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) \neq 0$ by assumption) to obtain

$$\Delta y = \frac{-\frac{\partial F}{\partial x_1}(\boldsymbol{a}, b)\Delta x_1 - \cdots - \frac{\partial F}{\partial x_p}(\boldsymbol{a}, b)\Delta x_p + o(1)\|\Delta x\|}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b) + o(1)}$$

as $\Delta\boldsymbol{x} \to \boldsymbol{0}$. By (), the last expression is

$$\Delta y = \|\Delta\boldsymbol{x}\| \frac{-\frac{\partial F}{\partial x_1}(\boldsymbol{a}, b)\frac{\Delta x_1}{\|\Delta\boldsymbol{x}\|} - \cdots - \frac{\partial F}{\partial x_p}(\boldsymbol{a}, b)\frac{\Delta x_p}{\|\Delta\boldsymbol{x}\|} + o(1)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b) + o(1)}$$

$$= \|\Delta\boldsymbol{x}\| \left[-\frac{\frac{\partial F}{\partial x_1}(\boldsymbol{a}, b)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b)}\frac{\Delta x_1}{\|\Delta\boldsymbol{x}\|} - \cdots - \frac{\frac{\partial F}{\partial x_p}(\boldsymbol{a}, b)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b)}\frac{\Delta x_p}{\|\Delta\boldsymbol{x}\|} + o(1)\right]$$

$$= -\frac{\frac{\partial F}{\partial x_1}(\boldsymbol{a}, b)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b)}\Delta x_1 - \cdots - \frac{\frac{\partial F}{\partial x_p}(\boldsymbol{a}, b)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b)}\Delta x_p + o(1)\|\Delta \boldsymbol{x}\|$$

as $\Delta \boldsymbol{x} \to \boldsymbol{0}$; thus f is differentiable at $\boldsymbol{a}$ and

$$\frac{\partial f}{\partial x_j}(\boldsymbol{a}) = -\frac{\frac{\partial F}{\partial x_j}(\boldsymbol{a}, b)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b)} \ .$$



The following technical strengthening will be used in the inductive proof of the general Implicit Function Theorem. It can be proved in the same way, or it follows immediately from VIII.5.3.7. by simply holding the other coordinates constant:

VIII.5.3.8. Lemma. Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Let $S = \{x_{k_1}, \ldots, x_{k_m}\}$ be a subset of $\{x_1, \ldots, x_p\}$. Suppose F is continuous on U and differentiable with respect to $S \cup \{y\}$ at $(\boldsymbol{a}, b)$, and that $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) \neq 0$. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

such that whenever f is a real-valued function on W with $f(\boldsymbol{a}) = b$ and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$ (there exist such functions by VIII.5.3.2.), then f is differentiable with respect to S at $\boldsymbol{a}$ and, for $x_j \in S$,

$$\frac{\partial f}{\partial x_j}(\boldsymbol{a}) = -\frac{\frac{\partial F}{\partial x_j}(\boldsymbol{a}, b)}{\frac{\partial F}{\partial y}(\boldsymbol{a}, b)} \ .$$

VIII.5.3.9. Theorem. [One-Dimensional Implicit Function Theorem, Differentiable Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Suppose F is differentiable at all points of U and $\frac{\partial F}{\partial y}$ is nonzero at all points of U. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(\boldsymbol{a}) = b$, and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$.

Additionally, f is differentiable at every point of W and, for all $\boldsymbol{x} \in W$ and $1 \leq j \leq p$,

$$\frac{\partial f}{\partial x_j}(\boldsymbol{x}) = -\frac{\frac{\partial F}{\partial x_j}(\boldsymbol{x}, f(\boldsymbol{x}))}{\frac{\partial F}{\partial y}(\boldsymbol{x}, f(\boldsymbol{x}))} \ .$$

PROOF: The existence of V, W, and f, and the uniqueness (and continuity) of f, follow immediately from VIII.5.3.4., and if $\boldsymbol{x}_0 \in V$, differentiability of f at $\boldsymbol{x}_0$ follows from VIII.5.3.7. applied at the point $(\boldsymbol{x}_0, f(\boldsymbol{x}_0))$. 👍

In the case $p = 1$, we get the version of the Implicit Function Theorem which justifies implicit differentiation in first-semester calculus (note that there is no truly "one-dimensional" version of the Implicit Function Theorem – one-variable implicit differentiation must be justified by the "two-dimensional" version):

VIII.5.3.10. COROLLARY. [ONE-VARIABLE IMPLICIT FUNCTION THEOREM, DIFFERENTIABLE VERSION] Let $F(x, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^2$ is open, and $(a, b) \in U$ with $F(a, b) = 0$. Suppose F is differentiable at all points of U and $\frac{\partial F}{\partial y}$ is nonzero at all points of U. Then there are:

(i) An open set V in $\mathbb{R}^2$ containing (a, b), with $V \subseteq U$;

(ii) An open interval W in $\mathbb{R}$ containing a;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(a) = b$, and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$.

Additionally, f is differentiable at every point of W and, for all $x \in W$,

$$f'(x) = -\frac{\frac{\partial F}{\partial x}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))} .$$

VIII.5.3.11. The fact that the slope of the tangent line to the level curve at (x, y) is $-\frac{\frac{\partial F}{\partial x}(x,y)}{\frac{\partial F}{\partial y}(x,y)}$ should come as no surprise, since this level curve should be perpendicular to $\nabla F(x, y) = \left(\frac{\partial F}{\partial x}(x, y), \frac{\partial F}{\partial y}(x, y)\right)$ (). But what is not obvious is that the level curve through (x, y) is a smooth curve with a tangent line, or indeed that the level set is a "curve" at all; this is the theoretical conclusion of the theorem (cf. VIII.5.1.9.).

The formula for $f'(x)$ given in the theorem is equivalent to but more efficient than the one obtained by first-semester calculus methods before partial derivatives are introduced (VIII.5.11.8.); indeed, the partial derivative version is the "right" way to do one-variable implicit differentiation.

We also get a mixed version in which differentiability with respect to only some of the variables is assumed and concluded:

VIII.5.3.12. THEOREM. [ONE-DIMENSIONAL IMPLICIT FUNCTION THEOREM, MIXED VERSION] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Let $S = \{x_{k_1}, \ldots, x_{k_m}\}$ be a subset of $\{x_1, \ldots, x_p\}$. Suppose F is differentiable with respect to $S \cup \{y\}$ at all points of U and $\frac{\partial F}{\partial y}$ is nonzero at all points of U. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(\boldsymbol{a}) = b$, and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$.

Additionally, f is continuous on W and differentiable with respect to S at every point of W and, for all $\boldsymbol{x} \in W$ and $x_j \in S$,

$$\frac{\partial f}{\partial x_j}(\boldsymbol{x}) = -\frac{\frac{\partial F}{\partial x_j}(\boldsymbol{x}, f(\boldsymbol{x}))}{\frac{\partial F}{\partial y}(\boldsymbol{x}, f(\boldsymbol{x}))} \ .$$

The continuous version (VIII.5.3.5.) and the differentiable version (VIII.5.3.9.) are the special cases $S = \emptyset$ and $S = \{x_1, \ldots, x_p\}$ respectively.

The Smooth Version

VIII.5.3.13. Theorem. [One-Dimensional Implicit Function Theorem, Smooth Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Suppose F is $\mathcal{C}^r$ on U $(1 \leq r \leq \infty)$ and $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) \neq 0$. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$, such that $\frac{\partial F}{\partial y}(\boldsymbol{x}, y) \neq 0$ for all $(\boldsymbol{x}, y) \in V$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(\boldsymbol{a}) = b$, and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$.

Additionally, f is $\mathcal{C}^r$ on W and, for all $\boldsymbol{x} \in W$ and $1 \leq j \leq p$,

$$\frac{\partial f}{\partial x_j}(\boldsymbol{x}) = -\frac{\frac{\partial F}{\partial x_j}(\boldsymbol{x}, f(\boldsymbol{x}))}{\frac{\partial F}{\partial y}(\boldsymbol{x}, f(\boldsymbol{x}))} \ .$$

Proof: By continuity, there is an open neighborhood U' of $(\boldsymbol{a}, b)$ contained in U such that $\frac{\partial F}{\partial y}$ is nonzero everywhere on U'. Everything then follows immediately from VIII.5.3.9. (applied with U' in place of U) except the fact that f is $\mathcal{C}^r$. We prove this by induction on r.

First suppose F is $\mathcal{C}^1$. Then, for $1 \leq j \leq p$, $\frac{\partial F}{\partial x_j}$ is continuous on V, and f is continuous on W, so $g_j(\boldsymbol{x}) = \frac{\partial F}{\partial x_j}(\boldsymbol{x}, f(\boldsymbol{x}))$ is continuous on W. Similarly, $g(\boldsymbol{x}) = \frac{\partial F}{\partial y}(\boldsymbol{x}, f(\boldsymbol{x}))$ is continuous and nonzero on W, so

$$\frac{\partial f}{\partial x_j}(\boldsymbol{x}) = -\frac{g_j(\boldsymbol{x})}{g(\boldsymbol{x})}$$

is also continuous on W for each j. Thus f is $\mathcal{C}^1$ on W.

Now assume the result for r, and suppose F is $\mathcal{C}^{r+1}$. Then by the inductive hypothesis f is $\mathcal{C}^r$, so g and each g_j are $\mathcal{C}^r$ by the higher-order Chain Rule (). We then obtain that $\frac{\partial f}{\partial x_j}$ is $\mathcal{C}^r$ for each j since it is a quotient of $\mathcal{C}^r$ functions with nonvanishing denominator. Thus f is $\mathcal{C}^{r+1}$. 👍

VIII.5.3.14. Corollary. [One-Variable Implicit Function Theorem, Smooth Version] Let $F(x,y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^2$ is open, and $(a,b) \in U$ with $F(a,b) = 0$. Suppose F is $\mathcal{C}^r$ on U $(1 \leq r \leq \infty)$ and $\frac{\partial F}{\partial y}(a,b) \neq 0$. Then there are:

(i) An open set V in $\mathbb{R}^2$ containing (a,b), with $V \subseteq U$, such that $\frac{\partial F}{\partial y}(x,y) \neq 0$ for all $(x,y) \in V$;

(ii) An open set W in $\mathbb{R}$ containing a;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(a) = b$, and $(x, f(x)) \in V$ and $F(x, f(x)) = 0$ for all $x \in W$.

Additionally, f is $\mathcal{C}^r$ on W and, for all $x \in W$,

$$f'(x) = -\frac{\frac{\partial F}{\partial x}(x, f(x))}{\frac{\partial F}{\partial y}(x, f(x))} .$$

We also get a mixed version which will be used in the inductive proof of the general theorem:

VIII.5.3.15. Theorem. [One-Dimensional Implicit Function Theorem, Mixed Smooth Version] Let $F(x_1, \ldots, x_p, y) : U \to \mathbb{R}$, where $U \subseteq \mathbb{R}^{p+1}$ is open, and $(\boldsymbol{a}, b) = (a_1, \ldots, a_p, b) \in U$ with $F(\boldsymbol{a}, b) = 0$. Let $S = \{x_{k_1}, \ldots, x_{k_m}\}$ be a subset of $\{x_1, \ldots, x_p\}$. Suppose F, $\frac{\partial F}{\partial y}$, and $\frac{\partial F}{\partial x_j}$ are continuous on U $(1 \leq r \leq \infty)$ for all $x_j \in S$, and $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) \neq 0$. Then there are:

(i) An open set V in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b)$, with $V \subseteq U$, such that $\frac{\partial F}{\partial y}(\boldsymbol{x}, y) \neq 0$ for all $(\boldsymbol{x}, y) \in V$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $f : W \to \mathbb{R}$ such that $f(\boldsymbol{a}) = b$, and $(\boldsymbol{x}, f(\boldsymbol{x})) \in V$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$.

Additionally, f and $\frac{\partial f}{\partial x_j}$ are continuous on W for all $x_j \in S$ and, for all $\boldsymbol{x} \in W$, $x_j \in S$, and $1 \leq j \leq p$,

$$\frac{\partial f}{\partial x_j}(\boldsymbol{x}) = -\frac{\frac{\partial F}{\partial x_j}(\boldsymbol{x}, f(\boldsymbol{x}))}{\frac{\partial F}{\partial y}(\boldsymbol{x}, f(\boldsymbol{x}))} .$$

Proof: The conclusion follows from VIII.5.3.12. just as VIII.5.3.13. follows from VIII.5.3.9.. 👍

VIII.5.4. Proof of the General Implicit Function Theorem

VIII.5.4.1. The most straightforward proof of the general Implicit Function Theorem is by induction on the number n of variables being solved for (this was also the first proof). The inductive proof requires a stronger continuity hypothesis on the partial derivatives of $\boldsymbol{F}$ than the one-dimensional version, as explained in VIII.5.4.3..

We will prove the following mixed version of the Implicit Function Theorem, which includes both the continuous version (VIII.5.1.6.) and the smooth ($\mathcal{C}^1$) version (VIII.5.1.7.):

VIII.5.4.2. THEOREM. [IMPLICIT FUNCTION THEOREM, MIXED VERSION] Let U be an open set in $\mathbb{R}^{p+n}$ and $(\boldsymbol{a},\boldsymbol{b}) \in U$. Let $\boldsymbol{F} : U \to \mathbb{R}^n$ with $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is continuous on U and that $D_{\boldsymbol{y}}\boldsymbol{F}$ exists and is continuous on U, i.e. that $\frac{\partial \boldsymbol{F}}{\partial y_j}$ exists and is continuous on U for $1 \le j \le n$. If $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible, i.e. $det(D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})) \neq 0$, then there are:

(i) An open set V in $\mathbb{R}^{p+n}$ containing $(\boldsymbol{a},\boldsymbol{b})$, with $V \subseteq U$, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$ is invertible for all $(\boldsymbol{x},\boldsymbol{y}) \in V$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $\boldsymbol{f} : W \to \mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{b}$ and, for all $\boldsymbol{x} \in W$, $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{0} \ .$$

Additionally, $\boldsymbol{f}$ is continuous on W. If additionally $S = \{x_{k_1}, \ldots, x_{k_m}\}$ is a subset of $\{x_1, \ldots, x_p\}$ and $\frac{\partial \boldsymbol{F}}{\partial x_{k_j}}$ exists and is continuous on U for each $x_{k_j} \in S$, then $\frac{\partial \boldsymbol{f}}{\partial x_{k_j}}$ exists and is continuous on W for each $x_{k_j} \in S$ and these partials satisfy

$$D_S\boldsymbol{f}(\boldsymbol{x}) = -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1} D_S\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$$

for all $\boldsymbol{x} \in W$, where $D_S\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$ is the $n \times m$ matrix

$$D_S\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \begin{bmatrix} \frac{\partial F_1}{\partial x_{k_1}}(\boldsymbol{x},\boldsymbol{y}) & \cdots & \frac{\partial F_1}{\partial x_{k_m}}(\boldsymbol{x},\boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial x_{k_1}}(\boldsymbol{x},\boldsymbol{y}) & \cdots & \frac{\partial F_n}{\partial x_{k_m}}(\boldsymbol{x},\boldsymbol{y}) \end{bmatrix} \ .$$

PROOF: The case $n = 1$ is VIII.5.3.15.. We now assume the theorem is true for n, and show it is true for $n+1$. So suppose $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{F}(x_1, \ldots, x_p, y_1, \ldots, y_{n+1})$ is a continuous function from an open set $U \subseteq \mathbb{R}^{p+n+1}$ to $\mathbb{R}^{n+1}$, $(\boldsymbol{a},\boldsymbol{b}) = (a_1, \ldots, a_p, b_1, \ldots, b_{n+1}) \in U$, that $D_{\boldsymbol{y}}\boldsymbol{F}$ and $D_S\boldsymbol{F}$ are continuous on U for some subset S of $\{x_1, \ldots, x_p\}$, and $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible. The strategy for finding the implicit function $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^{n+1}$ will be to show that for some i, $1 \le i \le n+1$, the y_k for $2 \le k \le n+1$ can be written as implicit functions g_k of $(x_1, \ldots, x_p, y_1)$ satisfying

$$F_j(x_1, \ldots, x_p, y_1, g_2(x_1, \ldots, x_p, y_1), \ldots, g_{n+1}(x_1, \ldots, x_p, y_1)) = 0$$

for $j \neq i$ and $(x_1, \ldots, x_p, y_1)$ close to $(a_1, \ldots, a_p, b_1)$ by the inductive hypothesis, and that y_1 can then be written as an implicit function h of $(x_1, \ldots, x_p)$ satisfying

$$F_i(x_1, \ldots, x_p, h(\boldsymbol{x}), g_2(\boldsymbol{x}, h(\boldsymbol{x})), \ldots, g_{n+1}(\boldsymbol{x}, h(\boldsymbol{x}))) = 0$$

for $\boldsymbol{x}$ near $\boldsymbol{a}$ by an application of VIII.5.3.15.. Then $\boldsymbol{f}$ can be defined by

$$\boldsymbol{f}(\boldsymbol{x}) = (h(\boldsymbol{x}), g_2(\boldsymbol{x}, h(\boldsymbol{x})), \ldots, g_{n+1}(\boldsymbol{x}, h(\boldsymbol{x}))) \ .$$

For the details, we will use the following notation. Write $J(\boldsymbol{x},\boldsymbol{y})$ for the determinant of the $(n+1)\times(n+1)$ matrix $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$. For each i, let $J_i(\boldsymbol{x},\boldsymbol{y})$ be the determinant of the i'th minor in the first column, i.e.

$$J_i(\boldsymbol{x},\boldsymbol{y}) = (-1)^{i+1} det \begin{bmatrix} \frac{\partial F_1}{\partial y_2}(\boldsymbol{x},\boldsymbol{y}) & \cdots & \frac{\partial F_1}{\partial y_{n+1}}(\boldsymbol{x},\boldsymbol{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_{n+1}}{\partial y_2}(\boldsymbol{x},\boldsymbol{y}) & \cdots & \frac{\partial F_{n+1}}{\partial y_{n+1}}(\boldsymbol{x},\boldsymbol{y}) \end{bmatrix}$$

with the i'th row (partials of F_i) deleted. Then

$$J(\boldsymbol{x},\boldsymbol{y}) = J_1(\boldsymbol{x},\boldsymbol{y})\frac{\partial F_1}{\partial y_1}(\boldsymbol{x},\boldsymbol{y}) + J_2(\boldsymbol{x},\boldsymbol{y})\frac{\partial F_2}{\partial y_1}(\boldsymbol{x},\boldsymbol{y}) + \cdots + J_{n+1}(\boldsymbol{x},\boldsymbol{y})\frac{\partial F_{n+1}}{\partial y_1}(\boldsymbol{x},\boldsymbol{y})$$

and so, since $J(\boldsymbol{a},\boldsymbol{b}) \neq 0$, there is an i, $1 \leq i \leq n+1$, such that $J_i(\boldsymbol{a},\boldsymbol{b}) \neq 0$ and $\frac{\partial F_i}{\partial y_1}(\boldsymbol{a},\boldsymbol{b}) \neq 0$. Then there is an open neighborhood U' of $(\boldsymbol{a},\boldsymbol{b})$ such that $J_i(\boldsymbol{x},\boldsymbol{y}) \neq 0$ and $\frac{\partial F_i}{\partial y_1}(\boldsymbol{x},\boldsymbol{y}) \neq 0$ for all $(\boldsymbol{x},\boldsymbol{y}) \in U'$. Apply the inductive hypothesis with $S' = S \cup \{y_1\}$ to obtain an open neighborhood V_1 of $(\boldsymbol{a},\boldsymbol{b})$ in $\mathbb{R}^{p+n+1}$, an open neighborhood W_1 of $(a_1,\ldots,a_p,b_1)$ in $\mathbb{R}^{p+1}$, and unique functions $g_k : W_1 \to \mathbb{R}$ $(2 \leq k \leq n+1)$ which satisfy, for $(x_1,\ldots,x_p,y_1) \in W_1$,

$$(x_1,\ldots,x_p,y_1,g_2(x_1,\ldots,x_p,y_1),\ldots,g_{n+1}(x_1,\ldots,x_p,y_1)) \in V_1$$

$$F_j(x_1,\ldots,x_p,y_1,g_2(x_1,\ldots,x_p,y_1),\ldots,g_{n+1}(x_1,\ldots,x_p,y_1)) = 0$$

for $j \neq i$, and g_k, $\frac{\partial g_k}{\partial y_1}$, and $\frac{\partial g_k}{\partial x_j}$ for $x_j \in S$ continuous on W_1 for each k.

Now, for $(x_1,\ldots,x_p,y_1) \in W_1$, set

$$G(x_1,\ldots,x_p,y_1) = F_i(x_1,\ldots,x_p,y_1,g_2(x_1,\ldots,x_p,y_1),\ldots,g_{n+1}(x_1,\ldots,x_p,y_1)) \ .$$

Then G is continuous on W_1, and $G(a_1,\ldots,a_p,b_1) = 0$. We show that G satisfies the hypotheses of VIII.5.3.15. on W_1. By the Chain Rule, $\frac{\partial G}{\partial y_1}$ exists on W_1 and

$$\frac{\partial G}{\partial y_1}(\boldsymbol{x},y_1) = \frac{\partial F_i}{\partial y_1}(\boldsymbol{x},y_1,g_2(\boldsymbol{x},y_1),\ldots,g_{n+1}(\boldsymbol{x},y_1)) + \frac{\partial F_i}{\partial y_2}(\boldsymbol{x},y_1,g_2(\boldsymbol{x},y_1),\ldots,g_{n+1}(\boldsymbol{x},y_1))\frac{\partial g_2}{\partial y_1}(\boldsymbol{x},y_1) +$$

$$\cdots + \frac{\partial F_i}{\partial y_{n+1}}(\boldsymbol{x},y_1,g_2(\boldsymbol{x},y_1),\ldots,g_{n+1}(\boldsymbol{x},y_1))\frac{\partial g_{n+1}}{\partial y_1}(\boldsymbol{x},y_1) \ .$$

Thus $\frac{\partial G}{\partial y_1}$ is continuous on W_1. It remains to show that $\frac{\partial G}{\partial y_1}(\boldsymbol{a},b_1) \neq 0$. If we set $G_i = G$ and define, for $j \neq i$,

$$G_j(x_1,\ldots,x_p,y_1) = F_j(x_1,\ldots,x_p,y_1,g_2(x_1,\ldots,x_p,y_1),\ldots,g_{n+1}(x_1,\ldots,x_p,y_1))$$

we similarly have

$$\frac{\partial G_j}{\partial y_1}(\boldsymbol{x},y_1) = \frac{\partial F_j}{\partial y_1}(\boldsymbol{x},y_1,g_2(\boldsymbol{x},y_1),\ldots,g_{n+1}(\boldsymbol{x},y_1)) + \frac{\partial F_j}{\partial y_2}(\boldsymbol{x},y_1,g_2(\boldsymbol{x},y_1),\ldots,g_{n+1}(\boldsymbol{x},y_1))\frac{\partial g_2}{\partial y_1}(\boldsymbol{x},y_1) +$$

$$\cdots + \frac{\partial F_j}{\partial y_{n+1}}(\boldsymbol{x},y_1,g_2(\boldsymbol{x},y_1),\ldots,g_{n+1}(\boldsymbol{x},y_1))\frac{\partial g_{n+1}}{\partial y_1}(\boldsymbol{x},y_1) \ .$$

But G_j is identically 0, so abbreviating notation we have

$$\frac{\partial G_i}{\partial y_1} = \frac{\partial F_i}{\partial y_1} + \frac{\partial F_i}{\partial y_2}\frac{\partial g_2}{\partial y_1} + \cdots + \frac{\partial F_i}{\partial y_{n+1}}\frac{\partial g_{n+1}}{\partial y_1}$$

$$0 = \frac{\partial F_j}{\partial y_1} + \frac{\partial F_j}{\partial y_2}\frac{\partial g_2}{\partial y_1} + \cdots + \frac{\partial F_j}{\partial y_{n+1}}\frac{\partial g_{n+1}}{\partial y_1}$$

for $j \neq i$. Multiplying the j'th equation by J_j and adding the equations together, we get

$$J_i \frac{\partial G_i}{\partial y_1} = \sum_{j=1}^{n+1} J_j \frac{\partial F_j}{\partial y_1} + \frac{\partial g_2}{\partial y_1} \sum_{j=1}^{n+1} J_j \frac{\partial F_j}{\partial y_2} + \cdots + \frac{\partial g_{n+1}}{\partial y_1} \sum_{j=1}^{n+1} J_j \frac{\partial F_j}{\partial y_{n+1}} \ .$$

The first sum is J, and the second sum is the determinant of the matrix

$$\begin{bmatrix} \frac{\partial F_1}{\partial y_2} & \frac{\partial F_1}{\partial y_2} & \cdots & \frac{\partial F_1}{\partial y_{n+1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_{n+1}}{\partial y_2} & \frac{\partial F_{n+1}}{\partial y_2} & \cdots & \frac{\partial F_n}{\partial y_{n+1}} \end{bmatrix}$$

expanded by minors on the first column. This matrix has two identical columns, hence has determinant 0 (), so the second sum is 0. Similarly, the other sums are 0. Thus

$$J_i(\boldsymbol{x}, \boldsymbol{y}) \frac{\partial G}{\partial y_1}(\boldsymbol{x}, y_1) = J(\boldsymbol{x}, \boldsymbol{y})$$

for all $(\boldsymbol{x}, \boldsymbol{y})$; so $\frac{\partial G}{\partial y_1}(\boldsymbol{a}, b_1) \neq 0$ since $J(\boldsymbol{a}, \boldsymbol{b}) \neq 0$.

Then by VIII.5.3.15. there is an open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$ and an open set V_2 in $\mathbb{R}^{p+1}$ containing $(\boldsymbol{a}, b_1)$, $V_2 \subseteq W_1$, and a unique function $h : W \to \mathbb{R}$, such that $(\boldsymbol{x}, h(\boldsymbol{x})) \in V_2$ and $G(\boldsymbol{x}, h(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$. Set

$$V = \{(x_1, \ldots, x_p, y_1, \ldots, y_n) \in V_1 : (x_1, \ldots, x_p, y_1) \in V_2\}$$

and define

$$\boldsymbol{f}(\boldsymbol{x}) = (h(\boldsymbol{x}), g_2(\boldsymbol{x}, h(\boldsymbol{x})), \ldots, g_{n+1}(\boldsymbol{x}, h(\boldsymbol{x})))$$

for $\boldsymbol{x} \in W$. Then, for $\boldsymbol{x} \in W$, we have $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in V$, and $F_i(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = G(\boldsymbol{x}, h(\boldsymbol{x})) = 0$. And if $j \neq i$,

$$F_j(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = F_j(\boldsymbol{x}, h(\boldsymbol{x}), g_2(\boldsymbol{x}, h(\boldsymbol{x})), \ldots, g_{n+1}(\boldsymbol{x}, h(\boldsymbol{x}))) = 0 \ .$$

Thus $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{0}$.

The uniqueness of $\boldsymbol{f}$ follows immediately from the uniqueness of the g_k and of h, and continuity of $\boldsymbol{f}$ follows from continuity of the g_k and h. Continuity of $\frac{\partial \boldsymbol{f}}{\partial x_j}$ on W for $x_j \in S$ then follows from continuity of the partials of the g_k and h for $x_j \in S$.

The formula for $D_S \boldsymbol{f}$ now follows easily from the Chain Rule. We write the case $S = \{x_1, \ldots, x_p\}$; the case of general S is identical with notational complications (or can be obtained immediately by restricting coordinates). Define $\boldsymbol{H} : W \to \mathbb{R}^{p+n}$ by $\boldsymbol{H}(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$. Then $\boldsymbol{H}$ is $\mathcal{C}^1$ on W, and

$$D\boldsymbol{H}(\boldsymbol{x}) = \begin{bmatrix} I_p \\ D\boldsymbol{f}(\boldsymbol{x}) \end{bmatrix}$$

for $\boldsymbol{x} \in W$, where I_p is the $p \times p$ identity matrix and $D\boldsymbol{f}(\boldsymbol{x})$ is $n \times p$. Then by the Chain Rule we have

$$D(\boldsymbol{F} \circ \boldsymbol{H})(\boldsymbol{x}) = D\boldsymbol{F}(\boldsymbol{H}(\boldsymbol{x}))D\boldsymbol{H}(\boldsymbol{x}) = \begin{bmatrix} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) & D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \end{bmatrix} \begin{bmatrix} I_p \\ Df(\boldsymbol{x}) \end{bmatrix}$$

$$= D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) + D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))D\boldsymbol{f}(\boldsymbol{x})$$

for $\boldsymbol{x} \in W$. But $\boldsymbol{F} \circ \boldsymbol{H}$ is identically zero on W, so $D(\boldsymbol{F} \circ \boldsymbol{H})(\boldsymbol{x}) = \boldsymbol{0}$ for all $\boldsymbol{x} \in W$. Since $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ is invertible for $\boldsymbol{x} \in W$, we have, for $\boldsymbol{x} \in W$,

$$D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) + D\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0} \ .$$



This completes the proof of VIII.5.4.2., and hence of the continuous version VIII.5.1.6. (the case $S = \emptyset$) and the $\mathcal{C}^1$ smooth version VIII.5.1.7. (the case $S = \{x_1, \ldots, x_p\}$). 

VIII.5.4.3. The reason this proof requires continuity of the partials, and not just that the partials exist and $J(\boldsymbol{x}, \boldsymbol{y}) \neq 0$ in a neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$, is that not only must J be nonzero, but at least one of the minors J_i must also be nonzero in a neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$ to apply the inductive assumption. The continuity can be replaced by an assumption that not only is the J nonzero in a neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$, but also a decreasing chain of minors of smaller and smaller size are nonzero in a neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$. This version was explicitly stated in [You09], but is quite clumsy so we have stuck to the cleaner statement of VIII.5.4.2., which is adequate for almost all applications. For a full extension of VIII.5.3.10. to the n-dimensional case, see VIII.5.7.7.. It does not appear to be known whether the assumption of the decreasing sequence of nonzero minors is automatic (cf. Exercise VIII.5.11.14.(b)).

VIII.5.4.4. To finish the proof of the smooth version, it remains only to show that if $\boldsymbol{F}$ is $\mathcal{C}^r$, then $\boldsymbol{f}$ is also $\mathcal{C}^r$. The entries of $D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ and $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ are $\mathcal{C}^{r-1}$ by assumption. Since the determinant of $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, f(\boldsymbol{x}))$ is a polynomial in the entries, it is also $\mathcal{C}^{r-1}$ on W, and since the determinant is never 0 on W, it follows that the reciprocal of the determinant is also $\mathcal{C}^{r-1}$ on W. Thus the entries of $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1}$ are also $\mathcal{C}^{r-1}$ on W by the formula in III.11.10.45., and therefore so are the entries of

$$D\boldsymbol{f}(\boldsymbol{x}) = -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$$

which is equivalent to $\boldsymbol{f}$ being $\mathcal{C}^r$ on W.

This completes the full proof of the smooth version VIII.5.1.7.. 

VIII.5.5. Proof of the Inverse Function Theorem

The Inverse Function Theorem can now be proved from the Implicit Function Theorem as in (). In this section we give an alternate direct proof, which "explains" better why the theorem is true and which generalizes to other settings, such as Banach spaces (). This proof uses the Contraction Mapping Principle (). This argument first appeared, in somewhat disguised form, in [Gou03]; cf. Exercise VIII.5.11.11..

To prove the Inverse Function Theorem, it suffices the prove the following theorem. Differentiability of the inverse function then follows from VIII.5.1.29., and the formula there shows that the inverse function is $\mathcal{C}^1$; and if $\boldsymbol{f}$ is $\mathcal{C}^r$, the inverse function is also $\mathcal{C}^r$ by the argument in VIII.5.4.4..

VIII.5.5.1. THEOREM. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in \mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^n$ a $\mathcal{C}^1$ function. If $D\boldsymbol{f}(\boldsymbol{a})$ is invertible, then there is a neighborhood V of $\boldsymbol{a}$, $V \subseteq U$, such that

(i) $\boldsymbol{f}$ is one-to-one on V.

(ii) $W = \boldsymbol{f}(V)$ is open in $\mathbb{R}^n$.

(iii) $D\boldsymbol{f}(\boldsymbol{x})$ is invertible for all $\boldsymbol{x} \in V$.

VIII.5.5.2. The first reduction is to make an affine change of variables to make $\boldsymbol{a} = \boldsymbol{0}$, $\boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{0}$, and $D\boldsymbol{f}(\boldsymbol{0}) = I_n$, which greatly simplifies the notation of the rest of the proof. We define

$$\boldsymbol{T}(\boldsymbol{x}) = D\boldsymbol{f}(\boldsymbol{a})^{-1}(\boldsymbol{x}) \ .$$

Then $\boldsymbol{T}$ is an invertible linear function from $\mathbb{R}^n$ to $\mathbb{R}^n$; $\boldsymbol{T}$ is $\mathcal{C}^1$ and $D\boldsymbol{T}(\boldsymbol{x}) = D\boldsymbol{f}(\boldsymbol{a})^{-1}$ for all $\boldsymbol{x}$. Define

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{T}(\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{a})) - \boldsymbol{c}$$

for $\boldsymbol{x} \in U - \boldsymbol{a}$, where $\boldsymbol{c} = \boldsymbol{T}(\boldsymbol{f}(\boldsymbol{a}))$; then by the Chain Rule $\boldsymbol{g}$ is $\mathcal{C}^1$ on $U - \boldsymbol{a}$ and

$$D\boldsymbol{g}(\boldsymbol{x}) = D\boldsymbol{T}(\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{a}))D\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{a}) = D\boldsymbol{f}(\boldsymbol{a})^{-1}D\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{a})$$

for $\boldsymbol{x} \in U - \boldsymbol{a}$. In particular, $D\boldsymbol{g}(\boldsymbol{0}) = I_n$.

It suffices to prove VIII.5.5.1. for $\boldsymbol{g}$ on $U - \boldsymbol{a}$, since we would then have

$$\boldsymbol{f}^{-1}(\boldsymbol{y}) = \boldsymbol{g}^{-1}(\boldsymbol{T}(\boldsymbol{y}) - \boldsymbol{c}) + \boldsymbol{a}$$

on $T^{-1}(W + \boldsymbol{c})$ with values in $V + \boldsymbol{a}$.

So, changing notation, in VIII.5.5.1. we will assume $\boldsymbol{a} = \boldsymbol{0}$, $\boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{0}$, and $D\boldsymbol{f}(\boldsymbol{0}) = I_n$.

VIII.5.5.3. Now define $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x})$. Then $\boldsymbol{g}$ is $\mathcal{C}^r$. Since $D\boldsymbol{g}(\boldsymbol{0}) = \boldsymbol{0}$, there is an $\epsilon > 0$ such that $\|Dg_k(\boldsymbol{x})\| < \frac{1}{2n}$ for all k whenever $\|\boldsymbol{x}\| < \epsilon$. Set $W_1 = B_\epsilon(\boldsymbol{0})$. If $\boldsymbol{x}, \boldsymbol{z} \in W_1$, then by the MVT () for each k there is a $\boldsymbol{c}_k$ between $\boldsymbol{x}$ and $\boldsymbol{z}$ with $g_k(x) - g_k(z) = Dg_k(\boldsymbol{c}_k)(\boldsymbol{x} - \boldsymbol{z})$; $\boldsymbol{c}_k \in W_1$ since W_1 is convex. Thus

$$\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{z})\| \leq \sum_{k=1}^{n} |g_k(x) - g_k(z)| = \sum_{k=1}^{n} |Dg_k(\boldsymbol{c}_k)(\boldsymbol{x} - \boldsymbol{z})| \leq \sum_{k=1}^{n} \|Dg_k(\boldsymbol{c}_k)\|\|\boldsymbol{x} - \boldsymbol{z}\| < \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|$$

for $\boldsymbol{x}, \boldsymbol{z} \in W_1$. In particular, $\|\boldsymbol{g}(\boldsymbol{x})\| < \frac{1}{2}\|\boldsymbol{x}\|$ for all $\boldsymbol{x} \in W_1$.

VIII.5.5.4. Now fix η, $0 < \eta < \epsilon$, and fix $\boldsymbol{y} \in \mathbb{R}^n$ with $\|y\| \leq \frac{\eta}{2}$. Define $\boldsymbol{g}_{\boldsymbol{y}} : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{y} + \boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y} + \boldsymbol{g}(\boldsymbol{x}) \ .$$

If $\|\boldsymbol{x}\| \leq \eta$, then $\|\boldsymbol{g}(\boldsymbol{x})\| < \frac{\eta}{2}$, so $\|\boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{x})\| < \eta$; thus $\boldsymbol{g}_{\boldsymbol{y}}$ maps $\bar{B}_\eta(\boldsymbol{0})$ into itself. If $\boldsymbol{x}, \boldsymbol{z} \in \bar{B}_\eta(\boldsymbol{0})$, then

$$\|\boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{x}) - \boldsymbol{g}_{\boldsymbol{y}}(\boldsymbol{z})\| = \|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{z})\| < \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|$$

and so $\boldsymbol{g_y}$ is a contractive mapping of the complete metric space $\bar{B}_\eta(\mathbf{0})$ to itself. By the Contractive Mapping Principle (), $\boldsymbol{g_y}$ has a unique fixed point $\boldsymbol{x} \in \bar{B}_\eta(\mathbf{0})$, i.e. there is a unique $\boldsymbol{x}$ with $\|\boldsymbol{x}\| \leq \eta$ and $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$. Thus, if $W = B_{\epsilon/2}(\mathbf{0})$, for any $\boldsymbol{y} \in W$ there is a unique $\boldsymbol{x} \in B_\epsilon(\mathbf{0})$ with $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{y}$ (such a $\boldsymbol{y}$ is in $\bar{B}_{\eta/2}(\mathbf{0})$ for $2\|\boldsymbol{y}\| \leq \eta < \epsilon$). Set $V = B_\epsilon(\mathbf{0}) \cap \boldsymbol{f}^{-1}(W)$; then V is an open set in $\mathbb{R}^n$ containing $\mathbf{0}$ and $\boldsymbol{f}$ maps V bijectively to W.

This completes the proof of VIII.5.5.1.. 👍

VIII.5.5.5. Of course, this argument is not a full proof of the Inverse Function Theorem; the arguments of VIII.5.1.29. and VIII.5.4.4. are also needed to finish the proof.

VIII.5.6. How Do We Find The Implicit/Inverse Function?

VIII.5.6.1. In the approach we have taken, the Implicit/Inverse Function Theorems are pure existence theorems: the desired functions are shown to exist (uniquely and differentiably) on *some* neighborhood, but we have no control over the size of the neighborhood and seemingly no effective way of finding (computing) the function. A philosophy expressed in [HH99, p. 148] (with which I only partially agree) is:

> "In the days before computers, finding solutions [to nonlinear equations] was virtually impossible; even in the good cases, where mathematicians could prove that solutions existed, they were usually not concerned with whether their proof could be turned into a practical algorithm to find the solutions in question. The advent of computers has made such an abstract approach unreasonable. Knowing that a system of equations has solutions is no longer enough; we want a practical algorithm that will enable us to solve them."

(In fact, this philosophy largely underlies most of that book.) By "solving" the equations the authors mean solving them numerically to arbitrary accuracy in a specific region; it is impossible in principle to find exact closed-form solutions in most cases.

VIII.5.6.2. Actually, in the Implicit/Inverse Function Theorems, since we know a value for the solution at one point and an expression for the derivative, the solution can be computed nearby by an iterative numerical integration (convert to a system of differential equations, cf. XI.1.10.12.), although this is not very efficient numerically and can take one out of the region where a unique solution exists (). But it turns out that with slightly more restrictive hypotheses, both an explicit region in which the solution exists and an iterative procedure, based on Newton's Method, that converges efficiently to the solution are possible.

VIII.5.7. How Much Can the Theorems be Generalized?

VIII.5.7.1. One might hope that the hypotheses of the Inverse Function Theorem can be relaxed to only require differentiability at $\boldsymbol{a}$, or at least nondegeneracy of the derivative at $\boldsymbol{a}$. But this is not enough to force the function to be locally injective around $\boldsymbol{a}$, even in one dimension:

VIII.5.7.2. Example. Let

$$f(x) = \begin{cases} x + x^2 \sin \frac{1}{x^2} & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases}$$

for $x \in \mathbb{R}$. Then f is differentiable on $\mathbb{R}$ and $f'(0) = 1$, but f is not injective (strictly increasing) on any neighborhood of 0. In fact, f' takes all real number values (in particular, both positive and negative values) in any neighborhood of 0.

VIII.5.7.3. Suppose we *assume* $\boldsymbol{f}$ is injective in an open neighborhood V of $\boldsymbol{a}$, with $f(V)$ open, and that $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$ and $D\boldsymbol{f}(\boldsymbol{a})$ is invertible. Then there is an inverse function $\boldsymbol{g}$ defined on $\boldsymbol{f}(V)$. By VIII.5.1.29., $\boldsymbol{g}$ is differentiable at $\boldsymbol{f}(\boldsymbol{a})$.

VIII.5.7.4. However, one of the main points of the Inverse Function Theorem is that nondegeneracy of the derivative of $\boldsymbol{f}$ at $\boldsymbol{a}$ forces $\boldsymbol{f}$ to be injective in a neighborhood V of $\boldsymbol{a}$, with $\boldsymbol{f}(V)$ open. Continuity of the partial derivatives is seemingly used mainly to insure that the derivative is nondegenerate in an entire neighborhood of $\boldsymbol{a}$ (this can fail if the derivative is discontinuous; cf. VIII.5.7.2.). So what if this is just assumed? In the one-dimensional case, if the derivative of a function is nonzero in an open interval around a, then $\boldsymbol{f}$ must be strictly monotone in that interval by () and hence locally injective. Remarkably, this remains true in higher dimensions:

VIII.5.7.5. THEOREM. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in U$, and $\boldsymbol{f} : U \to \mathbb{R}^n$. Suppose $\boldsymbol{f}$ is differentiable at every point of U and that $D\boldsymbol{f}(\boldsymbol{x})$ is invertible for all $\boldsymbol{x} \in U$. Then there is a neighborhood V of $\boldsymbol{a}$ such that $\boldsymbol{f}$ is one-to-one on V, $\boldsymbol{f}(V)$ is open in $\mathbb{R}^n$, and the inverse function $\boldsymbol{g} : \boldsymbol{f}(V) \to V$ is differentiable at every point of $\boldsymbol{f}(V)$ with

$$D\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x})) = D\boldsymbol{f}(\boldsymbol{x})^{-1}$$

for all $x \in V$.

VIII.5.7.6. Curiously, this theorem appears to have been first stated and proved by J. SAINT RAYMOND in 2002 [Ray02], although it is implicit in an earlier result of CHERNAVSKII in 1964 [Čer64]; see also [Hal74], [Cri92]), and http://terrytao.wordpress.com/2011/09/12/the-inverse-function-theorem-for-everywhere-differe It should be pointed out, however, that as a practical matter this result is of limited use beyond the case covered in the usual Inverse Function Theorem if $n > 1$, since in most cases the only reasonable way to verify that a function is differentiable is to show that the partial derivatives are continuous (cf. ()), although there are cases where the partial derivatives are continuous except at one point $\boldsymbol{a}$ and it can be shown by brute force that the function is differentiable at $\boldsymbol{a}$ (cf. ()), and this version but not the usual version can be applied in such a situation (provided, of course, that the nondegeneracy condition is satisfied).

There is a corresponding version of the Implicit Function Theorem, proved from VIII.5.7.5. the same way VIII.5.1.7. is deduced from VIII.5.1.34. (). The one-dimensional version of this result is VIII.5.3.10.. I am not aware of any analogous version of the Continuous or Mixed Implicit Function Theorems; cf. Exercises VIII.5.11.13. and VIII.5.11.14.(b). There is also a corresponding version of VIII.5.2.1..

VIII.5.7.7. THEOREM. Let U be an open set in $\mathbb{R}^{p+n}$ and $(\boldsymbol{a}, \boldsymbol{b}) \in U$. Let $\boldsymbol{F} : U \to \mathbb{R}^n$ with $\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is differentiable at every point of U. If $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ is invertible, i.e. $det(D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})) \neq 0$, for all $(\boldsymbol{x}, \boldsymbol{y}) \in U$, then there is an open set V in $\mathbb{R}^{p+n}$ containing $(\boldsymbol{a}, \boldsymbol{b})$, an open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$, and

a unique function $\boldsymbol{f} : W \to \mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{b}$ and, for all $\boldsymbol{x} \in W$, $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in V$ and $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{0}$. Additionally, $\boldsymbol{f}$ is differentiable at every point of W, and for any $\boldsymbol{x} \in W$, we have

$$D\boldsymbol{f}(\boldsymbol{x}) = -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \ .$$

VIII.5.7.8. It should be noted that the results of this section are technically not "generalizations" of the usual Implicit and Inverse Function Theorems – they are distinct results (but obviously closely related) since both the hypotheses and conclusions are weaker. A true generalization of a theorem is a new result for which the original theorem is a special case, generally consisting of obtaining the same, or stronger, conclusions under weaker hypotheses (here "weaker" means hypotheses which hold in actually or potentially more situations); obtaining stronger conclusions under the same hypotheses could also be considered a type of generalization, although such a result is usually just called a stronger theorem and not a generalization (() is a "generalization" of VIII.5.1.7. (in the case $r \geq 2$) in this sense). Thus VIII.5.3.12. is a true generalization of both VIII.5.3.4. and VIII.5.3.9., and VIII.5.4.2. a generalization of VIII.5.1.6. and VIII.5.1.7. ($\mathcal{C}^1$ version). The only other true generalizations I know of the Implicit and Inverse Function Theorems are extensions to maps between more general spaces like Banach spaces (cf. XXI.1.2.3.), or results like VIII.5.2.1..

VIII.5.8. The Domain and Range Straightening Theorems

From the Implicit and Inverse Function Theorems we can get a simple general local picture of a differentiable function, especially one of constant rank: up to a smooth change of coordinates, such a function of constant rank k from $\mathbb{R}^n$ to $\mathbb{R}^m$ locally consists simply of projection onto a k-dimensional subspace of $\mathbb{R}^n$ followed by a linear embedding of this subspace into $\mathbb{R}^m$.

The first result, a variation of VIII.5.2.1., is called the *Domain-Straightening Theorem* or *Local Submersion Theorem* in some references. If $\boldsymbol{f}$ is a smooth function from $U \subseteq \mathbb{R}^n$ to $\mathbb{R}^m$ and $\boldsymbol{a} \in U$, then $\boldsymbol{a}$ is called a *regular point* of $\boldsymbol{f}$ if the rank of $D\boldsymbol{f}(\boldsymbol{a})$ is m. (If $\boldsymbol{f}$ has a regular point, then necessarily $n \geq m$.) The set of regular points of $\boldsymbol{f}$ is an open set.

VIII.5.8.1. Theorem. [Domain-Straightening or Local Submersion Theorem] Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function ($1 \leq r \leq \infty$) from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^m$, $n \geq m$, and $\boldsymbol{a}$ a regular point of $\boldsymbol{f}$. Then there is an open neighborhood U of $\boldsymbol{a}$ and a $\mathcal{C}^r$ diffeomorphism $\boldsymbol{g}$ from a neighborhood U' of $(\boldsymbol{b}, \boldsymbol{0})$ in $\mathbb{R}^n \cong \mathbb{R}^m \times \mathbb{R}^{n-m}$ to U with $\boldsymbol{g}(\boldsymbol{b}, \boldsymbol{0}) = \boldsymbol{a}$ and $\boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in U'$, i.e.

$$\boldsymbol{f} \circ \boldsymbol{g}(x_1, \ldots, x_n) = (x_1, \ldots, x_m)$$

for all $(x_1, \ldots, x_n) \in U'$.

The name "Domain-Straightening Theorem" is appropriate since the theorem says that local coordinates can be chosen around $\boldsymbol{a}$ so that $\boldsymbol{f}$ just picks out the first m coordinates of a point, and thus the preimage of any point in $\mathbb{R}^m$ in a neighborhood of $\boldsymbol{b}$ is a piece of a vertical $(n-m)$-dimensional hyperplane. The case $n = m$ is just the Inverse Function Theorem. See () for an explanation of the name "Local Submersion Theorem."

Proof: By precomposing $\boldsymbol{f}$ with a translation, there is no loss of generality in assuming $\boldsymbol{a} = \boldsymbol{0}$. By assumption, the column rank of $D\boldsymbol{f}(\boldsymbol{0})$ is m, so the columns of $D\boldsymbol{f}(\boldsymbol{0})$ can be permuted so that the first m

columns are linearly independent. Thus $\boldsymbol{f}$ can be precomposed with a linear transformation T (permutation of coordinates) so that, if we write $\mathbb{R}^n$ as $\mathbb{R}^m \times \mathbb{R}^{n-m}$ and points in $\mathbb{R}^n$ as $(\boldsymbol{x}, \boldsymbol{y})$, with $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^{n-m}$, we have $D_{\boldsymbol{x}}(\boldsymbol{f} \circ T)(\boldsymbol{0}, \boldsymbol{0})$ invertible. Thus applying VIII.5.2.1. with the roles of $\boldsymbol{x}$ and $\boldsymbol{y}$ interchanged, there is an open neighborhood V of $(\boldsymbol{b}, \boldsymbol{0})$ in $\mathbb{R}^n$ and a $\mathcal{C}^r$ function $\boldsymbol{h} : V \to \mathbb{R}^m$ such that $\boldsymbol{h}(\boldsymbol{b}, \boldsymbol{0}) = (\boldsymbol{0}, \boldsymbol{0})$ and $(\boldsymbol{f} \circ T)(\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y}) = \boldsymbol{x}$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in V$. The function $\phi : V \to \mathbb{R}^n$ defined by $\phi(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{y})$ has $D\phi(\boldsymbol{0}, \boldsymbol{0})$ invertible, so by the Inverse Function Theorem there is an open neighborhood U' of $(\boldsymbol{b}, \boldsymbol{0})$ for which $\phi(U') \subseteq V$ and $\phi|_{U'}$ is a $\mathcal{C}^r$ diffeomorphism from U' to $\phi(U')$. Set $U = T(\phi(U'))$ and $\boldsymbol{g} = T \circ \phi$. 👍

VIII.5.8.2. Corollary. [Smooth Local Open Mapping Theorem] Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function ($r \geq 1$) from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$ ($n \geq m$), $\boldsymbol{a}$ a regular point of $\boldsymbol{f}$, and $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$. Then $\boldsymbol{a}$ has an open neighborhood U such that every open neighborhood of $\boldsymbol{a}$ contained in U maps onto an open neighborhood of $\boldsymbol{b}$, and $\boldsymbol{f}^{-1}(\boldsymbol{b}) \cap U$ is an $(n-m)$-dimensional $\mathcal{C}^r$ hypersurface in U. In particular, the range of $\boldsymbol{f}$ contains an open set around $\boldsymbol{f}(\boldsymbol{a})$.

Proof: Let U and $\boldsymbol{g}$ be as in VIII.5.8.1.. Then

$$\boldsymbol{f}^{-1}(\boldsymbol{b}) \cap U = \boldsymbol{g}((\boldsymbol{f} \circ \boldsymbol{g})^{-1}(\boldsymbol{b})) = \boldsymbol{g}\left(\{(b_1, \ldots, b_m, x_{m+1}, \ldots, x_n) : x_{m+1}, \ldots, x_n \in \mathbb{R}\} \cap U'\right)$$

where $\boldsymbol{b} = (b_1, \ldots, b_m)$. 👍

VIII.5.8.3. Conversely, it is a remarkable fact (**??**, a consequence of Sard's Theorem) that if the range of $\boldsymbol{f}$ contains a nonempty open set in $\mathbb{R}^m$, and $\boldsymbol{f}$ is sufficiently smooth, then $\boldsymbol{f}$ has "many" regular points.

By straightening only some of the coordinates in the domain, we obtain:

VIII.5.8.4. Theorem. Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function ($1 \leq r \leq \infty$) from an open subset Ω of $\mathbb{R}^n$ to $\mathbb{R}^m$, and $\boldsymbol{a}$ a point of the domain of $\boldsymbol{f}$. Suppose $D\boldsymbol{f}(\boldsymbol{a})$ has rank $\geq k$, $k < m$. Then there is an open neighborhood U of $\boldsymbol{a}$, a point $\boldsymbol{b} \in \mathbb{R}^k$, and a $\mathcal{C}^r$ diffeomorphism $\boldsymbol{g}$ from a neighborhood U' of $(\boldsymbol{b}, \boldsymbol{0})$ in $\mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k}$ to U with $\boldsymbol{g}(\boldsymbol{b}, \boldsymbol{0}) = \boldsymbol{a}$, an invertible linear transformation $T : \mathbb{R}^m \to \mathbb{R}^m$ (permutation of coordinates) with $T \circ \boldsymbol{f}(\boldsymbol{a}) = (\boldsymbol{b}, \boldsymbol{c})$ for some $\boldsymbol{c} \in \mathbb{R}^{m-k}$, and a $\mathcal{C}^r$ function $\boldsymbol{f}_0 : \mathbb{R}^n \to \mathbb{R}^{m-k}$ defined on a neighborhood of $(\boldsymbol{b}, \boldsymbol{0})$ such that

$$T \circ \boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{y}))$$

for all $(\boldsymbol{x}, \boldsymbol{y}) \in U' \subseteq \mathbb{R}^k \times \mathbb{R}^{n-k}$, i.e.

$$T \circ \boldsymbol{f} \circ \boldsymbol{g}(x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) = (x_1, \ldots, x_k, \boldsymbol{f}_0(x_1, \ldots, x_n))$$

for all $(x_1, \ldots, x_n) \in U'$. We have $\boldsymbol{f}_0(\boldsymbol{b}, \boldsymbol{0}) = \boldsymbol{c}$.

Proof: The matrix $D\boldsymbol{f}(\boldsymbol{a})$ has rank at least k, so by permuting rows we may make the first k rows linearly independent. Let T be such a coordinate permutation, and let ϕ be the function from Ω to $\mathbb{R}^k$ whose coordinate functions are the first k coordinate functions of $T \circ \boldsymbol{f}$. Set $\boldsymbol{b} = \phi(\boldsymbol{a})$. Then ϕ is $\mathcal{C}^r$ and satisfies the hypotheses of VIII.5.8.1., so there is a $\mathcal{C}^r$ diffeomorphism $\boldsymbol{g}$ from a neighborhood U' of $(\boldsymbol{b}, \boldsymbol{0})$ in $\mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k}$ to a neighborhood U of $\boldsymbol{a}$ with $\phi \circ \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in U'$. Then $T \circ \boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$ is of the form

$$(\phi \circ \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{y})) = (\boldsymbol{x}, \boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{y}))$$

for $(\boldsymbol{x}, \boldsymbol{y}) \in U'$, for some $\mathcal{C}^r$ function $\boldsymbol{f}_0 : U' \to \mathbb{R}^{m-k}$. 👍

VIII.5.8.5. Suppose $\boldsymbol{f}$ is as in the statement of the theorem. If the rank of $D\boldsymbol{f}(\boldsymbol{z})$ is exactly k for all $\boldsymbol{z}$ in a neighborhood of $\boldsymbol{a}$, we can say more. Set $\boldsymbol{h} = T \circ \boldsymbol{f} \circ \boldsymbol{g}$. Then $\boldsymbol{h}$ has rank exactly k in a neighborhood of $(\boldsymbol{b}, \boldsymbol{0})$. The derivative of $\boldsymbol{h}$ at $(\boldsymbol{x}, \boldsymbol{y})$ is of the form

$$\begin{bmatrix} I_k & 0 \\ B & A \end{bmatrix}$$

where $D\boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{y}) = [\, B \; A \,]$, B is $(m-k) \times k$, and A is $(m-k) \times (n-k)$. For such a matrix to have rank k, A must be 0: the first k columns are linearly independent, and if A has a nonzero entry the column it appears in will be linearly independent from the first k columns, a contradiction. Thus the partial derivatives of $\boldsymbol{f}_0$ in the $\boldsymbol{y}$ variables are identically zero near $(\boldsymbol{b}, \boldsymbol{0})$, i.e. $\boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{y})$ depends only on the $\boldsymbol{x}$ variables in a neighborhood of $(\boldsymbol{b}, \boldsymbol{0})$. Thus we obtain:

VIII.5.8.6. Theorem. [Rank Theorem] Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function $(1 \leq r \leq \infty)$ from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^m$, and $\boldsymbol{a}$ a point of the domain of $\boldsymbol{f}$. Suppose $D\boldsymbol{f}$ has constant rank k in a neighborhood of $\boldsymbol{a}$. Then there is an open neighborhood U of $\boldsymbol{a}$, a point $\boldsymbol{b} \in \mathbb{R}^k$, a $\mathcal{C}^r$diffeomorphism $\boldsymbol{g}$ from an open neighborhood U' of $(\boldsymbol{b}, \boldsymbol{0})$ in $\mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k}$ to U, an open neighborhood W of $\boldsymbol{b}$ in $\mathbb{R}^k$ such that $\boldsymbol{x} \in W$ whenever $(\boldsymbol{x}, \boldsymbol{y}) \in U'$, and a $\mathcal{C}^r$ function $\tilde{\boldsymbol{f}} : W \to \mathbb{R}^m$ such that

$$\boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = \tilde{\boldsymbol{f}}(\boldsymbol{x})$$

for all $(\boldsymbol{x}, \boldsymbol{y}) \in U'$.

In fact, since $\boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{0})$ for any $\boldsymbol{y}$, we can take $\tilde{\boldsymbol{f}}(\boldsymbol{x}) = T^{-1}(\boldsymbol{x}, \boldsymbol{f}_0(\boldsymbol{x}, \boldsymbol{0}))$.

In the case $k = n < m$ in VIII.5.8.4., the T is unnecessary and can be eliminated, and the Rank Theorem applies (the case $k = n = m$ is just the Inverse Function Theorem). In this case a better result holds, called the *Range-Straightening Theorem* or *Local Immersion Theorem*:

VIII.5.8.7. Theorem. [Range-Straightening or Local Immersion Theorem] Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function $(1 \leq r \leq \infty)$ from an open subset Ω of $\mathbb{R}^n$ to $\mathbb{R}^m$, and $\boldsymbol{a}$ a point of the domain of $\boldsymbol{f}$. Suppose $D\boldsymbol{f}(\boldsymbol{a})$ has rank n (then necessarily $n \leq m$). Set $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$. Then there are open neighborhoods U of $\boldsymbol{a}$ in $\mathbb{R}^n$ and V of $\boldsymbol{b}$ in $\mathbb{R}^m$, and a $\mathcal{C}^r$ diffeomorphism $\boldsymbol{h}$ from V to a neighborhood V' of $(\boldsymbol{a}, \boldsymbol{0})$ in $\mathbb{R}^m$ with $\boldsymbol{h}(\boldsymbol{b}) = (\boldsymbol{a}, \boldsymbol{0})$, such that

$$\boldsymbol{h} \circ \boldsymbol{f}(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{0})$$

for all $\boldsymbol{x} \in U$, i.e.

$$\boldsymbol{h} \circ \boldsymbol{f}(x_1, \ldots, x_n) = (x_1, \ldots, x_n, 0, \ldots, 0)$$

for all $(x_1, \ldots, x_n) \in U$.

The name "Range-Straightening Theorem" means that a smooth function from $\mathbb{R}^n$ to $\mathbb{R}^m$ $(n < m)$ which has rank n at a point $\boldsymbol{a}$ has local range which up to a smooth change of coordinates is a piece of a horizontal n-dimensional hyperplane in $\mathbb{R}^m$. See VIII.5.8.10. for a precise and more general result. See () for an explanation of the name "Local Immersion Theorem."

Proof: Since $D\boldsymbol{f}(\boldsymbol{a})$ has row rank n, there is an invertible linear transformation $T : \mathbb{R}^m \to \mathbb{R}^m$ (permutation of coordinates) so that the first n rows of $D(T \circ \boldsymbol{f})(\boldsymbol{a})$ are linearly independent. Define $\boldsymbol{F} : \Omega \times \mathbb{R}^{m-n} \subseteq \mathbb{R}^m \to$

$\mathbb{R}^m$ by $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{z}) = T(\boldsymbol{f}(\boldsymbol{x}))+(\boldsymbol{0},\boldsymbol{z})$ (where $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{z} \in \mathbb{R}^{m-n}$, and the last expression uses $\mathbb{R}^m \cong \mathbb{R}^n \times \mathbb{R}^{m-n}$). Then $D\boldsymbol{F}(\boldsymbol{a},\boldsymbol{0})$ is of the form

$$\begin{bmatrix} A & 0 \\ B & I_{m-n} \end{bmatrix}$$

where the first column is $D(T \circ \boldsymbol{f})(\boldsymbol{a})$; since A is invertible, $D\boldsymbol{F}(\boldsymbol{a},\boldsymbol{0})$ is invertible. Thus by the Inverse Function Theorem there is an inverse function $\boldsymbol{G}$ defined on a neighborhood V'' of $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{0}) = T(\boldsymbol{f}(\boldsymbol{a}))$ to a neighborhood V' of $(\boldsymbol{a},\boldsymbol{0})$. Set $V = T^{-1}(V'')$ and $\boldsymbol{h} = \boldsymbol{G} \circ T$. If $U = \{\boldsymbol{x} \in \Omega : (\boldsymbol{x},\boldsymbol{0}) \in V'\}$, then

$$\boldsymbol{h} \circ \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{G} \circ T \circ \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{G}(T(\boldsymbol{f}(\boldsymbol{x})) + (\boldsymbol{0},\boldsymbol{0})) = \boldsymbol{G}(\boldsymbol{F}(\boldsymbol{x},\boldsymbol{0})) = (\boldsymbol{x},\boldsymbol{0})$$

for all $\boldsymbol{x} \in U$. 👍

Putting together the Range-Straightening Theorem with the Rank Theorem (which depends on the Domain-Straightening Theorem), we obtain the ultimate result (this result is called the Rank Theorem or the Constant Rank Theorem in some references):

VIII.5.8.8. THEOREM. [STRAIGHTENING THEOREM] Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function ($1 \leq r \leq \infty$) from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^m$, and $\boldsymbol{a}$ a point of the domain of $\boldsymbol{f}$. Suppose $D\boldsymbol{f}$ has constant rank k in a neighborhood of $\boldsymbol{a}$. Then there is an open neighborhood U of $\boldsymbol{a}$ and an open neighborhood V of $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$ with $\boldsymbol{f}(U) \subseteq V$, and $\mathcal{C}^r$ diffeomorphisms $\boldsymbol{g}$ from a neighborhood U' of $(\boldsymbol{0},\boldsymbol{0})$ in $\mathbb{R}^n \cong \mathbb{R}^k \times \mathbb{R}^{n-k}$ to U and $\boldsymbol{h}$ from V to a neighborhood V' of $(\boldsymbol{0},\boldsymbol{0})$ in $\mathbb{R}^m \cong \mathbb{R}^k \times \mathbb{R}^{m-k}$ with $\boldsymbol{g}(\boldsymbol{0},\boldsymbol{0}) = \boldsymbol{a}$, $\boldsymbol{h}(\boldsymbol{b}) = (\boldsymbol{0},\boldsymbol{0})$, and

$$\boldsymbol{h} \circ \boldsymbol{f} \circ \boldsymbol{g}(x_1,\ldots,x_k,x_{k+1},\ldots,x_n) = (x_1,\ldots,x_k,0,\ldots,0) \ .$$

PROOF: By pre- and post-composing $\boldsymbol{f}$ with translations, there is no loss of generality in assuming that $\boldsymbol{a} = \boldsymbol{0}$ and that $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{0}) = \boldsymbol{0}$. By the Rank Theorem, there is a $\mathcal{C}^r$ diffeomorphism $\boldsymbol{g}$, defined on an open neighborhood U'' of $\boldsymbol{0}$ in $\mathbb{R}^n$, an open neighborhood W of $\boldsymbol{0}$ in $\mathbb{R}^k$, and a $\mathcal{C}^r$ function $\tilde{\boldsymbol{f}} : W \to \mathbb{R}^m$, such that $\boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) = \tilde{\boldsymbol{f}}(\boldsymbol{x})$ for all $(\boldsymbol{x},\boldsymbol{y}) \in U''$. Then $\tilde{\boldsymbol{f}}$ satisfies the hypotheses of the Range-Straightening Theorem, so there is a neighborhood W' of $\boldsymbol{0}$ in $\mathbb{R}^k$, contained in W, an open neighborhood V of $\boldsymbol{0}$ in $\mathbb{R}^m$, and a $\mathcal{C}^r$ diffeomorphism $\boldsymbol{h}$ from V to an open neighborhood V' of $\boldsymbol{0}$ such that

$$\boldsymbol{h} \circ \tilde{\boldsymbol{f}}(\boldsymbol{x}) = (\boldsymbol{x},\boldsymbol{0})$$

for all $\boldsymbol{x} \in W'$. Set

$$U' = \{(\boldsymbol{x},\boldsymbol{y}) \in U'' : \boldsymbol{x} \in W'\} \ .$$

Then U' is an open neighborhood of $\boldsymbol{0}$ in $\mathbb{R}^n$, and

$$\boldsymbol{h} \circ \boldsymbol{f} \circ \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x},\boldsymbol{0})$$

for all $(\boldsymbol{x},\boldsymbol{y}) \in U'$. Set $U = \boldsymbol{g}(U')$. 👍

VIII.5.8.9. Note that this theorem automatically applies if the rank of $D\boldsymbol{f}(\boldsymbol{a})$ is m (i.e. $\boldsymbol{a}$ is a regular point of $\boldsymbol{f}$), since it is rank $\leq m$ everywhere and at least m in some neighborhood of $\boldsymbol{a}$ (). In this case the $\boldsymbol{h}$ can be simply be taken to be a translation. Thus the Straightening Theorem is a generalization of the Domain-Straightening Theorem VIII.5.8.1.. A similar observation holds if $D\boldsymbol{f}(\boldsymbol{a})$ has rank n, so the Range-Straightening Theorem VIII.5.8.7. is also a special case of the Straightening Theorem. (And the Inverse Function Theorem is the case $k = n = m$.)

VIII.5.8.10. Corollary. Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function $(1 \leq r \leq \infty)$ from an open subset of $\mathbb{R}^n$ to $\mathbb{R}^m$, and $\boldsymbol{a}$ a point of the domain of $\boldsymbol{f}$. Suppose $D\boldsymbol{f}$ has constant rank k in a neighborhood of $\boldsymbol{a}$. Then there is an open neighborhood U of $\boldsymbol{a}$ such that $\boldsymbol{f}(U)$ is a k-dimensional $\mathcal{C}^r$ hypersurface in $\mathbb{R}^m$ passing through $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$.

Although the Straightening Theorem only applies at points where the function is locally of constant rank (i.e. has constant rank in a neighborhood of the point), this set of points is "generic":

VIII.5.8.11. Proposition. Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function $(1 \leq r \leq \infty)$ from an open subset U of $\mathbb{R}^n$ to $\mathbb{R}^m$. Then the set of points where $\boldsymbol{f}$ is locally of constant rank is a dense open subset of U.

Proof: For $0 \leq k \leq n$ let V_k be the set of points of U where $\boldsymbol{f}$ has rank $\geq k$. Then each V_k is open, and the set W of points where $\boldsymbol{f}$ is locally of constant rank is

$$W = \bigcup_{k=0}^{n} \left(V_k \setminus \bar{V}_{k+1}\right) \ .$$

Thus W is open, and $U \setminus W$ is

$$U \setminus W = U \cap \left[\bigcup_{k=0}^{n} \left(\bar{V}_k \setminus V_k\right)\right] \ .$$

Each $\bar{V}_k \setminus V_k = \partial V_k$ is closed and nowhere dense, hence its complement is dense and open; and a finite intersection of dense open sets is dense and open (XIII.2.3.10.). 👍

VIII.5.8.12. Remarks. (i) The only result in this section which applies to an $\boldsymbol{f}$ which is not necessarily of locally constant rank around $\boldsymbol{a}$ is VIII.5.8.4.. But the Straightening Theorem applies at "almost every" point by VIII.5.8.11..

(ii) The Domain-Straightening Theorem and Range-Straightening Theorem are stated in such a way that only composition of $\boldsymbol{f}$ on one side with a diffeomorphism is necessary. For the ultimate Straightening Theorem, composition on both sides is necessary in general.

(iii) Much of the notational complication of the results and proofs in this section comes from the need to carefully specify and name exact neighborhoods on which functions are defined and formulas are valid. It is probably a good idea for a reader to at first ignore these precise neighborhoods, replacing each instance simply by "on some neighborhood of $\boldsymbol{a}$ (etc.)" or just "near $\boldsymbol{a}$" (as is done in the text in a few places where precise specification of the neighborhood is unnecessary) to follow the essence of the argument without distraction; at the end of the day, the exact neighborhoods are unimportant, since the real conclusion is merely that there are *some* neighborhoods on which the results hold.

VIII.5.9. The Degenerate Case

We now put together the previous results to describe a common way the Implicit Function Theorem can be thought of and used. We will change notation a little at the beginning, and get back closer to the previous notation as we go. We will be a little informal during the discussion, but end up with a carefully stated theorem.

VIII.5.9.1. Suppose we have a general problem of solving a system of functional equations

$$\begin{aligned} F_1(u_1,\dots,u_q) &= v_1 \\ F_2(u_1,\dots,u_q) &= v_2 \\ &\vdots \\ F_m(u_1,\dots,u_q) &= v_m \end{aligned}$$

for some of the u's as functions of the v's and the other u's. The system can be abbreviated

$$\boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{v}$$

where $\boldsymbol{u} = (u_1,\dots,u_q) \in \mathbb{R}^q$, $\boldsymbol{v} = (v_1,\dots,v_m) \in \mathbb{R}^m$, $\boldsymbol{F} = (F_1,\dots,F_m) : \mathbb{R}^q \to \mathbb{R}^m$. We suppose $\boldsymbol{F}$ is defined and $\mathcal{C}^r$ $(r \geq 1)$ on an open set $U \subseteq \mathbb{R}^q$, and $\boldsymbol{u}_0 \in U$ with $\boldsymbol{F}(\boldsymbol{u}_0) = \boldsymbol{v}_0$. We want a solution of $\boldsymbol{F}(\boldsymbol{u}) = \boldsymbol{v}$ near $(\boldsymbol{u}_0, \boldsymbol{v}_0)$.

VIII.5.9.2. We will assume that $D\boldsymbol{F}(\boldsymbol{u})$ has rank n at all points in a neighborhood of $\boldsymbol{u}_0$ (this entails $n \leq \min(q,m)$). By reducing the size of U, we may and will assume that the rank is n for all $\boldsymbol{u} \in U$. We may have $n = q$ and/or $n = m$; in these cases, the following discussion will have to be slightly modified (simplified) notationally. The case $n = m$ is exactly the nondegenerate case treated in VIII.5.2.1.. We will try to solve for n of the u's in terms of the v's and the rest of the u's (linear algebra results show that we cannot hope to solve for more than n of the u's even if $\boldsymbol{F}$ is linear).

VIII.5.9.3. We may then reorder the v's and rename them $(z_1,\dots,z_n,w_1,\dots,w_k)$ $(k = m-n)$, writing $\boldsymbol{z} = (z_1,\dots,z_n) \in \mathbb{R}^n$, $\boldsymbol{w} = (w_1,\dots,w_k) \in \mathbb{R}^k$, $\Phi(\boldsymbol{u}) = (F_1(\boldsymbol{u}),\dots,F_n(\boldsymbol{u})) : U \to \mathbb{R}^n$, $\Psi(\boldsymbol{u}) = (F_{n+1}(\boldsymbol{u}),\dots,F_m(\boldsymbol{u})) : U \to \mathbb{R}^k$, so $\boldsymbol{v} = (\boldsymbol{z},\boldsymbol{w})$, $\boldsymbol{F} = (\Phi,\Psi)$, in such a way that $D\Phi(\boldsymbol{u}_0)$ has rank n. This is simply a matter of choosing n linearly independent rows in $D\boldsymbol{F}(\boldsymbol{u}_0)$, which can be done in several ways in general but there is always at least one way. We then have that $D\Phi(\boldsymbol{u})$ has rank n in a neighborhood of $\boldsymbol{u}_0$, and by further reducing the size of U we may and will assume it has rank n at all points of U. Write $\boldsymbol{v}_0 = (\boldsymbol{c},\boldsymbol{d})$.

VIII.5.9.4. The system $\Phi(\boldsymbol{u}) = \boldsymbol{z}$ then essentially falls under the case of VIII.5.2.1., but one more notational choice is needed. We reorder and rename the u's as $(x_1,\dots,x_p,y_1,\dots,y_n)$ $(p = q-n)$ and write $\boldsymbol{x} = (x_1,\dots,x_p) \in \mathbb{R}^p$, $\boldsymbol{y} = (y_1,\dots,y_n) \in \mathbb{R}^n$, $\boldsymbol{u} = (\boldsymbol{x},\boldsymbol{y})$, in such a way that $D_{\boldsymbol{y}}\Phi(\boldsymbol{u}_0)$ has rank n, i.e. is invertible. (This amounts to choosing n linearly independent columns in $D\Phi(\boldsymbol{u}_0)$, which can always be done, in many ways in general.) Write $\boldsymbol{u}_0 = (\boldsymbol{a},\boldsymbol{b})$. We have that $D_{\boldsymbol{y}}\Phi(\boldsymbol{u})$ is invertible for all $\boldsymbol{u}$ in a neighborhood of $\boldsymbol{u}_0$; we may again reduce U so this is true for all $\boldsymbol{u} \in U$. Our system

$$\Phi(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{z}$$

then exactly satisfies the hypotheses of VIII.5.2.1..

VIII.5.9.5. There is thus a unique function $\boldsymbol{G}$, defined in a neighborhood of $(\boldsymbol{a}, \boldsymbol{c})$ with range in a suitable neighborhood of $\boldsymbol{b}$, such that $\boldsymbol{G}(\boldsymbol{a}, \boldsymbol{c}) = \boldsymbol{b}$ and $\Phi(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) = \boldsymbol{z}$ for all $(\boldsymbol{x}, \boldsymbol{z})$ near $(\boldsymbol{a}, \boldsymbol{c})$. We have thus written $y_1, \ldots, y_n$ as functions of $\boldsymbol{x}$ and $\boldsymbol{z}$, i.e. as functions of $\boldsymbol{x}$ and $\boldsymbol{v}$, as desired.

VIII.5.9.6. We have so far not used the last k equations. So how do they come into the picture? For one thing, we could in general have chosen to use some of the w's in place of some of the z's, and if we did that we would have the y's (or, in general, possibly a different set of y's) written in a different way as functions of the x's and v's – in the degenerate case there is no uniqueness possible in general for how to "solve" the equations, even for a single choice of y's.

VIII.5.9.7. If we stick to the same choice of z's, w's, $x's$, and y's, though, the additional equations serve as a system of *constraints*: The equation $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{z}, \boldsymbol{w})$ does not have a solution for all $(\boldsymbol{z}, \boldsymbol{w})$ near $(\boldsymbol{c}, \boldsymbol{d})$, only for $(\boldsymbol{z}, \boldsymbol{w})$ satisfying a relation. By the Straightening Theorem (), the range of $\boldsymbol{F}$ on a suitably small neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$ (which we may take to be U by further reduction) is an n-dimensional $\mathcal{C}^r$-submanifold Γ of $\mathbb{R}^m$. For every $\boldsymbol{z}$ near $\boldsymbol{c}$ there is a unique point of Γ whose first n coordinates are $\boldsymbol{z}$; thus the last k coordinates $\boldsymbol{w}$ can be considered a function of $\boldsymbol{z}$, i.e. Γ is the graph of a function $\boldsymbol{H}$ from an open set in $\mathbb{R}^n$ containing $\boldsymbol{c}$ to $\mathbb{R}^k$, with $\boldsymbol{H}(\boldsymbol{c}) = \boldsymbol{d}$. Since Γ is a $\mathcal{C}^r$-manifold, it follows that $\boldsymbol{H}$ is also $\mathcal{C}^r$. We thus have the following theorem, leaving out the steps of choosing the x's, y's, z's, w's, in which nothing needs proof except some basic linear algebra facts, and the reductions in the size of U, which need ().

VIII.5.9.8. Theorem. [Implicit/Inverse Function Theorem, Degenerate Case] Let $n, r \in \mathbb{N}$, $p, k \in \mathbb{N} \cup \{0\}$, U an open set in $\mathbb{R}^{p+n}$, $\boldsymbol{F} = (\Phi, \Psi) : U \to \mathbb{R}^{n+k}$ with $\Phi : U \to \mathbb{R}^n$ and $\Psi : U \to \mathbb{R}^k$ $\mathcal{C}^r$ functions, $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$, $\boldsymbol{d} \in \mathbb{R}^k$, $(\boldsymbol{a}, \boldsymbol{b}) \in U$, $\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{c}, \boldsymbol{d})$. Suppose $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ and $D_{\boldsymbol{y}}\Phi(\boldsymbol{x}, \boldsymbol{y})$ have rank n for all $(\boldsymbol{x}, \boldsymbol{y}) \in U$. Then there are open neighborhoods $\mathcal{X}$ of $\boldsymbol{a}$, $\mathcal{Y}$ of $\boldsymbol{b}$, $\mathcal{Z}$ of $\boldsymbol{c}$, and $\mathcal{W}$ of $\boldsymbol{d}$, with $\mathcal{X} \times \mathcal{Y} \subseteq U$, and unique functions $\boldsymbol{G} : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ with $\boldsymbol{G}(\boldsymbol{a}, \boldsymbol{c}) = \boldsymbol{b}$, and $\boldsymbol{H} : \mathcal{Z} \to \mathcal{W}$ with $\boldsymbol{H}(\boldsymbol{c}) = \boldsymbol{d}$, such that the equation

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{z}, \boldsymbol{w})$$

or, equivalently, the system of equations

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{z}$$
$$\Psi(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{w}$$

is equivalent to the system

$$\boldsymbol{y} = \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})$$
$$\boldsymbol{w} = \boldsymbol{H}(\boldsymbol{z})$$

on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{W}$, in the sense that if $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{W}$, then $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w})$ satisfies one system if and only if it satisfies the other. Additionally, $\boldsymbol{G}$ and $\boldsymbol{H}$ are $\mathcal{C}^r$ on their domains.
If $p = 0$ the $\boldsymbol{x}$ and $\boldsymbol{a}$ are omitted from the statement, and if $k = 0$ the $\boldsymbol{w}$, $\boldsymbol{d}$, and $\boldsymbol{H}$ are eliminated, with the statement notation adjusted accordingly.

VIII.5.9.9. The case $k = 0$ is exactly VIII.5.2.1., and the case $p = k = 0$ is exactly the Inverse Function Theorem.

VIII.5.9.10. In the situation of VIII.5.9.1., if the rank of $D\boldsymbol{F}$ is not locally constant around $\boldsymbol{u}_0$, there is no solution to the system in general, at least in the sense we have discussed.

VIII.5.10. Analytic Versions

There are analytic versions of most of the results of this section:

VIII.5.10.1. Theorem. [Implicit Function Theorem, Real Analytic Version] Let U be an open set in $\mathbb{R}^{p+n}$ and $(\boldsymbol{a},\boldsymbol{b}) \in U$. Let $\boldsymbol{F} : U \to \mathbb{R}^n$ with $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is analytic on U. If $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible, i.e. $det(D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})) \neq 0$, then there are:

(i) An open set V in $\mathbb{R}^{p+n}$ containing $(\boldsymbol{a},\boldsymbol{b})$, with $V \subseteq U$, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$ is invertible for all $(\boldsymbol{x},\boldsymbol{y}) \in V$;

(ii) An open set W in $\mathbb{R}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $\boldsymbol{f} : W \to \mathbb{R}^n$ such that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{b}$ and, for all $\boldsymbol{x} \in W$, $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, f(\boldsymbol{x})) = \boldsymbol{0} \ .$$

Additionally, $\boldsymbol{f}$ is analytic on W, and for any $\boldsymbol{x} \in W$, we have

$$D\boldsymbol{f}(\boldsymbol{x}) = -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, f(\boldsymbol{x}))$$

(in particular, $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ is invertible for all $\boldsymbol{x} \in W$).

VIII.5.10.2. Theorem. [Real Analytic Inverse Function Theorem] Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{a} \in U$, and $\boldsymbol{f} : U \to \mathbb{R}^n$ an analytic function. If $D\boldsymbol{f}(\boldsymbol{a})$ is invertible, then there is a neighborhood V of $\boldsymbol{a}$, $V \subseteq U$, such that

(i) $\boldsymbol{f}$ is one-to-one on V.

(ii) $\boldsymbol{f}(V)$ is open in $\mathbb{R}^n$.

(iii) $D\boldsymbol{f}(\boldsymbol{x})$ is invertible for all $\boldsymbol{x} \in V$.

(iv) $\boldsymbol{f}^{-1}$ is analytic on $\boldsymbol{f}(V)$ and, for all $\boldsymbol{y} \in \boldsymbol{f}(V)$,

$$D(\boldsymbol{f}^{-1})(\boldsymbol{y}) = [D\boldsymbol{f}(\boldsymbol{f}^{-1}(\boldsymbol{y}))]^{-1} \ .$$

VIII.5.10.3. Theorem. [General Real Analytic Implicit/Inverse Function Theorem] Let $n \in \mathbb{N}$, $p \in \mathbb{N} \cup \{0\}$, $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$, U an open neighborhood of $(\boldsymbol{a},\boldsymbol{b})$ in $\mathbb{R}^{p+n}$, $\boldsymbol{F} : U \to \mathbb{R}^n$ an analytic function with $\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{c}$ and $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ invertible. Then there are:

(i) An open neighborhood V of $(\boldsymbol{a},\boldsymbol{b})$ in $\mathbb{R}^{p+n}$, contained in U, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$ is invertible for all $(\boldsymbol{x},\boldsymbol{y}) \in V$;

(ii) An open neighborhood W of $(\boldsymbol{a},\boldsymbol{c})$ in $\mathbb{R}^{p+n}$;

(iii) A unique function $\boldsymbol{G} : W \to \mathbb{R}^n$ satisfying $\boldsymbol{G}(\boldsymbol{a}, \boldsymbol{c}) = \boldsymbol{b}$ and, for all $(\boldsymbol{x}, \boldsymbol{z}) \in W$, $(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) = \boldsymbol{z} \ .$$

Additionally, $\boldsymbol{G}$ is analytic on W, and for all $(\boldsymbol{x}, \boldsymbol{z}) \in V$,

$$D\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z}) = \begin{bmatrix} & \vdots & \\ -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) & \vdots & D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z}))^{-1} \\ & \vdots & \end{bmatrix} \ .$$

VIII.5.10.4. Theorem. [Implicit Function Theorem, Complex Analytic Version] Let U be an open set in $\mathbb{C}^{p+n}$ and $(\boldsymbol{a}, \boldsymbol{b}) \in U$. Let $\boldsymbol{F} : U \to \mathbb{C}^n$ with $\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{0}$. Suppose $\boldsymbol{F}$ is analytic on U. If $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b})$ is invertible, i.e. $det(D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b})) \neq 0$, then there are:

(i) An open set V in $\mathbb{C}^{p+n}$ containing $(\boldsymbol{a}, \boldsymbol{b})$, with $V \subseteq U$, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ is invertible for all $(\boldsymbol{x}, \boldsymbol{y}) \in V$;

(ii) An open set W in $\mathbb{C}^p$ containing $\boldsymbol{a}$;

(iii) A unique function $\boldsymbol{f} : W \to \mathbb{C}^n$ such that $\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{b}$ and, for all $\boldsymbol{x} \in W$, $(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, f(\boldsymbol{x})) = \boldsymbol{0} \ .$$

Additionally, $\boldsymbol{f}$ is analytic on W, and for any $\boldsymbol{x} \in W$, we have

$$D\boldsymbol{f}(\boldsymbol{x}) = -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, f(\boldsymbol{x}))$$

(in particular, $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}))$ is invertible for all $\boldsymbol{x} \in W$).

VIII.5.10.5. Theorem. [Complex Analytic Inverse Function Theorem] Let U be an open set in $\mathbb{C}^n$, $\boldsymbol{a} \in U$, and $\boldsymbol{f} : U \to \mathbb{C}^n$ an analytic function. If $D\boldsymbol{f}(\boldsymbol{a})$ is invertible, then there is a neighborhood V of $\boldsymbol{a}$, $V \subseteq U$, such that

(i) $\boldsymbol{f}$ is one-to-one on V.

(ii) $\boldsymbol{f}(V)$ is open in $\mathbb{C}^n$.

(iii) $D\boldsymbol{f}(\boldsymbol{x})$ is invertible for all $\boldsymbol{x} \in V$.

(iv) $\boldsymbol{f}^{-1}$ is analytic on $\boldsymbol{f}(V)$ and, for all $\boldsymbol{y} \in \boldsymbol{f}(V)$,

$$D(\boldsymbol{f}^{-1})(\boldsymbol{y}) = [D\boldsymbol{f}(\boldsymbol{f}^{-1}(\boldsymbol{y}))]^{-1} \ .$$

VIII.5.10.6. Theorem. [General Complex Analytic Implicit/Inverse Function Theorem] Let $n \in \mathbb{N}$, $p \in \mathbb{N} \cup \{0\}$, $\boldsymbol{a} \in \mathbb{C}^p$, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{C}^n$, U an open neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$ in $\mathbb{C}^{p+n}$, $\boldsymbol{F} : U \to \mathbb{C}^n$ an analytic function with $\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{c}$ and $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{a}, \boldsymbol{b})$ invertible. Then there are:

(i) An open neighborhood V of $(\boldsymbol{a}, \boldsymbol{b})$ in $\mathbb{C}^{p+n}$, contained in U, such that $D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})$ is invertible for all $(\boldsymbol{x}, \boldsymbol{y}) \in V$;

(ii) An open neighborhood W of $(\boldsymbol{a}, \boldsymbol{c})$ in $\mathbb{C}^{p+n}$;

(iii) A unique function $\boldsymbol{G} : W \to \mathbb{C}^n$ satisfying $\boldsymbol{G}(\boldsymbol{a}, \boldsymbol{c}) = \boldsymbol{b}$ and, for all $(\boldsymbol{x}, \boldsymbol{z}) \in W$, $(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) \in V$ and

$$\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) = \boldsymbol{z} \ .$$

Additionally, $\boldsymbol{G}$ is analytic on W, and for all $(\boldsymbol{x}, \boldsymbol{z}) \in V$,

$$D\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z}) = \begin{bmatrix} & \vdots & \\ -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z}))^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) & \vdots & D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z}))^{-1} \\ & \vdots & \end{bmatrix} .$$

VIII.5.10.7. If $n = p = 1$, the unique analytic functions defined in VIII.5.10.1., VIII.5.10.2., VIII.5.10.4., and VIII.5.10.5. have Taylor series which, when relocated to 0, are the formal power series obtained in III.5.3.8. and III.5.3.7. respectively with $F = \mathbb{R}$ or $\mathbb{C}$. In other words, if the initial power series in III.5.3.8. and III.5.3.7. have positive radius of convergence, so does the power series in the conclusion. The same is true for general n and p, cf. III.5.3.8.(h). (There are also formal analogs of VIII.5.10.3. and VIII.5.10.6.; details are left to the reader.)

VIII.5.10.8. The complex analytic versions can be efficiently proved as follows. We outline the proof of VIII.5.10.6., of which VIII.5.10.4. and VIII.5.10.5. are corollaries. So begin with an analytic function $\boldsymbol{F} : U \to \mathbb{C}^n$ satisfying the hypotheses.

(1) Regard $\boldsymbol{F}$ as a $\mathcal{C}^\infty$ function from $\mathbb{R}^{2p+2n}$ to $\mathbb{R}^{2n}$.

(2) Apply VIII.5.2.1. to obtain the unique $\boldsymbol{G} : \mathbb{R}^{2p+2n} \to \mathbb{R}^{2n}$ with the right properties. Then $\boldsymbol{G}$ is $\mathcal{C}^\infty$.

(3) Check from the derivative formula of VIII.5.2.1. (cf. III.11.12.2.) that the partial derivatives of $\boldsymbol{G}$ with respect to any consecutive odd/even pair of coordinates satisfy the Cauchy-Riemann equations (XII.10.1.7.).

(4) Conclude from XII.10.1.6. that $\boldsymbol{G}$, regarded as a function from $\mathbb{C}^{p+n}$ to $\mathbb{C}^n$, is complex analytic.

VIII.5.10.9. The Taylor series of $\boldsymbol{G}$ around $(\boldsymbol{a}, \boldsymbol{c})$ can be computed by brute force algebraic manipulation of power series (there is a unique multivariable power series which formally satisfies the identity $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{z})) = \boldsymbol{z}$ for all $\boldsymbol{x}, \boldsymbol{z}$; cf. VIII.5.11.10.), although the calculation becomes rather ghastly. This calculation can be used to give an alternate proof of most of VIII.5.10.6.. But the calculation is not sufficient by itself: it must also be shown that the resulting series converges in a neighborhood of $(\boldsymbol{a}, \boldsymbol{c})$. Then the function it converges to must satisfy the identity in the conclusion, and is the only analytic solution (). Thus there is a unique analytic function satisfying the conclusion. But to conclude there is a unique *function* satisfying the identity, additional argument is needed (such as an application of VIII.5.2.1.).

VIII.5.10.10. There is an alternate approach which avoids all use of VIII.5.2.1. (or the other smooth results). Begin by proving the Complex Analytic Inverse Function Theorem VIII.5.10.5., where uniqueness comes along essentially for free, using the method of VIII.5.10.9.. Then conclude VIII.5.10.6. as in VIII.5.2.3..

VIII.5.10.11. For the real analytic versions, the following procedure seems to suffice (we again outline the proof of VIII.5.10.3.). Let $\boldsymbol{F} : U \to \mathbb{R}^n$ satisfy the hypotheses.

(1) Complexify $\boldsymbol{F}$ to a complex analytic function $\boldsymbol{F}_{\mathbb{C}}$ from a neighborhood of $(\boldsymbol{a}, \boldsymbol{b})$ in $\mathbb{C}^{p+n}$ to $\mathbb{C}^n$, using the same power series centered at $(\boldsymbol{a}, \boldsymbol{b})$.

(2) Apply VIII.5.10.6. to $\boldsymbol{F}_{\mathbb{C}}$ to obtain a complex analytic function $\boldsymbol{G}_{\mathbb{C}} : \mathbb{C}^{p+n} \to \mathbb{C}^n$.

(3) Restrict $\boldsymbol{G}_{\mathbb{C}}$ to $\mathbb{R}^{p+n}$ to get $\boldsymbol{G}$.

There is, however, a technical subtlety: it must be shown that $\boldsymbol{G}$ takes values in $\mathbb{R}^n$, not just in $\mathbb{C}^n$. This can be done by computing the Taylor series for $\boldsymbol{G}_{\mathbb{C}}$ around $(\boldsymbol{a}, \boldsymbol{c})$ as in VIII.5.10.9. and verifying that all coefficients (partial derivatives) are real.

VIII.5.10.12. There is a better approach: as in VIII.5.10.10., first prove the Real Analytic Inverse Function Theorem VIII.5.10.2.. This can be done as follows. Suppose $\boldsymbol{f}$ satisfies the hypotheses of VIII.5.10.2..

(1) Apply the Inverse Function Theorem to show that $\boldsymbol{f}$ is one-to-one on a neighborhood V_r of $\boldsymbol{a}$ with $\mathcal{C}^\infty$ inverse function $\boldsymbol{f}^{-1}$ on the neighborhood $W_r = \boldsymbol{f}(V_r)$ of $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$ in $\mathbb{R}^n$.

(2) Complexify $\boldsymbol{f}$ to $\boldsymbol{f}_{\mathbb{C}}$ on a neighborhood of $\boldsymbol{a}$ in $\mathbb{C}^n$ with the same Taylor series. Verify that $\boldsymbol{f}_{\mathbb{C}}$ satisfies the hypotheses of the Complex Analytic Inverse Function Theorem, hence is one-to-one on a neighborhood V_c of $\boldsymbol{a}$ in $\mathbb{C}^n$ with complex analytic inverse function $\boldsymbol{f}_{\mathbb{C}}^{-1}$ on the neighborhood $W_c = \boldsymbol{f}_{\mathbb{C}}(V_c)$ of $\boldsymbol{b}$ in $\mathbb{C}^n$.

(3) Use that $\boldsymbol{f}_{\mathbb{C}}$ is one-to-one and extends $\boldsymbol{f}$ to conclude that the restriction of $\boldsymbol{f}_{\mathbb{C}}^{-1}$ to the neighborhood $W = W_c \cap W_r$ of $\boldsymbol{b}$ in $\mathbb{R}^n$ agrees with the restriction of $\boldsymbol{f}^{-1}$ to W. Thus $\boldsymbol{f}^{-1}$ is real analytic on W.

Theorem VIII.5.10.2. can also be proved by the method of VIII.5.10.9. without any use of the usual Inverse Function Theorem.

Thus the real and complex analytic theorems are all proved (see Exercise VIII.5.11.20. for details). 👍

VIII.5.10.13. There are then real and complex analytic versions of the Straightening Theorems and the Rank Theorem; see Exercise VIII.5.11.20..

VIII.5.11. Exercises

VIII.5.11.1. Show that the hypothesis that $D_y\boldsymbol{F}(\boldsymbol{a},\boldsymbol{b})$ is invertible is not necessary for the conclusion of the Smooth Implicit Function Theorem, i.e. that the converse of the theorem does not hold. [Consider $F(x,y) = x^3 - y^3$.]

VIII.5.11.2. (a) If F is a function from $\mathbb{R}^2$ to $\mathbb{R}$, and the graph of the equation $F(x,y) = 0$ is a curve that crosses itself at $\boldsymbol{a}$ with different tangent lines, and F is differentiable at $\boldsymbol{a}$, show that $DF(\boldsymbol{a}) = 0$. [Consider directional derivatives.] This "explains" why the Implicit Function Theorem cannot apply at $\boldsymbol{a}$.

(b) If F is a $\mathcal{C}^1$-function from $\mathbb{R}^2$ to $\mathbb{R}$, and the graph of the equation $F(x,y) = 0$ is a curve with two segments passing through $\boldsymbol{a}$ with the same tangent line, not identically equal in any neighborhood of $\boldsymbol{a}$, show that $DF(\boldsymbol{a}) = 0$. [By making a linear change of variables, it may be assumed the tangent line at $\boldsymbol{a}$ is horizontal. Show that the Implicit Function Theorem cannot apply at $\boldsymbol{a}$.]

VIII.5.11.3. Let W be an open set in $\mathbb{R}^p$, and $\boldsymbol{f}$ a continuous function from W to $\mathbb{R}^n$. Define $\boldsymbol{F} : W \times \mathbb{R}^n \to \mathbb{R}^n$ by $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x})$. Let $\boldsymbol{a} \in W$, and set $\boldsymbol{b} = \boldsymbol{f}(\boldsymbol{a})$.

(a) Show that $\boldsymbol{F}$, $\boldsymbol{a}$, $\boldsymbol{b}$ satisfy the hypotheses of VIII.5.1.6..

(b) Show that the implicitly defined function is $\boldsymbol{f}$ (possibly restricted to a smaller neighborhood of $\boldsymbol{a}$).

Thus every continuous function is defined implicitly by a function as in VIII.5.1.6.. Since $\boldsymbol{f}$ need not be differentiable anywhere, no stronger conclusion can be obtained in VIII.5.1.6. without additional hypotheses, even in the case $p = n = 1$.

VIII.5.11.4. Show that under the hypotheses of VIII.5.3.2., even if $p = 1$, there is not necessarily a continuous function f on a neighborhood W of $\boldsymbol{a}$ with $f(\boldsymbol{a}) = b$ and $F(\boldsymbol{x}, f(\boldsymbol{x})) = 0$ for all $\boldsymbol{x} \in W$. [Construct a pseudo-arc P in $[0,1] \times (0,1) \subseteq \mathbb{R}^2$ containing only one point $(0,c)$ on the left edge and one point on the right edge, whose complement in $[0,1]^2$ has exactly two components, and let F be a continuous function from $[0,1]^2$ to $\mathbb{R}$ with $F(0,y) = y - c$, $F(x,1) = 1 - c$, $F(x,0) = -c$, and which is zero precisely on P. Note that $\frac{\partial F}{\partial y}(0,c)$ exists and equals 1.]

VIII.5.11.5. (a) Assuming the Implicit Function Theorem, use the formula for $D\boldsymbol{F}$ in VIII.5.2.4. to deduce the formula for the total derivative of the inverse function to $\boldsymbol{f}$.

(b) Assuming the Inverse Function Theorem, use the formula for $D\tilde{\boldsymbol{F}}$ in VIII.5.2.3. and () to deduce the formula for the total derivative $D\boldsymbol{f} = -(D_{\boldsymbol{y}}\boldsymbol{F})^{-1} D_{\boldsymbol{x}}\boldsymbol{F}$ of the implicitly defined function $\boldsymbol{f}$.

VIII.5.11.6. Use VIII.5.3.9. and the argument of VIII.5.2.4. to obtain the version of the one-dimensional Inverse Function Theorem described in VIII.5.7.4..

VIII.5.11.7. (Newton; cf. [Edw79]) Let $F(x,y) = \sum_{j,k} a_{jk} x^j y^k$ be a polynomial. Suppose x and y are (differentiable) functions of t and write $\dot{x}$ and $\dot{y}$ for $x'(t)$ and $y'(t)$ respectively. Suppose x and y are related by the equation $F(x,y) = 0$, i.e. x and y give parametrized motion along the "curve" $F(x,y) = 0$.

(a) Using the power, product, and chain rules, deduce that

$$\sum_{j,k} \left(\frac{j\dot{x}}{x} + \frac{k\dot{y}}{y} \right) a_{jk} x^j y^k = 0 \ .$$

(b) Interpret this equation as

$$\dot{x}\frac{\partial F}{\partial x} + \dot{y}\frac{\partial F}{\partial y} = 0 \ .$$

(c) If $\dot{x}$ and $\frac{\partial F}{\partial y}$ are nonzero, deduce the formula

$$\frac{\dot{y}}{\dot{x}} = -\frac{\partial F/\partial x}{\partial F/\partial y} \ .$$

(d) Deduce the last two formulas if F is just a differentiable function of x and y.

(e) Is this a proof of the one-variable Implicit Function Theorem VIII.5.3.10.? Why or why not?

VIII.5.11.8. Recall the way the formula for $f'(x)$ is found by implicit differentiation in first-semester calculus from an equation $F(x, y) = 0$: differentiate $G(x) = F(x, f(x))$ by the Chain Rule and other differentiation rules, set it equal to 0, and solve for $f'(x)$ (streamlined notation is usually used which amounts to this). Show that the same formula as the one in VIII.5.3.10. is always obtained.

VIII.5.11.9. In VIII.5.3.7., assume that the first-order partials of F are defined on U and continuous at $(\boldsymbol{a}, b)$ (which implies that F is differentiable at $(\boldsymbol{a}, b)$), and that $\frac{\partial F}{\partial y}$ is nonzero on U (which can be automatically arranged in the continuous partial case if $\frac{\partial F}{\partial y}(\boldsymbol{a}, b) \neq 0$ by replacing U by a smaller neighborhood of $(\boldsymbol{a}, b)$). Give a simplified proof as follows (note that the f is unique and continuous in this case by VIII.5.3.4.):

(a) For notational simplicity, assume $p = 1$. If Δx is sufficiently small, write $\Delta y = f(a + \Delta x) - f(a) = f(a + \Delta x) - b$. Then

$$0 = F(a + \Delta x, b + \Delta y) = F(a, b) = F(a + \Delta x, b + \Delta y) - F(a, b)$$
$$= [F(a + \Delta x, b + \Delta y) - F(a, b + \Delta y)] + [F(a, b + \Delta y) - F(a, b)] \ .$$

(b) By two applications of the MVT, there are $0 < \alpha, \beta < 1$ (depending on Δx) such that

$$F(a + \Delta x, b + \Delta y) - F(a, b + \Delta y) = \frac{\partial F}{\partial x}(a + \alpha\Delta x, b + \Delta y)\Delta x$$

$$F(a, b + \Delta y) - F(a, b) = \frac{\partial F}{\partial y}(a, b + \beta\Delta y)\Delta y \ .$$

(c) From (a) and (b),

$$\frac{\partial F}{\partial x}(a + \alpha\Delta x, b + \Delta y)\Delta x + \frac{\partial F}{\partial y}(a, b + \beta\Delta y)\Delta y = 0$$

so

$$\frac{\Delta y}{\Delta x} = -\frac{\frac{\partial F}{\partial x}(a + \alpha\Delta x, b + \Delta y)}{\frac{\partial F}{\partial y}(a, b + \beta\Delta y)} \ .$$

(d) By continuity,

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} -\frac{\frac{\partial F}{\partial x}(a + \alpha\Delta x, b + \Delta y)}{\frac{\partial F}{\partial y}(a, b + \beta\Delta y)}$$

exists and equals $-\frac{\frac{\partial F}{\partial x}(a,b)}{\frac{\partial F}{\partial y}(a,b)}$.

(e) Adapt the argument notationally to the case $p > 1$.

VIII.5.11.10. Use the Chain Rule to find formulas for the higher-order derivatives of implicitly defined functions.

VIII.5.11.11. Give a direct argument proving the Continuous Implicit Function Theorem along the lines of the proof of VIII.5.5.1. by proving the following version first proved in [Gou03]:

THEOREM. Let U be an open set in $\mathbb{R}^{p+n}$ containing $(\mathbf{0},\mathbf{0})$, and $\boldsymbol{G} : U \to \mathbb{R}^n$ a continuous function with $\boldsymbol{G}(\mathbf{0},\mathbf{0}) = \mathbf{0}$, such that $D_{\boldsymbol{y}}\boldsymbol{G}$ is continuous on U and $D_{\boldsymbol{y}}\boldsymbol{G}(\mathbf{0},\mathbf{0}) = \mathbf{0}$. Then there are open sets V in $\mathbb{R}^{p+n}$, $V \subseteq U$, $(\mathbf{0},\mathbf{0}) \in V$, and W in $\mathbb{R}^p$, $\mathbf{0} \in W$, and Y in $\mathbb{R}^n$, $\mathbf{0} \in Y$, such that for every $\boldsymbol{y} \in Y$ there is a unique $\boldsymbol{x} \in W$ with $(\boldsymbol{x},\boldsymbol{y}) \in V$ and $\boldsymbol{G}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y}$.

(a) Prove this theorem by using the Contraction Mapping Principle.

(b) If $U \subseteq \mathbb{R}^{p+n}$, $(\mathbf{0},\mathbf{0}) \in U$, and $\boldsymbol{F} : U \to \mathbb{R}^n$ is a continuous with $\boldsymbol{F}(\mathbf{0},\mathbf{0}) = \mathbf{0}$, such that $D_{\boldsymbol{y}}\boldsymbol{F}$ is continuous on U and $D_{\boldsymbol{y}}\boldsymbol{F}(\mathbf{0},\mathbf{0}) = I_n$, show that there is a unique implicitly defined function $\boldsymbol{f}$ on a neighborhood W of $\mathbf{0}$ in $\mathbb{R}^p$, taking values in a neighborhood Y of $\mathbf{0}$ in $\mathbb{R}^n$, such that $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{f}(\boldsymbol{x})) = \mathbf{0}$ for all $\boldsymbol{x} \in W$. [Set $\boldsymbol{G}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{F}(\boldsymbol{x},\boldsymbol{y})$.]

(c) Use an affine change of variables to deduce the existence of an implicit function under the hypotheses of VIII.5.1.6.. Conclude continuity of the implicit function from VIII.5.1.15..

See [**?**] for a nice geometric interpretation of this argument.

VIII.5.11.12. Define $F : \mathbb{R}^2 \to \mathbb{R}$ by

$$F(x,y) = \begin{cases} y + x^2 \cos\left(\frac{y}{x^4}\right) & \text{if} \quad x \neq 0 \\ y & \text{if} \quad x = 0 \end{cases} .$$

(a) Show that F is differentiable everywhere. In particular, F is continuous, and $\frac{\partial F}{\partial x}$ and $\frac{\partial F}{\partial y}$ exist everywhere.

(b) Show that $\frac{\partial F}{\partial y}$ is continuous in y for fixed x, and that $\frac{\partial F}{\partial y}(0,0) = 1$.

(c) Show that the equation $F(x,y) = 0$ cannot be uniquely solved for y as a differentiable function of x in any small neighborhood of $(0,0)$ (the smaller the neighborhood, the more differentiable solutions there are).

VIII.5.11.13. (a) Write a statement which is a common generalization of VIII.5.3.12. and VIII.5.7.7., along the lines of VIII.5.4.2..

***(b) Is this statement true? [cf. VIII.5.11.14.(b).]

VIII.5.11.14. (a) Show that under the hypotheses of VIII.5.7.5., if U is connected, then $det(D\boldsymbol{f}(\boldsymbol{x}))$ does not change sign on U, i.e. $\boldsymbol{f}$ either preserves or reverses orientation on all of U.

***(b) Is the assumption of a decreasing chain of minors as in VIII.5.4.3. automatic? (If so, an inductive proof of VIII.5.7.5. and the statement in VIII.5.11.13. can be given.)

VIII.5.11.15. Let U and V be open sets in $\mathbb{R}^n$, and $f : U \to V$ a bijection which is $\mathcal{C}^r$ $(1 \leq r \leq \infty)$. If f^{-1} is differentiable everywhere on V, show that f^{-1} is necessarily $\mathcal{C}^r$ on V.

VIII.5.11.16. Formulate and prove generalizations of the Continuous and Mixed Implicit Function Theorem similar to VIII.5.2.1..

VIII.5.11.17. Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$. Around which points can f be straightened?

VIII.5.11.18. Let $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^2$ be given by $\boldsymbol{f}(x) = (x^3, x^2)$. Then $\boldsymbol{f}$ is $\mathcal{C}^\infty$. Show that the range of $\boldsymbol{f}$ can be topologically straightened but not smoothly straightened around 0.

VIII.5.11.19. Define $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^2$ by

$$\boldsymbol{f}(x) = \left(e^{-1/x^2} \left[1 - \cos \frac{2\pi}{x} \right], e^{-1/x^2} \sin \frac{2\pi}{x} \right)$$

for $x \neq 0$, and $\boldsymbol{f}(0) = (0, 0)$.

(a) Show that $\boldsymbol{f}$ is a $\mathcal{C}^\infty$ function.

(b) Show that the image of $\left(-\frac{1}{n}, \frac{1}{n}\right)$ is topologically a Hawaiian earring (XIII.19.5.8.) for any $n \in \mathbb{N}$. Thus the range cannot even be topologically straightened around 0.

(c) Discuss straightening of the function $\boldsymbol{g} : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\boldsymbol{g}(x, y) = \left(e^{-1/x^2} \left[1 - \cos \frac{2\pi}{x} \right], y + e^{-1/x^2} \sin \frac{2\pi}{x} \right)$$

for $x \neq 0$ and $\boldsymbol{g}(0, y) = (0, y)$.

VIII.5.11.20. (a) Write out the details of the proofs of the real and complex analytic versions of the Implicit and Inverse Function Theorems.

(b) Compute enough terms of the Taylor series of the implicitly defined function to indicate the general procedure and the proof of uniqueness.

(c) Verify by calculation the derivative formulas for the complex analytic versions (for the real analytic versions, these formulas are trivial from the smooth versions).

(d) Formulate and prove real and complex analytic versions of the Straightening Theorems and the Rank Theorem.

VIII.5.11.21. Formulate and prove precise versions, both real and complex, of the statement that the roots of an n'th degree polynomial are continuous, and in fact analytic, functions of the coefficients, except at multiple roots. [Fix n, and consider the function

$$F(a_0, \ldots, a_n, x) = a_0 + a_1 x + \cdots + a_n x^n$$

and solve $F(a_0, \ldots, a_n, x) = 0$ for x.]

This suggests there are analytic formulas for the roots of a general n'th degree polynomial, although explicit expressions for the formulas may not be possible in general. There cannot be explicit formulas just involving finitely many algebraic operations, including roots, if $n \geq 5$.

Note that this argument applies to polynomials of degree $\leq n$, for any fixed n. Thus the analytic functions obtained are compatible for different n. Illustrate this explicitly for $n = 2$.

VIII.5.11.22. Let U be an open set in $\mathbb{R}^3$, and $F : U \to \mathbb{R}$ be a $\mathcal{C}^1$ function with $\frac{\partial F}{\partial x}$, $\frac{\partial F}{\partial y}$, and $\frac{\partial F}{\partial z}$ nonzero on U. Then the equation $F(x, y, z) = 0$ can be locally solved for each of x, y, z as functions of the other two. Show that

$$\frac{\partial x}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial x} = -1 \ .$$

(Give a careful explanation of what this equation means.) Thus algebraic manipulations of Leibniz notation must be done carefully. See also VIII.2.4.2..

VIII.5.11.23. Let K be the Cantor set, and let ϕ be a nonnegative $\mathcal{C}^\infty$ function from $\mathbb{R}$ to $\mathbb{R}$ which is zero exactly on $K \cup (-K)$ (V.17.2.9.). Set

$$F(x, y, z) = x^2 + y^2 - \phi(z)$$

$$G(x, y, z) = x^2 + y^2 + \phi(z)$$

and let $\Gamma = \{(x, y, z) : F(x, y, z) = 0\}$ and $\Delta = \{(x, y, z) : G(x, y, z) = 0\}$.

(a) Show that F and G are $\mathcal{C}^\infty$ on all of $\mathbb{R}^3$.

(b) Discuss the local structure of Γ and Δ around $\mathbf{0}$. Are they surfaces in any sense?

VIII.5.11.24. Formulate and prove the general version of VIII.5.1.25. for arbitrary $k, n, p \in \mathbb{N}$, $k \leq n$.

VIII.5.11.25. Deduce the following result from VIII.5.8.2.:

Corollary. [Smooth Open Mapping Theorem] Let $\boldsymbol{f}$ be a $\mathcal{C}^r$ function ($r \geq 1$) from an open set U in $\mathbb{R}^n$ to $\mathbb{R}^m$. Suppose $D\boldsymbol{f}$ has rank m at all points of U (this entails that $n \geq m$). Then $\boldsymbol{f}$ is an open mapping, i.e. if V is an open set in $\mathbb{R}^n$ contained in U, then $\boldsymbol{f}(V)$ is open in $\mathbb{R}^m$.

Compare with ().

VIII.5.11.26. Show that if $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear (and hence differentiable), the Rank Theorem for f follows easily from (and is in fact equivalent to) the following facts from linear algebra:

(i) Every subspace of a vector space has a complementary subspace ().

(ii) If $\mathcal{W}$ is a subspace of $\mathbb{R}^n$ complementary to the null space $\mathcal{N}(f)$, then the restriction of f to $\mathcal{W}$ is a linear isomorphism of $\mathcal{W}$ onto the range of f ().

(iii) Two vector spaces of the same dimension are isomorphic.

VIII.6. Constrained Extrema and Lagrange Multipliers

It is often necessary to find the extrema of a function over a surface in $\mathbb{R}^3$, or a higher-dimensional version of the situation. There is a clever way to do this in many cases, using the method of *Lagrange multipliers*. The theory is related to the Implicit Function Theorem.

VIII.6.1. The Two-Dimensional Case

VIII.6.2. The Higher-Dimensional Case with One Constraint

VIII.6.3. The Higher-Dimensional Case with Multiple Constraints

VIII.6.4. Degenerate Cases

VIII.6.5. Exercises

VIII.6.5.1. Consider the following problem treated in first-semester calculus courses:

> Find the most economical shape of a cylindrical can, i.e. the shape making the surface area as small as possible for a fixed volume.

In other words, if r is the radius of the base and h the height, minimize

$$S = 2\pi r^2 + 2\pi r h$$

subject to the constraint

$$V = \pi r^2 h = \text{ constant }.$$

(a) The standard way to do it is to solve the constraint equation for h and then find the minimum of

$$S(r) = 2\pi r^2 + 2\pi r \left(\frac{V}{\pi r^2}\right) = 2\pi r^2 + \frac{2V}{r}$$

for $r > 0$. Solve the problem this way and show that h must be $2r$.

(b) Solve the problem somewhat more easily using Lagrange multipliers (this cannot be done in first-semester calculus!)

(c) Critique the following alternate solution which appeared in [Gri27]:

"[Implicitly] Differentiate both of these equations at once with respect to r [regarding h as a function of r].

$$\frac{dS}{dr} = 4\pi r + 2\pi h + 2\pi r \frac{dh}{dr} = 0,$$

$$\frac{dV}{dr} = 2\pi r h + \pi r^2 \frac{dh}{dr} = 0.$$

[$\frac{dV}{dr} = 0$ since V is constant, and $\frac{dS}{dr} = 0$ at a minimum.] From [the second] we find

$$\frac{dh}{dr} = -\frac{2h}{r}.$$

This substituted in [the first] gives

$$4\pi r + 2\pi h - 4\pi h = 0;$$

whence $h = 2r$. The height must equal the diameter."

This is the correct answer. When does this method work?

VIII.7. Surface Area and Surface Integrals

VIII.7.1. Surface Area

Giving a proper definition of area for a surface in $\mathbb{R}^3$ is much trickier than defining arc length of a curve. The roughly analogous process simply does not work.

VIII.7.1.1. To define the arc length of a curve, one chooses a large number of closely spaced points along the curve, approximates the curve by the straight line segments between consecutive points, and adds up the lengths of these segments to give an approximation to the arc length. As the lengths of these approximating segments approaches zero uniformly, the approximations to the arc length approach a limit (actually the supremum) which is the reasonable candidate for the arc length of the curve. See ().

We could try to do something similar with a surface. Choose a lot of closely spaced points which can be connected by straight line segments forming planar triangles, and use these triangles as a piecewise-linear approximation to the surface and the sum of the areas as an approximation to the area of the surface, and then take the limit as the diameters of the triangles approach zero.

VIII.7.1.2. There are two potential problems:

(i) It is not obvious that such triangular approximations can be made for any surface. This turns out not to be a serious problem, since it is a (nontrivial) theorem that every surface can be triangulated ().

(ii) It is not obvious that the limit exists in this case.

In fact, the limit does *not* exist in general! So this method of defining surface area is doomed to failure.

VIII.7.1.3. Example. We describe the standard example where things go terribly wrong. This phenomenon was discovered by Archimedes in ancient Greece.

We want to find the surface area of a (right) cylinder of height h whose base has radius r. By cutting the cylinder along a vertical line and "unrolling" it, we can easily guess that the surface area should be $2\pi rh$.

Instead, we try the procedure described above. To make calculations easier, take the case $r = h = 1$. Fix two large natural numbers m and n. Draw $m+1$ evenly spaced horizontal circles on the cylinder including the two end circles, and on each place n evenly spaced points so alternate circles have the points placed halfway between the ones on the adjacent circles. Approximate by triangles whose vertices are two consecutive points on one circle and the point halfway between on an adjacent circle. This approximates the cylinder by an "accordion" surface (Figure VIII.20), which is sometimes called a *Schwarz lantern* since it was described by H. Schwarz.

Let the sum of the areas of the triangles be $A_{m,n}$. Do we have that

$$\lim_{m,n\to\infty} A_{m,n} = 2\pi \ ?$$

(This means: for any $\epsilon > 0$, is there an N such that $|A_{m,n} - 2\pi| < \epsilon$ whenever $m, n > N$.)

It is easy to see that the answer is no. Fix m and n. Each triangle is isosceles, with a base of length $2\sin\frac{\pi}{n}$ and altitude greater than $1 - \cos\frac{\pi}{n}$. Since there are $2mn$ triangles, we have that

$$A_{m,n} > 2mn \sin\frac{\pi}{n}\left(1 - \cos\frac{\pi}{n}\right) \ .$$

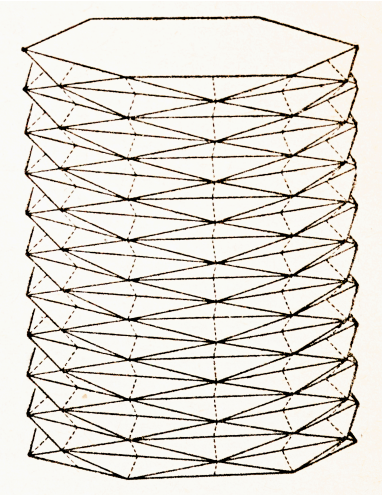


Figure VIII.20: Schwarz lantern.

If n is large, we have $\sin\frac{\pi}{n} \approx \frac{\pi}{n}$ and $1-\cos\frac{\pi}{n} \approx \frac{\pi^2}{2n^2}$, so the area of each triangle is roughly at least $\frac{\pi^3}{2n^3}$. Thus, if $m \gg n^2$, $A_{m,n}$ is large, so cannot approach 2π.

In fact, with slightly more care the altitude of each triangle can be computed to be $\sqrt{(1-\cos\frac{\pi}{n})^2+\frac{1}{m^2}}$, so we have an exact formula

$$A_{m,n} = 2mn\sin\frac{\pi}{n}\sqrt{\left(1-\cos\frac{\pi}{n}\right)^2+\frac{1}{m^2}}\ .$$

Suppose m and n go to infinity in such a way that

$$\lim_{m,n\to\infty}\frac{m}{n^2} = q$$

(this can be arranged for any $q \in [0,+\infty]$). Then

$$\lim_{m,n\to\infty} A_{m,n} = \lim_{m,n\to\infty} 2mn\sin\frac{\pi}{n}\sqrt{\left(1-\cos\frac{\pi}{n}\right)^2+\frac{1}{m^2}}$$

$$= \lim_{m,n\to\infty} 2\pi\left[\frac{n}{\pi}\sin\frac{\pi}{n}\right]\sqrt{m^2\frac{\pi^4}{4n^4}\left[\frac{4n^4}{\pi^4}\left(1-\cos\frac{\pi}{n}\right)^2\right]+1} = 2\pi\sqrt{\frac{\pi^4}{4}q^2+1}$$

since the limit of the expressions in square brackets is 1. Thus the area $A_{m,n}$ approaches the right answer 2π if and only if $q=0$, i.e. if and only if $m \to \infty$ more slowly than n^2. The limit can be anything in $[2\pi,+\infty]$ by choosing q properly.

VIII.7.1.4. It is actually possible to compute the surface area of any surface, or surfacelike subset, of $\mathbb{R}^3$ as a limit, although in a more complicated way. Indeed, the area of such a set is just its 2-dimensional Hausdorff measure (XVII.6.3.). But for smooth or piecewise-smooth surfaces, we will proceed in a different way which not only gives a workable definition, but also a practical method of calculation in reasonable cases.

VIII.8. Power Series and Analytic Functions of Several Variables

The theory of power series and analytic functions in several variables is similar to the theory in one variable with some technical complications. In this section, we will take the path of least resistance to the most important aspects of the theory, avoiding some of the (sometimes interesting) technicalities that arise. As in the one-variable case, we can do the real and complex theories in parallel; we let $\mathbb{F}$ stand for either $\mathbb{R}$ or $\mathbb{C}$.

VIII.8.1. Multivariable Power Series

VIII.8.1.1. In general, we want to consider power series of the form

$$\sum_{k_1=0}^{\infty}\sum_{k_2=0}^{\infty}\cdots\sum_{k_n=0}^{\infty} a_{k_1,\ldots,k_n}x_1^{k_1}x_2^{k_2}\cdots x_n^{k_n}$$

where the $a_{k_1,\ldots,k_n}$ are scalars (real or complex) and $x_1,\ldots,x_n$ are independent variables. The notation can become oppressive working with a general n, although multiindex notation (VIII.8.1.5.) can help, so sometimes we will notationally stick to the case $n=2$, writing x and y (or z and w in the complex case) for the independent variables. Just as in multivariable calculus, the case $n=2$ exhibits all the essential features of the theory for general n in contrast with the case $n=1$. Thus a power series in two variables will be written

$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a_{jk}x^j y^k$$

or more generally

$$\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a_{jk}(x-x_0)^j(y-y_0)^k\ .$$

VIII.8.1.2. Even in the two-variable case, there is no natural ordering on the terms, so we will regard the infinite series as an unordered sum

$$\sum_{(j,k)} a_{jk}(x-x_0)^j(y-y_0)^k$$

where the sum is over $\mathbb{N}_0^2$ (recall the notation $\mathbb{N}_0=\mathbb{N}\cup\{0\}$). In general, we will have an unordered sum over $\mathbb{N}_0^n$. Regarding multivariable power series as unordered sums means that we will only consider absolute convergence, and leads to technical simplifications in the theory without essential loss of generality.

VIII.8.1.3. Since we are considering multivariable power series as unordered sums, there is no harm in allowing more than one term with the same powers $(x-x_0)^j(y-y_0)^k$; we may even have infinitely many terms with the same pair (j,k) of powers, and it is convenient to allow this. Any such sum can be converted to *standard form* by first summing together all terms with powers a fixed (j,k), and then summing over (j,k); the standard form is a drastic revision of the original power series. As long as the original power series converges (as an unordered sum, i.e. converges absolutely), the standard form conversion and series will also converge (IV.2.10.20.). So we can make a formal definition:

VIII.8.1.4. DEFINITION. Let Ω be an index set, and $\phi = (\phi_1, \ldots, \phi_n) : \Omega \to \mathbb{N}_0^n$ a function. A *multivariable power series* in variables $X_1, \ldots, X_n$ indexed by Ω is an unordered sum

$$\sum_{\omega \in \Omega} a_\omega X_1^{\phi_1(\omega)} X_2^{\phi_2(\omega)} \cdots X_n^{\phi_n(\omega)}$$

where the a_ω are scalars. If $(c_1, \ldots, c_n)$ is an n-tuple of scalars, a multivariable power series in $x_1, \ldots, x_n$ centered at $(c_1, \ldots, c_n)$, indexed by Ω, is an unordered sum

$$\sum_{\omega \in \Omega} a_\omega (x_1 - c_1)^{\phi_1(\omega)} (x_2 - c_2)^{\phi_2(\omega)} \cdots (x_n - c_n)^{\phi_n(\omega)} \ .$$

If $n = 2$, a multivariable power series in x and y centered at (x_0, y_0), indexed by Ω, is an unordered sum

$$\sum_{\omega \in \Omega} a_\omega (x - x_0)^{\phi_1(\omega)} (y - y_0)^{\phi_2(\omega)} \ .$$

The multivariable power series is in *standard form* if ϕ is one-to-one; in this case, we may take $\Omega = \mathbb{N}_0^n$ and write the power series as

$$\sum_{(k_1, \ldots, k_n) \in \mathbb{N}_0^n} a_{k_1, \ldots, k_n} (x_1 - c_1)^{k_1} \cdots (x_n - c_n)^{k_n} \ .$$

Any convergent multivariable power series has a unique drastic revision which is in standard form.

VIII.8.1.5. It is convenient to use *multi-index notation*: if $\alpha = (m_1, \ldots, m_n) \in \mathbb{N}_0^n$, and $\boldsymbol{x} = (x_1, \ldots, x_n)$, write $\boldsymbol{x}^\alpha = x_1^{m_1} x_2^{m_2} \cdots x_n^{m_n}$. We can then write the series

$$\sum_{\omega \in \Omega} a_\omega (x_1 - c_1)^{\phi_1(\omega)} (x_2 - c_2)^{\phi_2(\omega)} \cdots (x_n - c_n)^{\phi_n(\omega)}$$

centered at $\boldsymbol{c} = (c_1, \ldots, c_n)$ as

$$\sum_{\omega \in \Omega} a_\omega (\boldsymbol{x} - \boldsymbol{c})^{\phi(\omega)}$$

and a standard form multivariable power series as

$$\sum_{\alpha \in \mathbb{N}_0^n} a_\alpha (\boldsymbol{x} - \boldsymbol{c})^\alpha \ .$$

VIII.8.1.6. If $\boldsymbol{c} = (c_1, \ldots, c_n) \in \mathbb{F}^n$, the power series

$$\sum_{\omega \in \Omega} a_\omega (\boldsymbol{x} - \boldsymbol{c})^{\phi(\omega)}$$

converges at $\boldsymbol{b} = (b_1, \ldots, b_n) \in \mathbb{F}^n$ if the unordered sum

$$\sum_{\omega \in \Omega} a_\omega (\boldsymbol{b} - \boldsymbol{c})^{\phi(\omega)}$$

converges (i.e. converges absolutely). We then have by comparison:

VIII.8.1.7. PROPOSITION. Let

$$\sum_{\omega\in\Omega} a_\omega(\boldsymbol{x}-\boldsymbol{c})^{\phi(\omega)}$$

be a multivariable power series. If the series converges at $\boldsymbol{b} = (b_1,\ldots,b_n)$, then it converges at $\boldsymbol{x} = (x_1,\ldots,x_n)$ for any $(x_1,\ldots,x_n) \in \mathbb{F}^n$ with $|x_1-c_1| \le |b_1-c_1|,\ldots,|x_n-c_n| \le |b_n-c_n|$.

VIII.8.1.8. Thus, if the series converges for some $\boldsymbol{b}$ with $b_k \ne c_k$ for all k, then it converges at all points of the rectangle (or *polydisk* if $\mathbb{F} = \mathbb{C}$)

$$R = \{\boldsymbol{x} : |x_k - c_k| \le |b_k - c_k| \text{ for all } 1 \le k \le n\}$$

centered at $\boldsymbol{c}$, and thus defines a function from R to $\mathbb{F}$.

VIII.8.1.9. EXAMPLE. A one-variable power series converges precisely on an interval (or disk in the complex case). But the situation is more complicated with multivariable power series: the set of convergence need not be a rectangle (or polydisk). For example, consider the real 2-variable power series

$$\sum_{k=0}^{\infty} x^k y^k \ .$$

This series converges (to $\frac{1}{1-xy}$) if and only if $|xy| < 1$, i.e. on the open subset of $\mathbb{R}^2$ containing $(0,0)$ bounded by the hyperbolas $y = \pm\frac{1}{x}$.

The best simple general statement we can make is that for any fixed y, a 2-variable power series converges at (x,y) exactly for all x in a (possibly degenerate or empty) interval (disk) centered at x_0, and similarly for fixed x. Thus if S is the set of convergence, every horizontal cross section is an interval (disk) centered at x_0, and every vertical cross section is an interval (disk) centered at y_0. These statements are true since if one variable is fixed, the series is an ordinary power series in the other variable. A similar statement holds in higher dimensions (fixing all but one variable). Thus, in general, the set of convergence is a union of rectangles (polydisks) centered at $\boldsymbol{c}$.

VIII.8.1.10. PROPOSITION. Let

$$\sum_{\omega\in\Omega} a_\omega(\boldsymbol{x}-\boldsymbol{c})^{\phi(\omega)}$$

be a multivariable power series which converges on a rectangle (polydisk)

$$R = \{\boldsymbol{x} : |x_k - c_k| \le r_k \text{ for all } 1 \le k \le n\}$$

centered at $\boldsymbol{c}$, where $r_1,\ldots,r_n > 0$. Let f be the function on R defined by the power series. For each finite $E \subseteq \Omega$, define

$$f_E(\boldsymbol{x}) = \sum_{\omega\in E} a_\omega(\boldsymbol{x}-\boldsymbol{c})^{\phi(\omega)}$$

for $\boldsymbol{x} \in R$. Then $f_E \to f$ uniformly on R, i.e. for every $\epsilon > 0$ there is a finite subset F of Ω such that, whenever E is a finite subset of Ω containing F,

$$|f(\boldsymbol{x}) - f_E(\boldsymbol{x})| < \epsilon$$

for all $\boldsymbol{x} \in R$.

PROOF: This is essentially the Weierstrass M-test (). Set $\boldsymbol{r} = (r_1, \ldots, r_n)$. The unordered sum

$$\sum_{\omega \in \Omega} a_\omega \boldsymbol{r}^{\phi(\omega)}$$

converges; hence for $\epsilon > 0$ there is a finite subset $F \subseteq \Omega$ such that

$$\sum_{\omega \in \Omega \setminus F} |a_\omega| \boldsymbol{r}^{\phi(\omega)} < \epsilon \ .$$

Thus, if E is a finite subset of Ω containing F, we have

$$|f(\boldsymbol{x}) - f_E(\boldsymbol{x})| = \left| \sum_{\omega \in \Omega \setminus E} a_\omega (\boldsymbol{x} - \boldsymbol{c})^{\phi(\omega)} \right|$$

$$\leq \sum_{\omega \in \Omega \setminus E} |a_\omega| |x_1 - c_1|^{\phi_1(\omega)} |x_2 - c_2|^{\phi_2(\omega)} \cdots |x_n - c_n|^{\phi_n(\omega)} \leq \sum_{\omega \in \Omega \setminus F} |a_\omega| \boldsymbol{r}^{\phi(\omega)} < \epsilon \ .$$



VIII.8.1.11. COROLLARY. If a multivariable power series converges on a rectangle (polydisk) R, it defines a continuous function on R.

We now show that a multivariable power series can be differentiated term-by-term.

VIII.8.1.12. THEOREM. Let

$$\sum_{\omega \in \Omega} a_\omega (\boldsymbol{x} - \boldsymbol{c})^{\phi(\omega)}$$

be a multivariable power series which converges on a rectangle (polydisk)

$$R = \{\boldsymbol{x} : |x_k - c_k| \leq r_k \text{ for all } 1 \leq k \leq n\}$$

centered at $\boldsymbol{c}$, where $r_1, \ldots, r_n > 0$. Let f be the function on R defined by the power series. Then, for each k, $1 \leq k \leq n$, the unordered sum

$$\sum_{\omega \in \Omega} a_\omega \frac{\partial}{\partial x_k} \left((\boldsymbol{x} - \boldsymbol{c})^{\phi(\omega)} \right) = \sum_{\omega \in \Omega} \phi_k(\omega) a_\omega (x_1 - c_1)^{\phi_1(\omega)} (x_2 - c_2)^{\phi_2(\omega)} \cdots (x_k - c_k)^{\phi_k(\omega) - 1} \cdots (x_n - c_n)^{\phi_n(\omega)} \tag{VIII.2}$$

converges u.c. to $\frac{\partial f}{\partial x_k}$ on the open rectangle (polydisk)

$$R^\circ = \{\boldsymbol{x} : |x_k - c_k| < r_k \text{ for all } 1 \le k \le n\} .$$

PROOF: This can be quickly reduced to the one-variable case. Fix k. For each $j \neq k$, fix x_j with $|x_j - c_j| \le r_j$. For any x_k, $|x_k - c_k| \le r_k$, the unordered sum

$$\sum_{\omega \in \Omega} a_\omega (\boldsymbol{x} - \boldsymbol{c})^{\phi(\omega)}$$

converges; hence it can be computed by summing in any manner. Form the drastic revision

$$\sum_{m \in \mathbb{N}_0} b_m (x_k - c_k)^m$$

obtained by first summing all terms with factor $(x_k - c_k)^m$ for fixed m. This is an unordered power series in one variable $x_k - c_k$ which converges for $|x_k - c_k| \le r_k$; thus the associated ordinary power series

$$\sum_{m=0}^{\infty} b_m (x - c_k)^m \tag{VIII.3}$$

has radius of convergence at least r_k, and converges to $f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n)$ for $|x - c_k| \le r_k$. Thus, by V.15.3.23., the series of term-by-term derivatives converges absolutely to

$$\frac{d}{dx}\left(f(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n)\right) = \frac{\partial f}{\partial x_k}(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n)$$

for $|x - c_k| < r_k$, i.e. VIII.3 converges to $\frac{\partial f}{\partial x_k}$ on R°.

We have not yet shown that the original multivariable series VIII.2 converges. But this can be shown by a majorization argument. The above argument can be used starting with the series of absolute values

$$\sum_{\omega \in \Omega} |a_\omega| (\boldsymbol{x} - \boldsymbol{c})^{\phi(\omega)}$$

which converges on R to a function F. If we choose the x_j so that $x_j - c_j \ge 0$ for all j, then the unordered sum

$$\sum_{\omega \in \Omega} \phi_k(\omega) |a_\omega| (x_1 - c_1)^{\phi_1(\omega)} (x_2 - c_2)^{\phi_2(\omega)} \cdots (x_k - c_k)^{\phi_k(\omega) - 1} \cdots (x_n - c_n)^{\phi_n(\omega)} \tag{VIII.4}$$

has a drastic revision which converges (to $\frac{\partial F}{\partial x_k}$); thus the series VIII.4 also converges since all the terms are nonnegative. This implies that VIII.2 converges on R° by comparison. Since it has a drastic revision which converges to $\frac{\partial f}{\partial x_k}$ on R°, VIII.2 also converges to $\frac{\partial f}{\partial x_k}$ on R°.

This argument shows pointwise convergence of VIII.2 on R°; u.c. convergence is then automatic by VIII.8.1.10.. 👍

VIII.8.1.13. Corollary. Let

$$\sum_{\omega\in\Omega} a_\omega(\boldsymbol{x}-\boldsymbol{c})^{\phi(\omega)}$$

be a multivariable power series which converges on an open rectangle (polydisk)

$$R^\circ = \{\boldsymbol{x} : |x_k - c_k| < r_k \text{ for all } 1 \le k \le n\}$$

centered at $\boldsymbol{c}$, where $r_1, \dots, r_n > 0$. Let f be the function on R° defined by the power series. Then f is $\mathcal{C}^\infty$ on R°, and the partial derivatives of all orders of f are obtained by term-by-term partial differentiation, i.e. for any $m_1, \dots, m_n \ge 0$ the unordered sum

$$\sum_{\omega\in\Omega} a_\omega \frac{\partial^{m_1+\cdots+m_n}}{\partial x_1^{m_1}\cdots\partial x_n^{m_n}}\left((\boldsymbol{x}-\boldsymbol{c})^{\phi(\omega)}\right)$$

converges u.c. to $\dfrac{\partial^{m_1+\cdots+m_n} f}{\partial x_1^{m_1}\cdots\partial x_n^{m_n}}$ on R°.

Proof: Every point of R° is contained in the interior of a closed polydisc centered at $(c_1, \dots, c_n)$ contained in R°. Apply VIII.8.1.12. repeatedly, shrinking the polydisk slightly each time. 👍

VIII.8.1.14. We extend the multi-index notation: if $\alpha = (m_1, \dots, m_n) \in \mathbb{N}_0^n$, write

$$|\alpha| = m_1 + \cdots + m_n$$

$$\alpha! = m_1! m_2! \cdots m_n!$$

$$\frac{\partial^\alpha}{\partial x_\alpha} = \frac{\partial^{|\alpha|}}{\partial x_\alpha} = \frac{\partial^{m_1+\cdots+m_n}}{\partial x_1^{m_1}\cdots\partial x_n^{m_n}} .$$

We can then rephrase VIII.8.1.13.: under the hypotheses, for every α the series

$$\sum_{\omega\in\Omega} a_\omega \frac{\partial^\alpha}{\partial x_\alpha}\left((\boldsymbol{x}-\boldsymbol{c})^{\phi(\omega)}\right)$$

converges u.c. on R° to $\dfrac{\partial^\alpha f}{\partial x_\alpha}$.

VIII.8.1.15. Equality of mixed partials (VIII.2.2.2.) for functions defined by convergent power series is an immediate corollary of VIII.8.1.13., giving a simpler alternate proof in this case.

VIII.8.1.16. Note that when $\mathbb{F} = \mathbb{C}$, these partial derivatives are interpreted in the complex sense: fix the remaining variables and take the derivative as in XII.1.1.2.. Thus a complex-valued function defined on a polydisk by a multivariable power series has partial derivatives of all orders in the complex sense, which are also given by multivariable power series obtained by term-by-term differentiation.

VIII.8.2. Analytic Functions of Several Variables

In this section, we describe some of the basics of the theory of real and complex analytic functions of several variables. Some aspects of the theory are virtually identical to the one-variable case, but there are also significant differences.

VIII.8.2.1. DEFINITION. Let U be an open set in $\mathbb{F}^n$. A function $f : U \to \mathbb{F}$ is *analytic* if, for every $\boldsymbol{c} \in U$, f is represented in a neighborhood of $\boldsymbol{c}$ by a multivariable power series centered at $\boldsymbol{c}$. A function $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{F}^m$ is analytic if each f_k is analytic.

VIII.8.2.2. As in the one-variable case, this definition is a local one: a function f on an open set U is analytic on U if and only if it is analytic on a neighborhood of $\boldsymbol{c}$ for each $\boldsymbol{c} \in U$. In particular, if $U = \cup_j U_j$ is a union of open sets, then f is analytic on U if and only if it is analytic on each U_j.

An immediate corollary of VIII.8.1.13. is:

VIII.8.2.3. COROLLARY. Let U be an open set in $\mathbb{F}^n$, and f analytic on U. Then f is $\mathcal{C}^\infty$ on U, and for any multi-index α, $\frac{\partial^\alpha f}{\partial x_\alpha}$ is analytic on U.

VIII.8.2.4. An analytic function f is "jointly analytic" in its variables. If all but one variable is fixed, it is analytic in the remaining variable in the usual sense of (), since the multivariable power series for f around any point has a drastic revision obtained by collecting together all terms of like power in the remaining variable, which will converge in a neighborhood of the point in $\mathbb{F}$. (More generally, if some of the variables are fixed, the function is analytic in the remaining variables.) Thus an analytic function is "separately analytic" in its variables. But a separately real analytic function is not (jointly) analytic in general. For example,

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases}$$

is separately analytic, but not even continuous at $(0,0)$ (). This function is analytic on $\mathbb{R}^2 \setminus \{(0,0)\}$. The function

$$g(x,y) = \begin{cases} xye^{-1/(x^2+y^2)} & \text{if} \quad (x,y) \neq (0,0) \\ 0 & \text{if} \quad (x,y) = (0,0) \end{cases}$$

is separately analytic, $\mathcal{C}^\infty$ on $\mathbb{R}^2$, and analytic on $\mathbb{R}^2 \setminus \{(0,0)\}$, but not analytic on any neighborhood of $(0,0)$.

The situation is completely different in the complex case: a complex-valued function on an open subset of $\mathbb{C}^n$ which is separately analytic is in fact analytic (XII.10.1.5.).

As in the one-variable case, the definition of analytic function is hard to check, since the condition of the definition must be verified for every $\boldsymbol{c} \in U$. Thus the next result is crucial in obtaining a useful theory:

VIII.8.2.5. THEOREM. Let U be an open set in $\mathbb{F}^n$. Then any multivariable power series which converges on U defines an analytic function on U.

PROOF: Let

$$\sum_{\omega\in\Omega} a_\omega(\boldsymbol{x}-\boldsymbol{c})^{\phi(\omega)}$$

be a multivariable power series which converges on U. We may assume that U is the interior of the set on which the series converges, which is a union of polydisks centered at $\boldsymbol{c}$ (since this is an open set containing U), and in particular that $\boldsymbol{c}\in U$. 👍

Chapter IX

Riemann Integration

There are four general approaches to integration: the Riemann integral and the "modern" approaches of Lebesgue, Daniell, and Denjoy-Perron-Henstock-Kurzweil. In this chapter, we discuss the first; the others are treated in Chapter XVIII. The idea is to associate to a real- (or complex)-valued function f defined on a set X, and a subset A of X, a number $\mathcal{I}_A(f)$, usually denoted by some version of the notation $\int_A f$, reflecting the "contribution" of the function f on the set A.

IX.1. Introduction

Integration is a basic component of real analysis. The goal of integration is to systematically assign to a real-valued function f on a set X, and a subset A of X, a number $\mathcal{I}_A(f)$ called the *integral of f over A* (the notation $\int_A f$, or some variation, is usually used; the symbol $\int$ is called the *integral sign*), which is usually thought of as a "normalized sum" of the values of f on the points of the subset A. This number has various interpretations in applied situations as total distance (regarding f as giving instantaneous velocity), total mass (regarding f as giving density), etc. Integration in analysis is thus almost always (and always in this book) a generalization or analog of what in elementary calculus is called the "definite integral." (We will also have occasion to consider analogs of the "indefinite integral" of calculus, but these will have different names and will not be called integrals.)

IX.1.1. Origin of Integration Theories

IX.1.1.1. All modern theories of integration are in a sense outgrowths of the so-called Riemann integral (actually developed first by Cauchy for continuous functions and generalized by Riemann to also handle moderately discontinuous functions), the type of integration treated in elementary calculus. The Riemann integral was the first reasonably satisfactory theory of integration, and is still by far the theory most widely used in applications.

IX.1.1.2. For theoretical purposes, Riemann integration has two serious inadequacies:

1. The class of Riemann integrable functions is too small. For a function to be Riemann integrable on an interval, it must be bounded and "almost continuous" (see () for a precise statement). It is particularly useful to be able to integrate certain unbounded functions. This is reasonably, but not entirely, satisfactorily

achieved by using various notions of improper Riemann integrals and principal values. It is also important to be able to integrate over unbounded intervals, which can be done through other types of improper integrals, although not in an entirely satisfactory way. At least for theoretical purposes, it is also important to be able to integrate functions which are too discontinuous to be Riemann integrable, and to be able to integrate functions over subsets of the real numbers more general than intervals or finite unions of intervals (although the practical value of this type of integration is sometimes somewhat overemphasized as an inadequacy of Riemann integration).

2. It is crucial in many settings to have limit and approximation theorems, i.e. to know that under reasonable conditions, if $f_n \to f$ in some sense, we have $\mathcal{I}_A(f_n) \to \mathcal{I}_A(f)$. It is quite difficult to establish good theorems of this sort for Riemann integration which are applicable even if we know that the limit function is Riemann integrable (which is often not the case due to the limitations described above). It turns out that to get reasonable convergence theorems, more general integration theories are needed.

IX.1.1.3. Another limitation of Riemann integration is that it only (directly) works on $\mathbb{R}$, or on $\mathbb{R}^n$. It is very important to have integration theories which work on more general spaces, and it turns out that modern approaches give this flexibility with essentially no additional work.

IX.1.1.4. There are three principal modern approaches to integration, each with many variations. We first discuss the basics of Riemann integration in a form allowing the various generalizations to be made naturally.

IX.1.1.5. The Riemann integral has gotten somewhat of a bad rap: it is fashionable in some circles to downplay or even ignore Riemann integration in favor of more modern integration theories. There are certainly inadequacies in Riemann integration which must be addressed, and there will be no suggestion here that modern integration theories are not important. However, Riemann integration is far from obsolete! Here are some senses in which Riemann integration (and even pre-Riemann Cauchy integration) is still central:

(i) Although there are versions of the Fundamental Theorem of Calculus for more general types of integration, it is still really a theorem about Riemann (or Cauchy) integration, and most applications are to computing Riemann integrals.

(ii) All numerical integration is based on the Riemann integral.

(iii) Integrals in vector calculus, complex analysis, and differential geometry (e.g. line and surface integrals, contour integrals, integrals of differential forms) are Riemann integrals.

(iv) Many, if not most, integrals arising in physics and engineering are Riemann integrals.

(v) It is often pointed out that one rarely if ever actually computes a Lebesgue (or other modern) integral; the integral is approximated by a Riemann integral which is computed either by the Fundamental Theorem of Calculus or by numerical integration techniques.

Thus it is worthwhile to give a full treatment of the Riemann integral as well as more modern theories.

IX.1.2. The Cauchy Integral for Continuous Functions

IX.1.2.1. Actually, we will begin with a sort of "pre-Riemann integral," due to CAUCHY, in which there seems to be only one natural approach, which is fairly simple: the case of integrating a continuous function on a closed bounded interval. All theories of integration essentially reduce to the same thing in this case. Actually this case, although quite special, is by far the most important one in a practical sense – most integrals encountered in practice are of this form, or slight variations of it which can be treated on an *ad hoc* basis. In fact, although the integral in this form was not given a rigorous approach until CAUCHY'S work, it had been used by mathematicians dating back at least to ARCHIMEDES in ancient Greece; the fundamental and brilliant insight of both NEWTON and LEIBNIZ which got the whole subject of calculus off the ground was the inverse relation between integration and differentiation.

IX.1.2.2. The integral is intended to denote the "area" of the region under the graph and above the x-axis, between the lines $x = a$ and $x = b$ (Figure IX.1). To obtain an integral with good properties, the parts of the region below the axis, where the function is negative, must be interpreted as "negative area"; such regions are shown in tan. The integral thus denotes a "signed area."

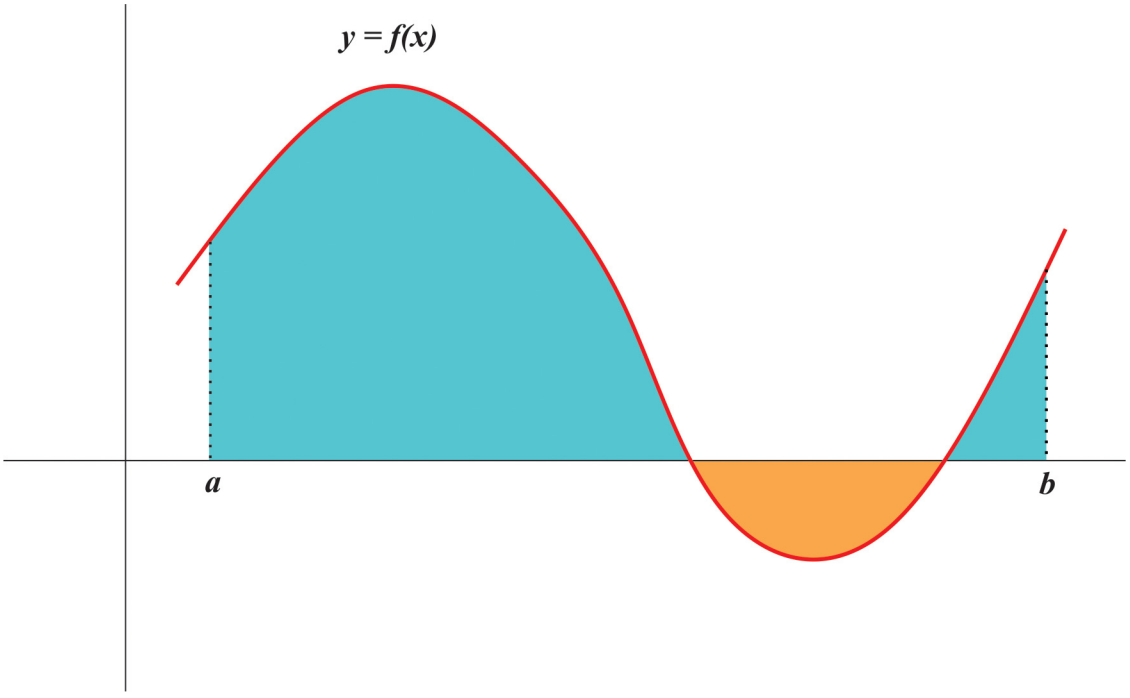


Figure IX.1: The Integral of a Continuous Function

IX.1.2.3. Suppose f is a continuous real-valued function on a closed bounded interval $J = [a, b]$. Write $\ell(J) = b - a$ for the length of J. We want to define the integral $\mathcal{I}_J(f)$ of f over J. There are three fundamental properties the integral ought to have:

(i) If $f \le g$ on J, then $\mathcal{I}_J(f) \le \mathcal{I}_J(g)$.

(ii) If f takes the constant value c on J, then $\mathcal{I}_J(f) = c \cdot \ell(J) = c(b-a)$.

(iii) If J is partitioned into nonoverlapping intervals $J_1, \dots, J_n$ ("nonoverlapping" means the subintervals only have endpoints in common), then

$$\mathcal{I}_J(f) = \sum_{k=1}^{n} \mathcal{I}_{J_k}(f) \ .$$

These three properties are enough to completely characterize the integral in this case, and to suggest an explicit method of defining and calculating its value.

IX.1.2.4. The continuity of f means that if $\mathcal{P} = \{J_1, \dots, J_n\}$ is a sufficiently fine partition of J into nonoverlapping subintervals, then f is almost constant on each J_k; thus, if t_k is any number (*tag*) in J_k, we must have

$$\mathcal{I}_J(f) \approx \sum_{k=1}^{n} f(t_k)\ell(J_k)$$

(such a sum is called a *Riemann sum* for $\mathcal{I}_J(f)$, although they were introduced by Cauchy, and Riemann did not even make extensive use of them) with the approximation becoming better and better as the partition becomes finer no matter how the t_k are chosen (within J_k). See Figure IX.2.

IX.1.2.5. This can be described more carefully in a slightly different way (this was in effect Riemann's general approach, although he phrased things somewhat differently; this formulation was explicitly introduced by Darboux after Riemann's death): if m_k and M_k are the infimum and supremum of f on J_k (in this case actually the minimum and maximum on J_k respectively), and $\omega(f, J_k) = M_k - m_k$ is the *oscillation* of f over J_k, then we define the *upper and lower estimates* $\overline{\mathcal{I}}(f, \mathcal{P})$ and $\underline{\mathcal{I}}(f, \mathcal{P})$ by

$$\overline{\mathcal{I}}(f, \mathcal{P}) = \sum_{k=1}^{n} M_k \ell(J_k) \quad \text{and} \quad \underline{\mathcal{I}}(f, \mathcal{P}) = \sum_{k=1}^{n} m_k \ell(J_k)$$

and we have that

$$\underline{\mathcal{I}}(f, \mathcal{P}) \le \mathcal{I}_J(f) \le \overline{\mathcal{I}}(f, \mathcal{P})$$

and

$$\overline{\mathcal{I}}(f, \mathcal{P}) - \underline{\mathcal{I}}(f, \mathcal{P}) = \sum_{k=1}^{n} \omega(f, J_k)\ell(J_k) \ .$$

(See Figures IX.3 and IX.4.) As the partition $\mathcal{P}$ becomes finer, we have that the $\underline{\mathcal{I}}(f, \mathcal{P})$ increase, the $\overline{\mathcal{I}}(f, \mathcal{P})$ decrease, and the $\omega(f, J_k)$ approach 0, so the upper and lower estimates approach the same limiting value (there are of course some technicalities to check here), which is the only reasonable value for $\mathcal{I}_J(f)$. We also have

$$\underline{\mathcal{I}}(f, \mathcal{P}) \le \sum_{k=1}^{n} f(x_k)\ell(J_k) \le \overline{\mathcal{I}}(f, \mathcal{P})$$

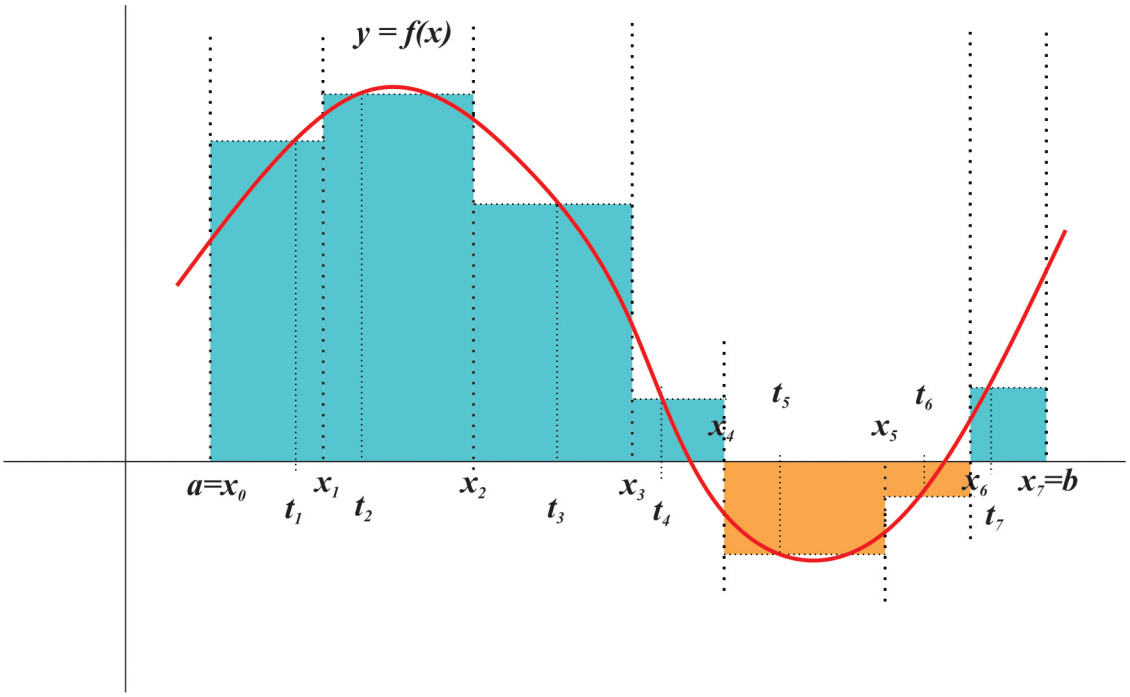


Figure IX.2: The Riemann Sum for a Tagged Partition

for any Riemann sum, so we have

$$\left|\mathcal{I}_J(f) - \sum_{k=1}^{n} f(x_k)\ell(J_k)\right| \leq \overline{\mathcal{I}}(f,\mathcal{P}) - \underline{\mathcal{I}}(f,\mathcal{P}) = \sum_{k=1}^{n} \omega(f,J_k)\ell(J_k)$$

for any Riemann sum, and thus the Riemann sums converge to $\mathcal{I}_J(f)$ as the partitions become arbitrarily fine (in the sense that the *mesh* of the partition, defined as the maximum of the lengths of the J_k, goes to 0).

IX.1.2.6. Before the rigor introduced in the nineteenth century, and informally still to this day, instead of an actual (finite) partition of J one would think of "partitioning" the interval J into an infinite number of subintervals of "infinitesimal" length dx (Figure IX.5). Then on the infinitesimal interval containing the number x, the function value would differ only infinitesimally from the number $f(x)$, so the integral over this infinitesimal subinterval would be $f(x) \cdot dx$ (up to a "second-order infinitesimal"). The total integral would then be the "sum" of these. This approach was taken by Leibniz, who developed the standard notation used for the integral even today: he denoted this "sum" $\mathbf{S}$, and wrote

$$\mathcal{I}_J(f) = \mathbf{S}_a^b f(x)\,dx\ .$$

He actually used a "long S" and almost immediately elongated it further to the familiar integral sign $\int$. (Summation notation using $\sum$ was introduced by Euler, and did not exist in Leibniz's time.)

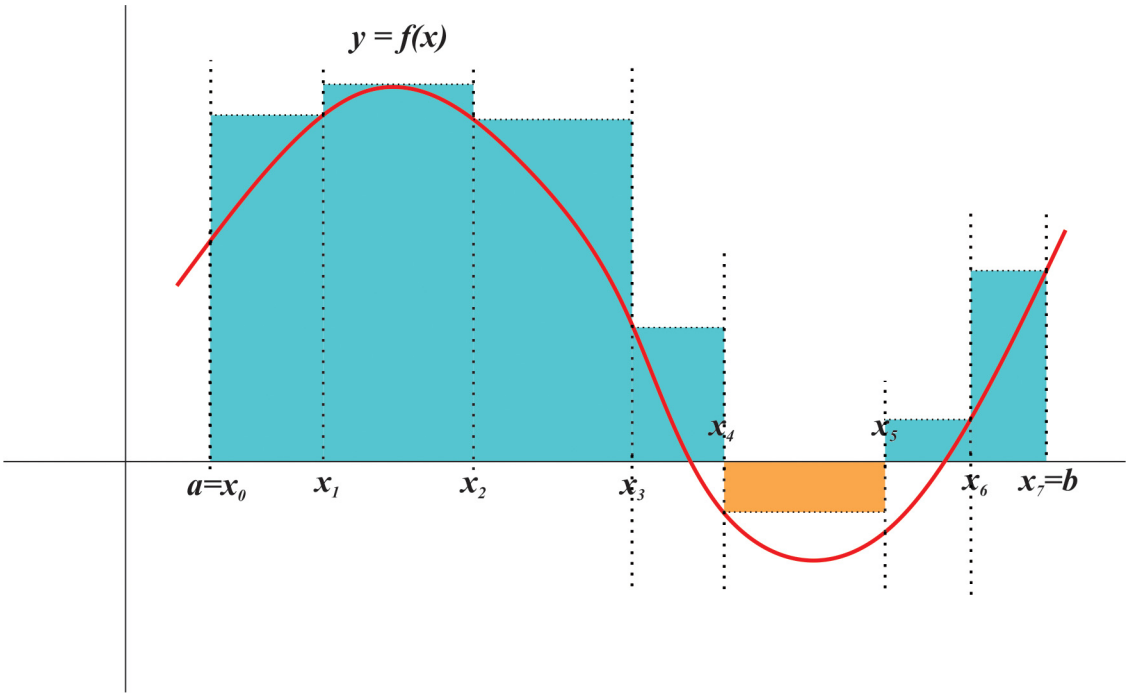


Figure IX.3: The Upper Estimate $\overline{\mathcal{I}}(f, \mathcal{P})$ for a Partition $\mathcal{P}$

IX.1.3. The Riemann Integral

IX.1.3.1. CAUCHY'S approach works in somewhat greater generality, as developed by RIEMANN. In order to have the upper and lower estimates converge together, we do not actually have to have the oscillations $\omega(f, J_k)$ approach zero as the partitions become fine; we need only that the sum ("weighted local oscillation")

$$\sum_{k=1}^{n} \omega(f, J_k)\ell(J_k)$$

becomes small, which only requires that the oscillation $\omega(f, J_k)$ be small for "most" k. See Figures IX.6 and IX.7. If f has isolated discontinuities, the weighted local oscillation is still small for fine partitions (Figure IX.8).

IX.1.3.2. Thus, if f is bounded, and continuous except at finitely many points (or, more generally, except on a set of Lebesgue measure 0), the Cauchy approach works perfectly well, either via the weighted local oscillation idea or via Riemann sums. (If f is unbounded, the weighted local oscillation approach does not make sense, and it is also easily seen that the Riemann sum approach cannot work either, at least the way we have done it using the mesh of the partitions.) The functions for which the approach works are called *Riemann-integrable functions* on J; it turns out that they are precisely the bounded functions which are continuous "almost everywhere" (except on a set of Lebesgue measure 0).

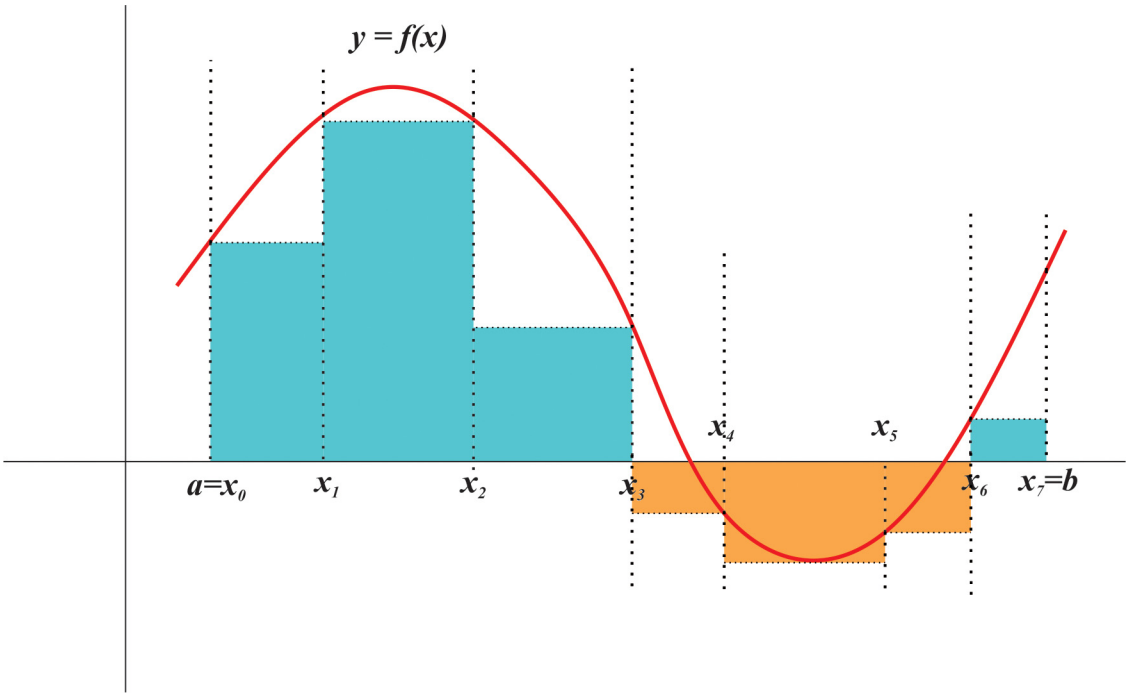


Figure IX.4: The Lower Estimate $\underline{\mathcal{I}}(f, \mathcal{P})$ for a Partition $\mathcal{P}$

IX.1.4. The Lebesgue Approach

IX.1.4.1. The first and most widely used modern approach began with LEBESGUE'S work in the early twentieth century. LEBESGUE'S idea was to consider more general partitions of the interval J on which the integral is to be defined, where the subsets A_k are not necessarily subintervals but can be more complicated, but where as in the Cauchy approach the function is almost constant on each A_k. The integral $\mathcal{I}_J(f)$ is then approximated by a sum over the sets in the partition, where the term corresponding to A_k is the almost constant value of f on A_k times the "size" of A_k.

IX.1.4.2. It is easy to find such sets A_k for any bounded function. (For motivational purposes, we will consider only bounded functions, although the actual development of the theory outlined below works equally well for unbounded functions and even extended real-valued functions.) In fact, it is not important that the domain of the function be an interval or even a set of real numbers; we can consider a bounded function f from any set X to $\mathbb{R}$ (bounded just means the range is a bounded subset of $\mathbb{R}$). Suppose $|f(x)| \leq M$ for all $x \in X$. Then, for any natural number n, we set

$$A_k = \left\{ x \in X : \frac{k-1}{n} < f(x) \leq \frac{k}{n} \right\}$$

for integers k with $|k| \leq nM$. This is a finite partition of X into pairwise disjoint subsets, and $\omega(f, A_k) \leq \frac{1}{n}$.

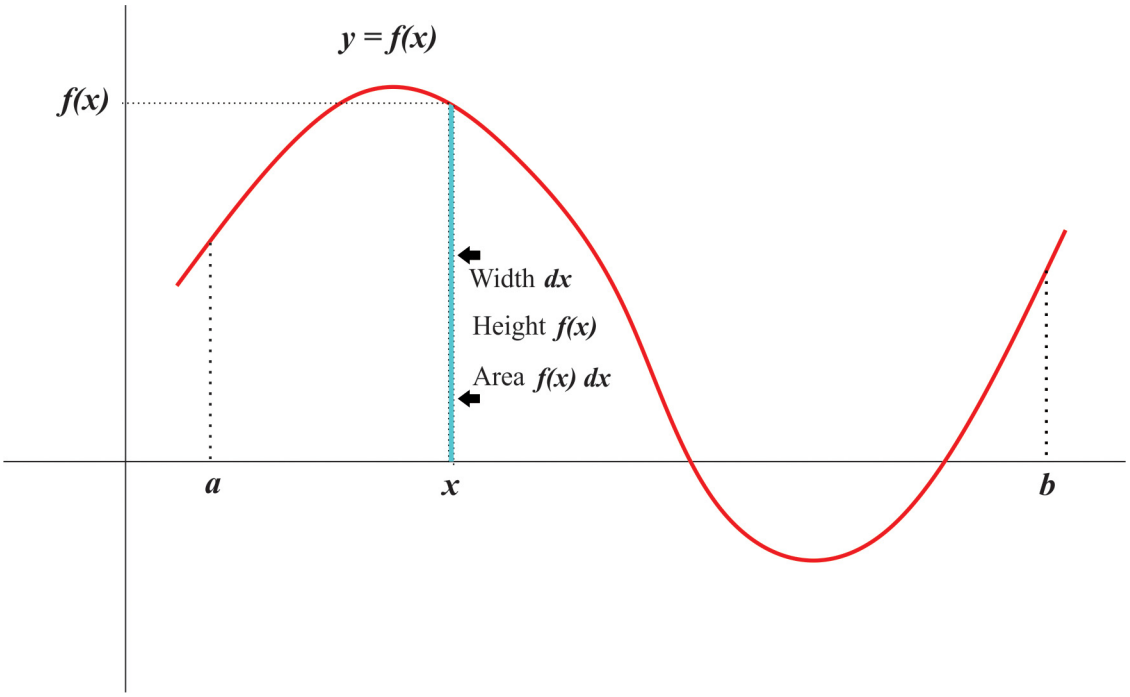


Figure IX.5: Partition into Rectangles of "Infinitesimal" Width

IX.1.4.3. We need one more ingredient, though: a "size" for each A_k. The reasonable way to do this is to require the A_k to be measurable sets for a measure μ on X; then $\mu(A_k)$ can be taken to be the "size" of A_k, and we get an approximation

$$\mathcal{I}_X(f,\mu) \approx \sum_k \frac{k}{n}\mu(A_k)$$

(the notation reflects the fact that the integral is with respect to the measure μ) which should become exact in the limit as $n \to \infty$. The restriction on f is then that the sets A_k must be μ-measurable for any choices of n and k; such an f is called a *μ-measurable function*. In most applications, including Lebesgue measure on $\mathbb{R}$, μ-measurability is a very mild restriction on f and is satisfied by any function which can be explicitly described. (There is an additional technicality: the sum may not be well defined if $\mu(A_k)$ is infinite for both positive and negative k's. This technicality forces a slightly different approach to the actual integral.)

IX.1.4.4. So in this approach, one begins with a measure space (LEBESGUE at first only considered Lebesgue measure on $\mathbb{R}$, but the general theory is nearly identical). As mentioned, for technical reasons, we will not proceed exactly as above, but the motivational description is an accurate reflection of the essence of the process.

IX.1.4.5. If μ is a measure on a set X, we build up the integral corresponding to μ in stages. We first consider indicator functions (often called characteristic functions, but this term also has other uses in

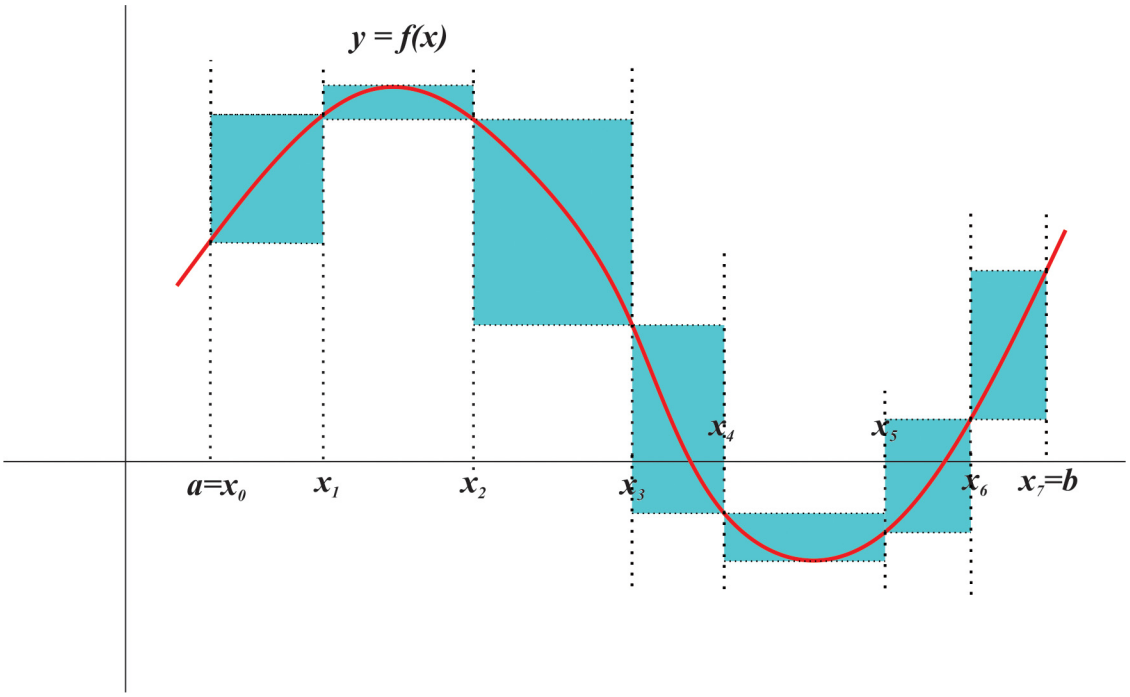


Figure IX.6: The Weighted Local Oscillation for a Partition

analysis). If $A \subseteq X$, let χ_A denote the function from X to $\mathbb{R}$ defined by

$$\chi_A(x) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases} .$$

This χ_A is called the *indicator function* of the set A. It is reasonable to define the integral of χ_A to be the measure $\mu(A)$ of the set A (provided, of course, that A is a μ-measurable set, or equivalently that χ_A is a μ-measurable function).

IX.1.4.6. Next, we consider *simple functions*, finite nonnegative linear combinations of indicator functions of μ-measurable sets. Such a function has a representation

$$s = \sum_{k=1}^{n} a_k \chi_{A_k}$$

where the A_k are μ-measurable sets and the a_k are nonnegative real numbers. These are precisely the nonnegative μ-measurable functions taking only finitely many values. It then makes sense to define the integral by

$$\mathcal{I}_X(s, \mu) = \sum_{k=1}^{n} a_k \mu(A_k)$$

and indeed this is the only possible definition to obtain a linear integral. (There is a technicality here: it must be shown that this number is well defined by s, since the representation $s = \sum a_k \chi_{A_k}$ is not unique.)

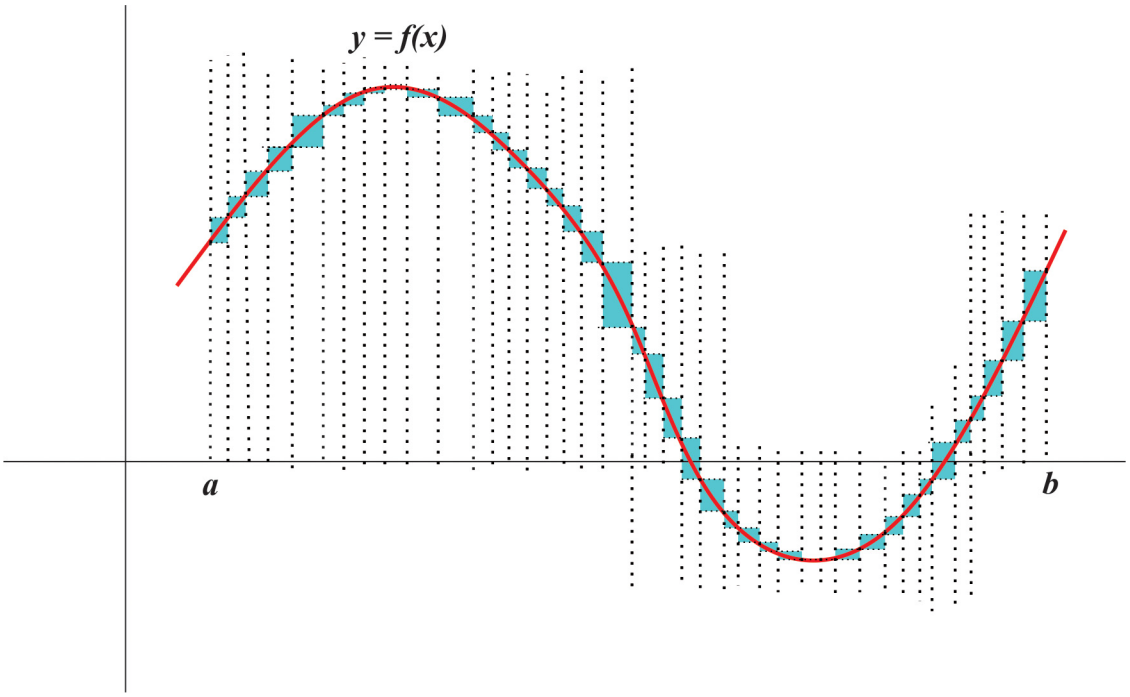


Figure IX.7: The Weighted Local Oscillation for a Finer Partition

IX.1.4.7. If f is a nonnegative μ-measurable function, we then define

$$\mathcal{I}_X(f,\mu) = \sup\{\mathcal{I}_X(s,\mu) : s \text{ simple}, 0 \le s \le f\} .$$

This definition even makes sense for extended real-valued functions. Even if f is real-valued, $\mathcal{I}_X(f,\mu)$ can equal $+\infty$. We say a nonnegative function f is μ-*integrable* if f is μ-measurable and $\mathcal{I}_X(f,\mu) < \infty$.

IX.1.4.8. Finally, if f is a real-valued μ-measurable function, and there are nonnegative μ-integrable functions g and h with $f = g - h$, we say f is μ-integrable and set

$$\mathcal{I}_X(f,\mu) = \mathcal{I}_X(g,\mu) - \mathcal{I}_X(h,\mu)$$

(it must be shown that this is well defined by f independent of the choices of g and h).

IX.1.4.9. We can extend the notion of integration over the whole measure space X to integration over any μ-measurable set A by setting

$$\mathcal{I}_A(f,\mu) = \mathcal{I}_X(f\chi_A,\mu)$$

where $f\chi_A$ is the function which is equal to f on A and is identically zero outside A; this definition even makes sense if f is only defined on the subset A.

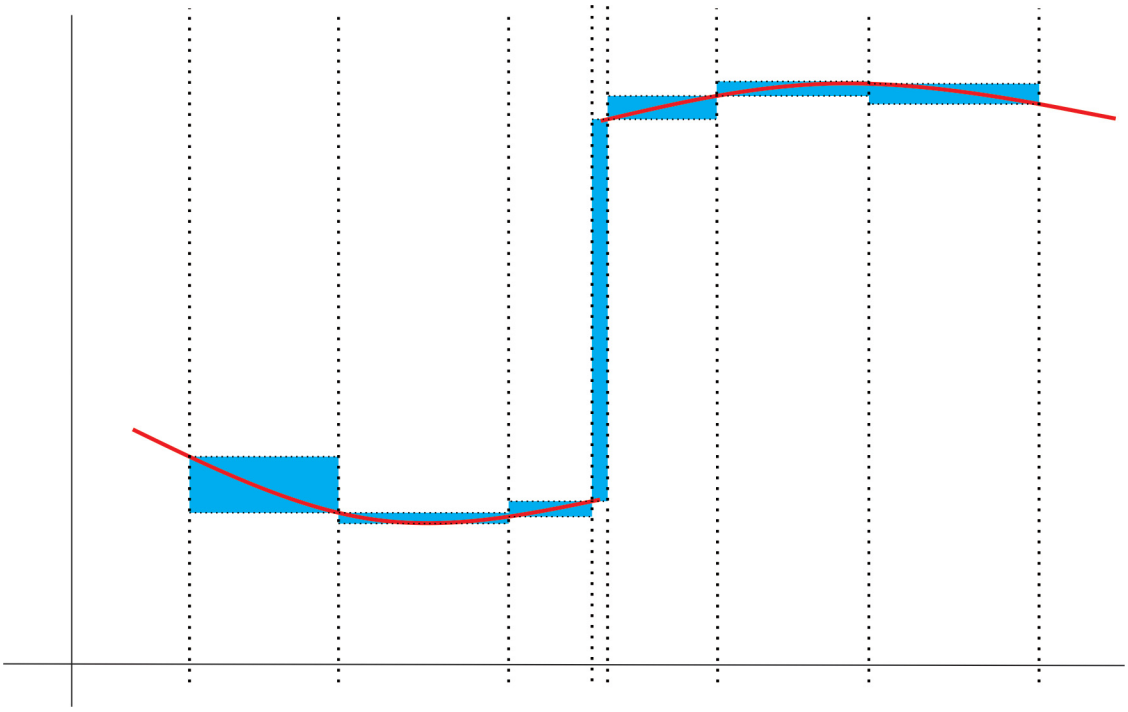


Figure IX.8: The Weighted Local Oscillation Around an Isolated Discontinuity

IX.1.4.10. It turns out that this integral $\mathcal{I}_X$ has all the nice properties one would expect, and almost all one might hope for, in an integral. In the case where μ is Lebesgue measure λ on $\mathbb{R}$, this integral extends the Riemann integral in the sense that any Riemann-integrable function f on a closed bounded interval J is Lebesgue-integrable (λ-integrable) on J and the the Riemann integral $\mathcal{I}_J(f)$ of f over J agrees with the Lebesgue integral $\mathcal{I}_J(f, \lambda)$ of f over J. But many more functions are Lebesgue integrable than are Riemann-integrable, including many unbounded functions.

IX.1.4.11. The Lebesgue approach to integration on a general measure space is also very important. For example, if $(\Omega, \mathcal{E}, \mathbf{P})$ is a probability space, a $\mathbf{P}$-measurable real-valued function X on Ω is called a *random variable*, and if X is nonnegative or $\mathbf{P}$-integrable, its integral is called the *expected value* of the random variable X.

IX.1.4.12. There are a couple of drawbacks to the Lebesgue approach, however. In order to define the Lebesgue integral on $\mathbb{R}$, Lebesgue measure must first be constructed, which is a moderately complicated task. The Lebesgue integral on a general measure space is somewhat simpler, but preliminaries are not completely avoided: besides the simple basics of the theory of measure spaces from (), one needs to know about measurable functions, where the details () are somewhat more challenging.

IX.1.4.13. The other drawback of Lebesgue integration from a certain point of view is that it is an "absolute" theory of integration: if f is measurable, then f is integrable if and only if $|f|$ is integrable.

Roughly, this means that both the positive and negative parts of f are not too large. There are situations where the improper Riemann integral of a function f taking both positive and negative values is defined, yet f is not Lebesgue integrable. On the other hand, the "absolute" property is what makes Lebesgue integration well-behaved; it is a sort of analog of absolute convergence of infinite series, which forces all sorts of nice behavior which does not carry over to conditionally convergent series. (The connection between integration and infinite series is much more than just an analog; the theory of infinite series can be interpreted as a special case of the theory of measure and integration – see ().)

IX.1.5. The Daniell Approach

IX.1.5.1. Either measure or integration can be regarded as the primitive notion, and the other one can be derived from it. We have seen how integration can be derived from measure via the Lebesgue approach. In Euclidean space, there is a more direct approach simply consisting of increasing dimension by one: if f is a nonnegative real-valued function on a subset E of $\mathbb{R}^n$, the integral $\int_E f\, d\lambda^{(n)}$ is simply the $(n+1)$-dimensional Lebesgue measure of the "subgraph"

$$\{(x_1, \ldots, x_n, y) : (x_1, \ldots, x_n) \in E, 0 \leq y \leq f(x_1, \ldots, x_n)\} \subseteq \mathbb{R}^{n+1}$$

(in particular, the integral of a nonnegative measurable function over a subset of $\mathbb{R}$ is just the measure of a corresponding subset of the plane). Conversely, measure can be recovered from integration: the measure of a subset E of $\mathbb{R}^n$ (or of a general measure space) is just the integral of the indicator function χ_E over $\mathbb{R}^n$.

IX.1.5.2. Taking integration as the primitive concept and defining measure from it is the approach taken by Daniell. In this approach, we begin with a sufficiently large supply of real-valued functions on a set X for which there is an "integral" defined having reasonable linearity, positivity, and continuity properties. The actual mechanism by which the integral is defined is unimportant; it can even just be an abstract real-valued functional defined on the set of functions. If necessary, the integral is then extended in a systematic way to a larger class of real-valued functions on X including indicator functions of enough subsets of X. A measure μ on X can then be defined by setting $\mu(E)$ for a subset E of X to be the integral of χ_E. A good motivating example (which was Daniell's own starting point) is to take X to be an interval $[a, b]$ in $\mathbb{R}$, the functions to be the real-valued continuous functions on $[a, b]$, and the integral the Cauchy integral. The measure obtained by the Daniell process is then just Lebesgue measure on $[a, b]$. This turns out to actually be a rather painless way to define Lebesgue measure.

IX.1.5.3. An important generalization of the motivating example is the *Riesz Representation Theorem*: if X is a compact Hausdorff space and ϕ is a positive linear functional on the set $C(X)$ of real-valued (or complex-valued) continuous functions on X, i.e. an assignment of a number $\phi(f)$ to each continuous function f with the properties that

(i) $\phi(f + g) = \phi(f) + \phi(g)$ for all $f, g \in C(X)$.

(ii) $\phi(\alpha f) = \alpha\phi(f)$ for every $f \in C(X)$ and every scalar α.

(iii) $\phi(f) \geq 0$ if f takes only nonnegative real values.

then there is a finite measure μ on the σ-algebra of subsets of X generated by the open sets (the *Borel sets*) such that $\phi(f) = \int_X f\, d\mu$ for all f. The measure μ is constructed by the Daniell process of extending the integral (linear functional) to a larger class of functions including indicator functions of Borel sets.

IX.1.6. Direct Generalizations of Riemann Integration

IX.1.6.1. A third approach to generalizing Riemann integration is to directly modify the definition of the Riemann integral to obtain a more-inclusive integral on $\mathbb{R}$. This approach was notably used by DENJOY and PERRON, and culminated relatively recently (1957) in the so-called HENSTOCK-KURZWEIL (H-K) integral, which agrees with the integrals of DENJOY and PERRON but has a much better definition.

IX.1.6.2. The H-K integral is easier to define than the Lebesgue integral (Lebesgue measure is not needed), and almost as easy as the Riemann integral; and it remedies both shortcomings of the Riemann integral. In fact, it turns out to be almost the same as the Lebesgue integral, and even slightly more general: a nonnegative function is H-K integrable if and only if it is Lebesgue integrable, but a function taking both positive and negative values can be H-K integrable but not Lebesgue integrable if its positive and negative parts are too big but "cancel out" in a suitable way.

IX.1.6.3. The H-K integral is defined using the Riemann sum approach. The main difference from Riemann integration is a much more delicate notion of the "fineness" of a partition into subintervals than just its mesh. We will not give a precise description here (see ()), but the fineness of a partition depends not only on the intervals J_k but also on which tags $t_k \in J_k$ are chosen. It is then a somewhat subtle matter to show that arbitrarily fine partitions exist. Once this has been done, the rest of the development formally follows the theory of Riemann integration very closely. But many more functions turn out to be H-K integrable.

IX.1.6.4. The H-K integral is nice to work with in many ways. For example, integrals which would naturally be called "improper H-K integrals" turn out to be honest H-K integrals. One of the nicest properties of H-K integration is that the following version of the Fundamental Theorem of Calculus, which fails spectacularly for Riemann integration and even fails for Lebesgue integration, holds for H-K integration:

IX.1.6.5. THEOREM. Let f be a differentiable function on a closed bounded interval $[a, b]$. Then f' is H-K integrable on $[a, b]$ (and hence on any subinterval), and for every $x \in [a, b]$ we have

$$f(x) = f(a) + \int_a^x f'(t)\, dt$$

where the integral is the H-K integral of f' over $[a, x]$.

IX.1.6.6. On the other hand, the H-K integral is strictly an extension of Riemann integration on $\mathbb{R}$, and does not generalize the way Lebesgue integration does to other measure spaces. (There is a version of H-K integration on $\mathbb{R}^n$, which is not nearly as satisfactory as on $\mathbb{R}$, but nothing more general.)

IX.1.7. Oriented and Unoriented Integration

IX.1.7.1. There are two general kinds of integration: oriented integration and unoriented integration (they are not very different, differing only in sign, but the sign is often crucial in applications). The two types both appear in Riemann integration: if I is a closed bounded interval $[a, b]$, with $a < b$, then the integral

$$\int_a^b f(x)\,dx$$

of a function f from I to $\mathbb{R}$ can be regarded either as an integral *over* I ("unoriented") or as an integral *from* a *to* b ("oriented"). The expression

$$\int_b^a f(x)\,dx$$

appears to be rather artificial and intrinsically meaningless, and the definition

$$\int_b^a f(x)\,dx = -\int_a^b f(x)\,dx$$

is seemingly arbitrary, although it is useful in a number of natural contexts, for example making the formula

$$\int_a^c f(x)\,dx = \int_a^b f(x)\,dx + \int_b^c f(x)\,dx$$

valid for any a, b, c no matter what their order (so long as they are all contained in some interval on which f is integrable). Another important use of this convention is that if f is continuous on an interval I and $x_0 \in I$, not necessarily the left endpoint, then the formula

$$F(x) = \int_{x_0}^x f(t)\,dt$$

defines the unique antiderivative of f on I satisfying $F(x_0) = 0$.

IX.1.7.2. Integrating from right to left in $\mathbb{R}$ seems unnatural since there is a natural orientation of $\mathbb{R}$. However, in Euclidean space there is no natural orientation; there are two equally natural possibilities (choice of a coordinate system specifies one of the two orientations; cf. VI.3.1.1., ()). Complicated subsets of $\mathbb{R}^n$ do not have any obvious intrinsic notion of orientation. Thus in the Lebesgue approach to integration it is natural to use unoriented integration. And when the generalization to abstract measure spaces is made, there is no notion of orientation available at all, so unoriented integration is the only possibility. Thus the theory of integration in measure theory is (almost) exclusively unoriented integration.

IX.1.7.3. On the other hand, in vector calculus and analysis such integrals as line and surface integrals are important, and in Complex Analysis contour integrals are basic; such integrals incorporate a natural orientation, and the orientation is crucial in many results. In fact, any result relating integration and differentiation along the lines of the Fundamental Theorem of Calculus, e.g. the various versions of Stokes' Theorem, necessarily involves oriented integration, since differentiation itself requires orientation (this is obscured for functions from $\mathbb{R}$ to $\mathbb{R}$ since $\mathbb{R}$ has a natural orientation, but orientations must be chosen to define the derivative of a real-valued function on a one-dimensional Euclidean space or a function between two one-dimensional Euclidean spaces). It turns out that many deep results in geometry and topology involve oriented integration.

IX.1.7.4. As a general (but not universal) rule, oriented integration is usually done only for smooth objects and is primarily used in geometry, while unoriented integration is done for not necessarily smooth or even continuous objects and is primarily used in analysis. That these statements are not categorically true is partly due to the fact that the boundary between analysis and geometry is quite blurred. We will examine important examples of both types of integration.

IX.2. The Riemann Integral

We first examine the integral commonly called the Riemann integral. This integral was actually developed informally by NEWTON and LEIBNIZ (based in part on ideas going back to ancient Greece), and formally developed by CAUCHY for continuous functions and expanded and put on a "modern" foundation by RIEMANN, with important contributions by others, notably DIRICHLET and DARBOUX. See [**?**] for a detailed historical account of the development of this integral.

It is a curious fact that there is no consensus among mathematicians at the present time about the proper definition of the Riemann integral and of Riemann-integrable functions. Several rather different-looking (but equivalent) definitions can be found in standard references and texts, with other variations given as theorems. While it is pretty easy to see that any two of these definitions give the same *value* for the integral of any function satisfying both definitions, it is not entirely obvious that they apply to exactly the same functions (i.e. give the same class of integrable functions). We will discuss the various definitions and show that they are actually all equivalent. A normal pedagogical approach would be to pick one definition as *the* definition, and prove that the others are equivalent characterizations; but the lack of a standard definition makes the several-definition approach more reasonable in this case.

The multitude of definitions can be somewhat of a conceptual obstacle to focusing on the essential nature of the Riemann integral. However, the situation can, and should, be viewed positively: the fact that Riemann-integrability can be viewed in a number of ways is of theoretical importance and an indication that it is a natural concept. And the large supply of equivalent definitions is very useful in practice: in any situation, we are free to use whichever definition is most convenient to work with at the moment.

The various definitions are all precise versions of the following informal definition:

DEFINITION. (INFORMAL) Let f be a real-valued function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J if the sums

$$\sum_{k=1}^{n} c_k \ell(J_k)$$

where $\{J_1, \ldots, J_n\}$ is a partition of J into subintervals and c_k is an approximation to the value of f on J_k, converge to a limiting value L as the partition becomes sufficiently fine. The limiting value L is called the *(Riemann) integral of f over J*, denoted

$$\int_a^b f(x)\,dx \ .$$

There are, of course, several aspects of this "definition" which must be made precise to have a real definition. It is not entirely clear how to make some parts of the definition precise, particularly the part about the partition becoming "sufficiently fine." (With the right interpretation of this phrase, the statement actually becomes a definition of the Henstock-Kurzweil integral.)

The integral can be interpreted as the "area" of the region under the graph and above the x-axis, between the lines $x = a$ and $x = b$. Parts of the region below the axis, where the function is negative, are interpreted as "negative area"; such regions are shown in tan. The integral thus denotes a "signed area."

IX.2.1. Partitions, Tags, Riemann and Darboux Sums

In this subsection we make a number of definitions and set up the notation we will use for Riemann integration.

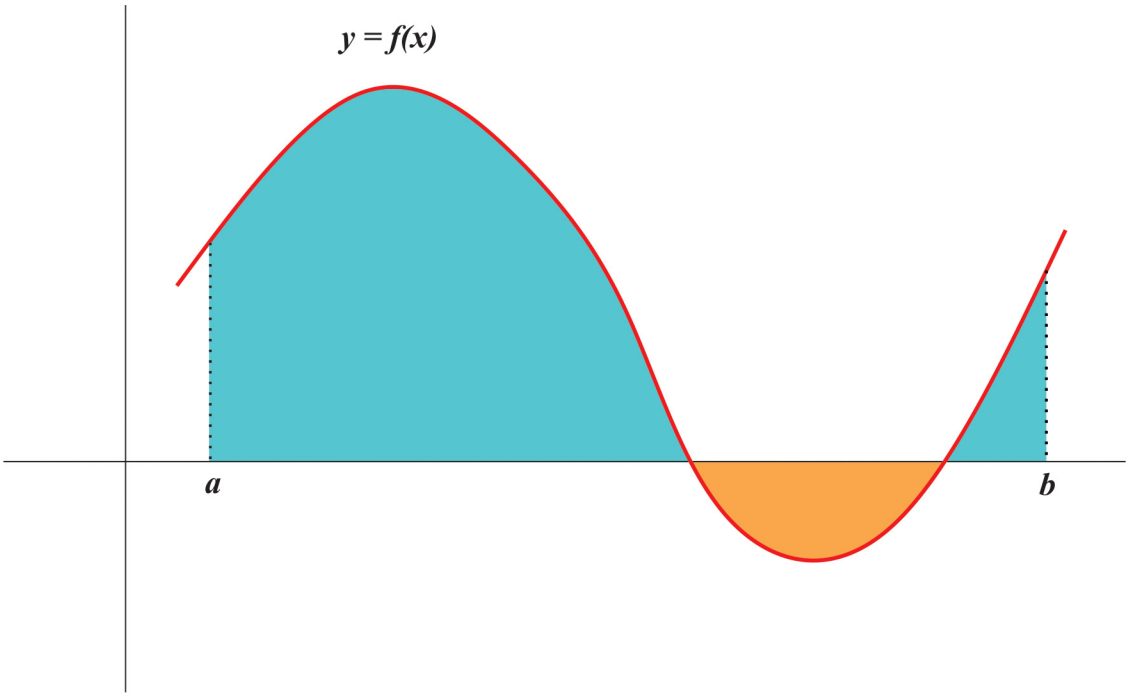


Figure IX.9: The Integral of a Continuous Function

Partitions

IX.2.1.1. DEFINITION. Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$. A *partition* of $[a, b]$ is a division of $[a, b]$ into a finite number of subintervals $\{[x_{k-1}, x_k] : 1 \leq k \leq n\}$ with only endpoints in common. A partition $\mathcal{P}$ is usually described by its set of endpoints $\{x_0, x_1, \ldots, x_n\}$, where $a = x_0 < x_1 < \cdots < x_n = b$. We can also write $\mathcal{P} = \{J_1, \ldots, J_n\}$, where $J_k = [x_{k-1}, x_k]$ is the k'th subinterval. We sometimes write Δx_k for $\ell(J_k) = x_k - x_{k-1}$, the length of the k'th subinterval of the partition.

The *norm* or *mesh* of a partition $\mathcal{P} = \{x_0, x_1, \ldots, x_n\}$ is

$$\|\mathcal{P}\| = \max\{x_k - x_{k-1} : 1 \leq k \leq n\}.$$

If $\mathcal{P} = \{x_0, \ldots, x_n\}$ and $\mathcal{Q} = \{x'_0, \ldots, x'_m\}$ are partitions of $[a, b]$, we say that $\mathcal{Q}$ is *finer* than $\mathcal{P}$, or a *refinement* of $\mathcal{P}$, if $\mathcal{P} \subseteq \mathcal{Q}$, i.e. if each x_k is one of the x'_j, or equivalently if each interval in the partition $\mathcal{Q}$ is a subinterval of one of the intervals in $\mathcal{P}$.

A *tag* in an interval $[c, d]$ is a specified number in $[c, d]$. A *tagged partition* is a partition $\{x_0, x_1, \ldots, x_n\}$ with a specified tag t_i in $[x_{i-1}, x_i]$ for each i. A tagged partition is usually denoted by

$$\dot{\mathcal{P}} = \{x_0, x_1, \ldots, x_n; t_1, \ldots, t_n\} .$$

The norm or mesh of a partition is a measure of how fine the partition is. There is a more sophisticated and flexible notion of fineness of a partition based on gauges ().

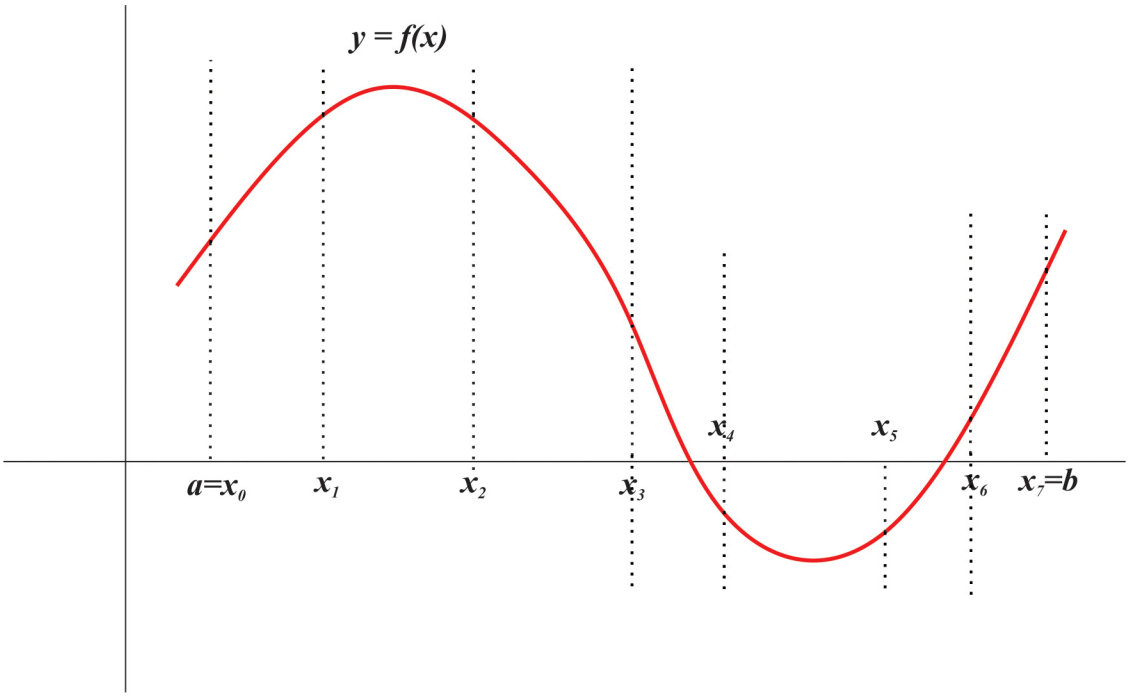


Figure IX.10: A partition with Seven Subintervals

Note that if $\mathcal{P}$ and $\mathcal{Q}$ are two partitions of $[a, b]$, neither is a refinement of the other in general. But:

IX.2.1.2. PROPOSITION. Two partitions $\mathcal{P}$ and $\mathcal{Q}$ of an interval $[a, b]$ have a common refinement.

PROOF: Just take $\mathcal{P} \cup \mathcal{Q}$, the partition obtained by using all the endpoints in both partitions. 👍

Riemann Sums

IX.2.1.3. DEFINITION. Let f be a real-valued function on an interval $[a, b]$, and let

$$\dot{\mathcal{P}} = \{x_0, x_1, \ldots, x_n; t_1, \ldots, t_n\}$$

be a tagged partition of $[a, b]$. The *Riemann sum* for f with respect to $\dot{\mathcal{P}}$ is

$$\mathcal{R}(f, \dot{\mathcal{P}}) = \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}) = \sum_{k=1}^{n} f(t_k)\,\Delta x_k \ .$$

Note that Riemann sums are only defined for tagged partitions, and the numerical value depends on the choice of tags as well as the partition.

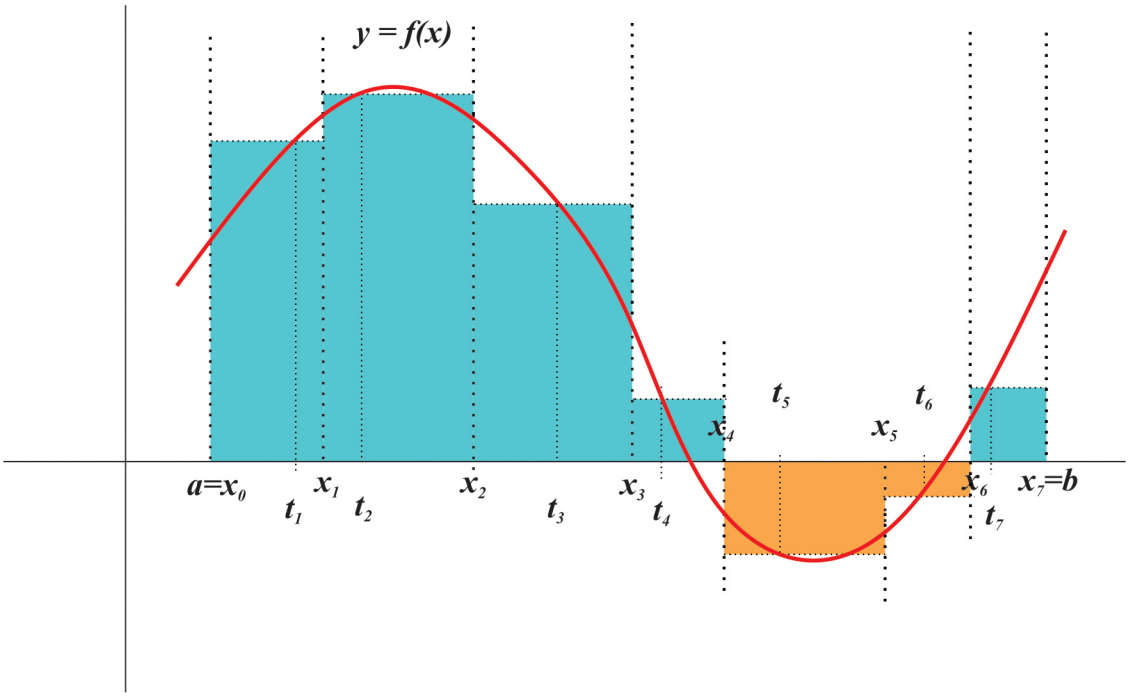


Figure IX.11: The Riemann Sum for a Tagged Partition

Darboux Sums and Oscillation

Two closely related sums which do not depend on tags are the upper and lower Darboux sums:

IX.2.1.4. DEFINITION. Let f be a bounded function on an interval $[a,b]$. If $\mathcal{P} = \{x_0, \ldots, x_n\}$ is a partition of $[a,b]$, for $1 \le k \le n$ set

$$M_k = \sup_{x_{k-1} \le x \le x_k} f(x) \text{ and } m_k = \inf_{x_{k-1} \le x \le x_k} f(x) \ .$$

Then the *upper and lower Darboux sums* for f with respect to $\mathcal{P}$ are

$$\overline{\mathcal{S}}(f,\mathcal{P}) = \sum_{k=1}^{n} M_k(x_k - x_{k-1}) = \sum_{k=1}^{n} M_k \, \Delta x_k$$

$$\underline{\mathcal{S}}(f,\mathcal{P}) = \sum_{k=1}^{n} m_k(x_k - x_{k-1}) = \sum_{k=1}^{n} m_k \, \Delta x_k$$

(written $U(f,\mathcal{P})$ and $L(f,\mathcal{P})$ in many references).

The relation between Darboux and Riemann sums is:

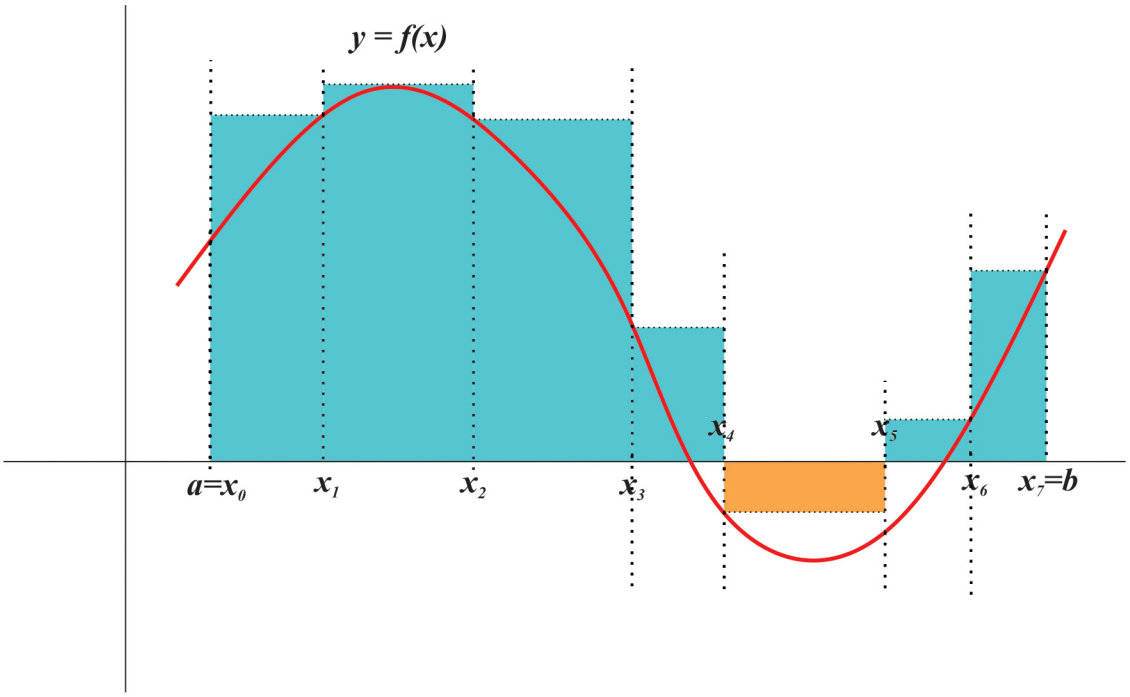


Figure IX.12: The Upper Darboux Sum for a Partition

IX.2.1.5. PROPOSITION. Let f be a bounded function on an interval $[a,b]$, and let $\mathcal{P}$ be a partition of $[a,b]$. Then

(i) If $\dot{\mathcal{P}}$ is any tagging of $\mathcal{P}$, then

$$\underline{\mathcal{S}}(f,\mathcal{P}) \leq \mathcal{R}(f,\dot{\mathcal{P}}) \leq \overline{\mathcal{S}}(f,\mathcal{P}) \ .$$

(ii) For any $\epsilon > 0$ there are taggings $\dot{\mathcal{P}}$ and $\ddot{\mathcal{P}}$ of $\mathcal{P}$ such that

$$\mathcal{R}(f,\dot{\mathcal{P}}) > \overline{\mathcal{S}}(f,\mathcal{P}) - \epsilon$$

$$\mathcal{R}(f,\ddot{\mathcal{P}}) < \underline{\mathcal{S}}(f,\mathcal{P}) + \epsilon \ .$$

PROOF: Part (i) is obvious. For (ii), let $\mathcal{P} = \{x_0,\ldots,x_n\}$. For each k, choose $t_k \in [x_{k-1},x_k]$ with

$$f(t_k) > M_k - \frac{\epsilon}{n\Delta x_k}$$

and set $\dot{\mathcal{P}} = \{x_0,\ldots,x_k;t_1,\ldots,t_k\}$. The construction of $\ddot{\mathcal{P}}$ is similar. 👍

As the partition $\mathcal{P}$ becomes finer, the upper Darboux sums decrease and the lower Darboux sums decrease:

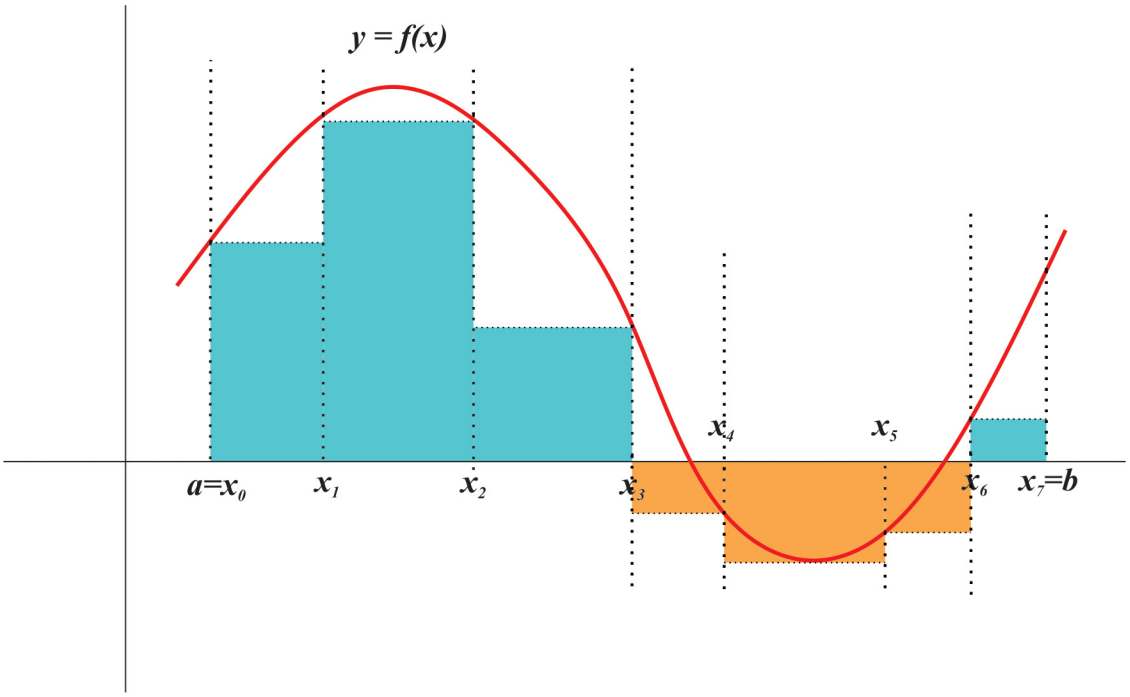


Figure IX.13: The Lower Darboux Sum for a Partition

IX.2.1.6. PROPOSITION. Let f be a bounded function on an interval $[a, b]$, and $\mathcal{P}$ and $\mathcal{Q}$ partitions of $[a, b]$. If $\mathcal{Q}$ is finer than $\mathcal{P}$, then

$$\underline{\mathcal{S}}(f, \mathcal{P}) \leq \underline{\mathcal{S}}(f, \mathcal{Q}) \leq \overline{\mathcal{S}}(f, \mathcal{Q}) \leq \overline{\mathcal{S}}(f, \mathcal{P}) \ .$$

There is no such simple relation for Riemann sums.

Interior Darboux Sums

For technical purposes, there is a useful variation on the notion of a Darboux sum:

IX.2.1.7. DEFINITION. Let f be a bounded function on an interval $[a, b]$. If $\mathcal{P} = \{x_0, \ldots, x_n\}$ is a partition of $[a, b]$, for $1 \leq k \leq n$ set

$$M_k^\circ = \sup_{x_{k-1} < x < x_k} f(x) \text{ and } m_k^\circ = \inf_{x_{k-1} < x < x_k} f(x) \ .$$

Then the *upper and lower interior Darboux sums* for f with respect to $\mathcal{P}$ are

$$\overline{\mathcal{S}}_\circ(f, \mathcal{P}) = \sum_{k=1}^{n} M_k^\circ (x_k - x_{k-1}) = \sum_{k=1}^{n} M_k^\circ \, \Delta x_k$$

$$\underline{\mathcal{S}}_\circ(f,\mathcal{P}) = \sum_{k=1}^{n} m_k^\circ(x_k - x_{k-1}) = \sum_{k=1}^{n} m_k^\circ\,\Delta x_k\ .$$

The difference is that M_k° and m_k° are the supremum and infimum of f on the *interior* of the k'th subinterval.

IX.2.1.8. For any f and $\mathcal{P}$, we obviously have $m_k \le m_k^\circ \le M_k^\circ \le M_k$, so

$$\underline{\mathcal{S}}(f,\mathcal{P}) \le \underline{\mathcal{S}}_\circ(f,\mathcal{P}) \le \overline{\mathcal{S}}_\circ(f,\mathcal{P}) \le \overline{\mathcal{S}}(f,\mathcal{P})\ .$$

If f is continuous on $[a,b]$, then $M_k^\circ = M_k$ and $m_k^\circ = m_k$ for all k, so we have equality at the ends, but the interior Darboux sums for step functions often differ from the usual Darboux sums. We will obtain an approximate equivalence between the notions in ().

The Metaprinciple

There is a simple but useful observation about the behavior of Riemann sums, Darboux sums, and interior Darboux sums:

IX.2.1.9. Proposition. Let f be a bounded function on an interval $[a,b]$, $\mathcal{P}$ a partition of $[a,b]$, and $\dot{\mathcal{P}}$ any tagging of $\mathcal{P}$. We have the following relations for any constant c:

$$\begin{aligned}
\mathcal{R}(f+c,\dot{\mathcal{P}}) &= \mathcal{R}(f,\dot{\mathcal{P}}) + c(b-a)\\
\overline{\mathcal{S}}(f+c,\mathcal{P}) &= \overline{\mathcal{S}}(f,\mathcal{P}) + c(b-a)\\
\underline{\mathcal{S}}(f+c,\mathcal{P}) &= \underline{\mathcal{S}}(f,\mathcal{P}) + c(b-a)\\
\overline{\mathcal{S}}_\circ(f+c,\mathcal{P}) &= \overline{\mathcal{S}}_\circ(f,\mathcal{P}) + c(b-a)\\
\underline{\mathcal{S}}_\circ(f+c,\mathcal{P}) &= \underline{\mathcal{S}}_\circ(f,\mathcal{P}) + c(b-a)\\
\mathcal{R}(-f,\dot{\mathcal{P}}) &= -\mathcal{R}(f,\dot{\mathcal{P}})\\
\overline{\mathcal{S}}(-f,\mathcal{P}) &= -\underline{\mathcal{S}}(f,\mathcal{P})\\
\underline{\mathcal{S}}(-f,\mathcal{P}) &= -\overline{\mathcal{S}}(f,\mathcal{P})\\
\overline{\mathcal{S}}_\circ(-f,\mathcal{P}) &= -\underline{\mathcal{S}}_\circ(f,\mathcal{P})\\
\underline{\mathcal{S}}_\circ(-f,\mathcal{P}) &= -\overline{\mathcal{S}}_\circ(f,\mathcal{P})\ .
\end{aligned}$$

IX.2.1.10. Thus, in analyzing and comparing such sums, it suffices to consider nonnegative functions by adding a suitable constant; and results about upper Darboux sums have an exact analog for lower Darboux sums which do not have to be proved separately, but simply follow by taking negations.

Relation between Darboux Sums and Interior Darboux Sums

The next result is a good example of the metaprinciple in action.

IX.2.1.11. PROPOSITION. Let f be a bounded function on an interval $[a,b]$, and $\mathcal{P}$ a partition of $[a,b]$. Then for every $\epsilon > 0$ there is a partition $\mathcal{Q}$ of $[a,b]$ such that

$$\underline{\mathcal{S}}_\circ(f,\mathcal{P}) - \epsilon \le \underline{\mathcal{S}}(f,\mathcal{Q}) \le \overline{\mathcal{S}}(f,\mathcal{Q}) \le \overline{\mathcal{S}}_\circ(f,\mathcal{P}) + \epsilon \ .$$

PROOF: By the Metaprinciple, we need show there is a $\mathcal{Q}$ giving the last inequality: then by negation, there is a $\mathcal{Q}'$ giving the first inequality, and a common refinement of $\mathcal{Q}$ and $\mathcal{Q}'$ gives both. Furthermore, in finding $\mathcal{Q}$ for the last inequality we may assume that f is nonnegative.

Let $\mathcal{P} = \{x_0, \dots, x_n\}$. We will obtain $\mathcal{Q}$ by shortening each interval in $\mathcal{P}$ and adding additional small intervals around x_k for each k: given $\epsilon > 0$, choose $\delta > 0$ small enough that $\delta < \ell(J_k)/2$ for all k, and so that $\delta < \frac{\epsilon}{2M(n+1)}$, where M is a bound for f on $[a,b]$, and let

$$\mathcal{Q} = \{x_0, x_0+\delta, x_1-\delta, x_1+\delta, \dots, x_n-\delta, x_n\} \ .$$

Then for each k we have

$$\sup_{x_{k-1}+\delta \le x \le x_k - \delta} f(x) \le M_k^\circ$$

and so by adding up separately over the truncated original intervals and the new intervals we have

$$\overline{\mathcal{S}}(f,\mathcal{Q}) \le \sum_{k=1}^n M_k^\circ[\ell(J_k) - 2\delta] + \sum_{k=0}^n M(2\delta) \le \overline{\mathcal{S}}_\circ(f,\mathcal{P}) + \epsilon \ .$$



Oscillation

The oscillation of a bounded real-valued function on any set can be defined:

IX.2.1.12. DEFINITION. Let f be a real-valued function on a set X, and $S \subseteq X$. If f is bounded on S, the number

$$\omega(f,S) = \sup_S f - \inf_S f$$

is called the *oscillation* of f on S.

IX.2.1.13. More generally, if f is a function from a set X to a metric space (Y,ρ), and f is bounded on a subset S of X, then we can define

$$\omega(f,S) = \sup\{\rho(f(s),f(t)) : s,t \in S\} \ .$$

If X is a topological space and $x_0 \in X$, then f is continuous at x_0 if and only if

$$\lim \omega(f,\mathcal{U}) = 0$$

as $\mathcal{U}$ ranges over the directed set of neighborhoods of x_0. In particular, if $X \subseteq \mathbb{R}$, then $f : X \to \mathbb{R}$ is continuous at $x_0 \in X$ if and only if

$$\lim_{\epsilon \to 0} \omega(f, X \cap (x-\epsilon, x+\epsilon)) = 0 \ .$$

IX.2.1.14. The connection with integration is that if f is a bounded function on an interval $[a, b]$, and $\mathcal{P}$ is a partition of $[a, b]$, with M_k and m_k defined as in IX.2.1.4., then

$$\omega(f, J_k) = M_k - m_k$$

and so we have

$$\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) = \sum_{k=1}^{n} \omega(f, J_k)\ell(J_k) \ .$$

The last number is called the *weighted local oscillation* of f with respect to $\mathcal{P}$.

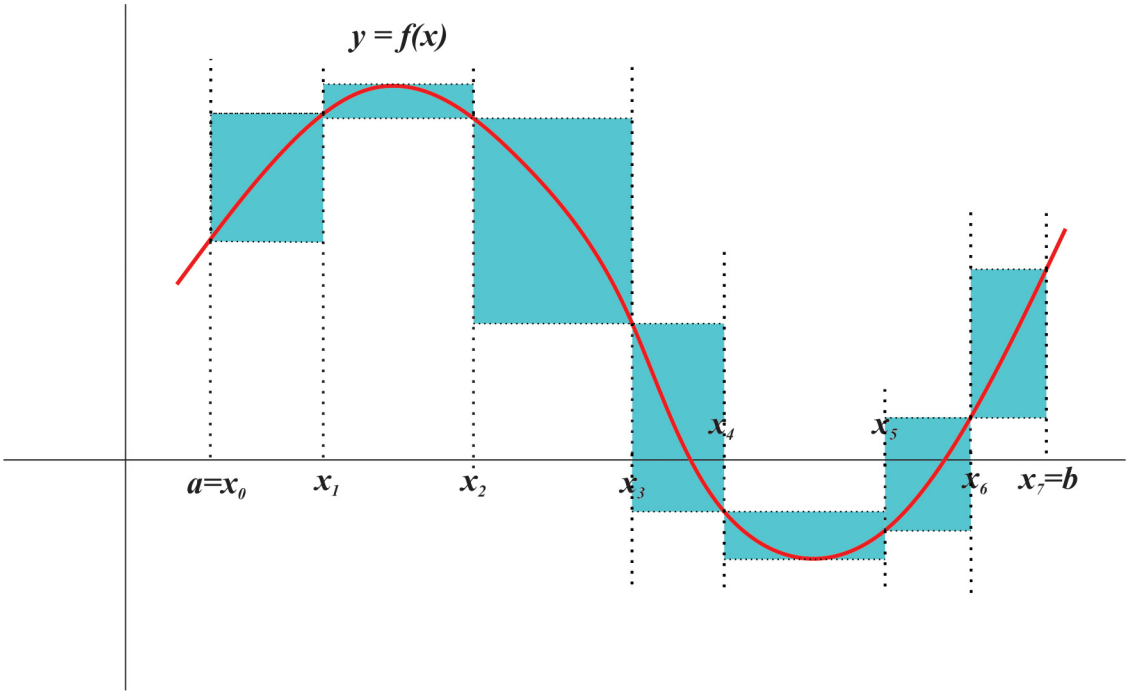


Figure IX.14: The Weighted Local Oscillation for a Partition

IX.2.1.15. The weighted local oscillation does not increase, and typically decreases, as the partition becomes finer (Figure IX.15).

IX.2.2. Step Functions and their Integrals

IX.2.3. Definitions of the Riemann Integral

We will give five definitions of Riemann integrability. All have slight rephrasings which are easily seen to be equivalent to the basic definition. There are other definitions which can be found in the literature, but the ones we give are the "extreme" ones and others can be seen to fit between our definitions.

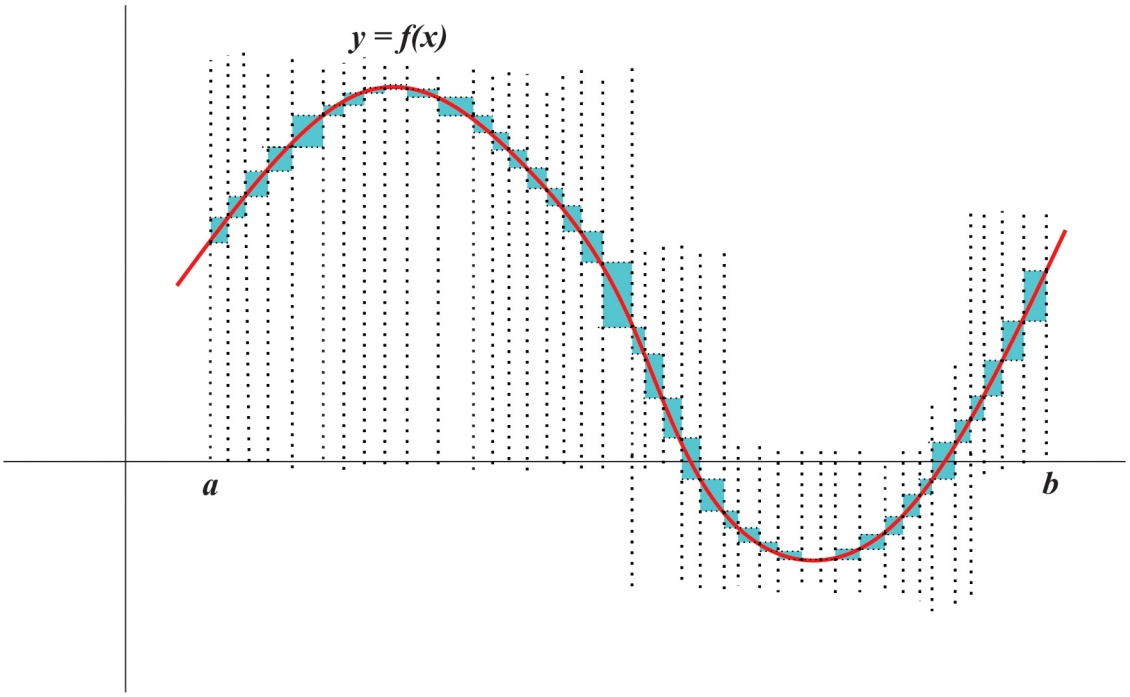


Figure IX.15: The Weighted Local Oscillation for a Finer Partition

The basic definitions are in terms of (1) Riemann sums, (2) upper and lower Darboux sums (or oscillation), or (3) approximation above and below by step functions. The other variation is (A) requiring existence of one partition forcing good approximation or (B) requiring good approximation for all sufficiently fine partitions.

Definition 1

IX.2.3.1. DEFINITION. [DEFINITION 1A] Let f be a function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J in sense 1A, with Riemann integral $R \in \mathbb{R}$, if for every $\epsilon > 0$ there is a partition $\mathcal{P} = \{x_0, \ldots, x_n\}$ of J such that, for every choice of tags $t_1, \ldots, t_n$ for $\mathcal{P}$, we have

$$\left| R - \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}) \right| < \epsilon \ .$$

IX.2.3.2. DEFINITION. [DEFINITION 1B] Let f be a function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J in sense 1B, with Riemann integral $R \in \mathbb{R}$, if for every $\epsilon > 0$ there is a $\delta > 0$ such that, for every tagged partition $\dot{\mathcal{P}} = \{x_0, \ldots, x_n; t_1, \ldots, t_n\}$ of J with $\|\mathcal{P}\| < \delta$, we have

$$\left| R - \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1}) \right| < \epsilon \ .$$

Definition 2

IX.2.3.3. DEFINITION. [DEFINITION 2A] Let f be a bounded function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J in sense 2A if for every $\epsilon > 0$ there is a partition $\mathcal{P}$ of J such that

$$\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) < \epsilon \ .$$

IX.2.3.4. DEFINITION. [DEFINITION 2B] Let f be a bounded function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J in sense 2B if for every $\epsilon > 0$ there is a $\delta > 0$ such that, for every partition $\mathcal{P}$ of J with $\|\mathcal{P}\| < \delta$, we have

$$\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) < \epsilon \ .$$

Definitions 2A and 2B can be rephrased in terms of "weighted local oscillations":

IX.2.3.5. DEFINITION. [DEFINITION 2A′] Let f be a bounded function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J in sense 2A′ if for every $\epsilon > 0$ there is a partition $\mathcal{P}$ of J such that

$$\sum_{k=1}^{n} \omega(f, J_k)\ell(J_k) < \epsilon$$

where J_k is the k'th subinterval of $\mathcal{P}$.

IX.2.3.6. DEFINITION. [DEFINITION 2B′] Let f be a bounded function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J in sense 2B′ if for every $\epsilon > 0$ there is a $\delta > 0$ such that, for every partition $\mathcal{P}$ of J with $\|\mathcal{P}\| < \delta$, we have

$$\sum_{k=1}^{n} \omega(f, J_k)\ell(J_k) < \epsilon$$

where J_k is the k'th subinterval of $\mathcal{P}$.

The fact that Definitions 2A′ and 2B′ are equivalent to 2A and 2B respectively is obvious.

There is a less obvious rephrasing of Definition 2A using interior Darboux sums (IX.2.1.7.) which also turns out to be equivalent and technically useful:

IX.2.3.7. DEFINITION. [DEFINITION 2A″] Let f be a bounded function on a closed bounded interval $J = [a, b]$. Then f is Riemann-integrable on J in sense 2A″ if for every $\epsilon > 0$ there is a partition $\mathcal{P}$ of J such that

$$\overline{\mathcal{S}}_\circ(f, \mathcal{P}) - \underline{\mathcal{S}}_\circ(f, \mathcal{P}) < \epsilon \ .$$

IX.2.3.8. Since for any f and any partition $\mathcal{P}$ we have

$$\underline{\mathcal{S}}(f,\mathcal{P}) \le \underline{\mathcal{S}}_\circ(f,\mathcal{P}) \le \overline{\mathcal{S}}_\circ(f,\mathcal{P}) \le \overline{\mathcal{S}}(f,\mathcal{P})$$

it follows that if f is Riemann-integrable in sense 2A, then it is Riemann-integrable in sense 2A$''$. The converse implication follows easily from IX.2.1.11., so we have:

IX.2.3.9. PROPOSITION. Let f be a bounded function on a closed bounded interval $[a,b]$. Then f is Riemann-integrable in sense 2A$''$ if and only if f is Riemann-integrable in sense 2A.

IX.2.3.10. We could also give an obvious similar version 2B$''$ of 2B. However, although 2B $\Rightarrow$ 2B$''$ is obvious, it is difficult to prove directly that 2B$''$ $\Rightarrow$ 2B. After the fact it will be seen to be true since we will have 2B$''$ $\Rightarrow$ 2A$''$ $\Rightarrow$ 2A $\Rightarrow$ 2B.

Definition 3

IX.2.3.11. DEFINITION. [DEFINITION 3A] Let f be a function on a closed bounded interval $J=[a,b]$. Then f is Riemann-integrable on J in sense 3A if for every $\epsilon>0$ there are step functions ϕ and ψ on J such that $\phi \le f \le \psi$ on J and

$$\int_a^b (\psi-\phi) = \int_a^b \psi - \int_a^b \phi < \epsilon\ .$$

IX.2.3.12. One could give a corresponding Definition 3B, but the statement is awkward and does not appear in the literature to my knowledge. But there is a commonly used rephrasing of Definition 3A, which is fairly obviously equivalent to 3A. First note that since a step function is bounded, any function satisfying Definition 3A is automatically bounded. Then make the following definition:

IX.2.3.13. DEFINITION. Let f be a bounded function on a closed bounded interval $J=[a,b]$. Define

$$\underline{\int_a^b} f(x)\,dx = \sup\left\{\int_a^b \phi : \phi \text{ step},\ \phi \le f\right\}$$

$$\overline{\int_a^b} f(x)\,dx = \inf\left\{\int_a^b \psi : \psi \text{ step},\ f \le \psi\right\}\ .$$

IX.2.3.14. DEFINITION. [DEFINITION 3A$'$] Let f be a function on a closed bounded interval $J=[a,b]$. Then f is Riemann-integrable on J in sense 3A$'$ if

$$\underline{\int_a^b} f(x)\,dx = \overline{\int_a^b} f(x)\,dx\ .$$

The common value is called the *Riemann integral* of f from a to b.

IX.2.3.15. To see that Definitions 3A and 3A′ are equivalent, note first that if f is any bounded function on J and ϕ, ψ are step functions with $\phi \le f \le \psi$, then $\int_a^b \phi \le \int_a^b \psi$, and thus by () we have

$$\int_a^b \phi \le \underline{\int_a^b} f(x)\,dx \le \bar{\int_a^b} f(x)\,dx \le \int_a^b \psi\ .$$

Thus, if f is Riemann integrable in sense 3A, then for any $\epsilon > 0$ we have

$$0 \le \bar{\int_a^b} f(x)\,dx - \underline{\int_a^b} f(x)\,dx < \epsilon$$

so f is Riemann integrable in sense 3A′. Conversely, if f is Riemann integrable in sense 3A′, and $\epsilon > 0$, choose step functions ϕ and ψ on J with $\phi \le f \le \psi$ and

$$\int_a^b \phi > \underline{\int_a^b} f(x)\,dx - \frac{\epsilon}{2} \quad \text{and} \quad \int_a^b \psi < \bar{\int_a^b} f(x)\,dx + \frac{\epsilon}{2}\ .$$

Then

$$\int_a^b \psi - \int_a^b \phi < \epsilon$$

so f is Riemann integrable in sense 3A.

One useful technical fact is the invariance of all the definitions under adding constants:

IX.2.3.16. Proposition. Let f be a function on a closed bounded interval $J = [a, b]$, let c be a constant, and $g = f + c$. If f is Riemann-integrable in any of the senses, then g is also Riemann-integrable in the same sense, and the Riemann integral of g is the Riemann integral of f plus $c(b-a)$. Riemann and Darboux sums for g are also the same as for f, plus $c(b-a)$.

Using this proposition, the theory of Riemann integration can be reduced to the case of nonnegative functions, since any function which is bounded below can be converted to a nonnegative function by adding a constant.

IX.2.4. Comparing the Definitions

The main result is:

IX.2.4.1. Theorem. Let f be a function on a closed bounded interval $J = [a, b]$. Then

(i) If f is Riemann-integrable on J in sense 1A or 1B, then f is bounded on J.

(ii) If f is Riemann-integrable in sense 2A or 2B, there is a unique real number I such that

$$\underline{\mathcal{S}}(f, \mathcal{P}) \le I \le \overline{\mathcal{S}}(f, \mathcal{P})$$

for every partition $\mathcal{P}$ of J.

(iii) If f is Riemann-integrable on J in any of the senses 1A, 1B, 2A, 2B, 2A′, 2B′, 3A, or 3A′, it is Riemann-integrable on J in all these senses, and the Riemann integral of f over J is well defined. In particular, the number R in definitions 1A and 1B is unique, and coincides with the number I of (ii) and with $\int_a^b f(x)\,dx$ of 3A′.

IX.2.4.2. A few implications are obvious. If f is Riemann-integrable in sense 1B, it is clearly Riemann-integrable in sense 1A. Similarly, if f is Riemann-integrable in sense 2B, it is also Riemann-integrable in sense 2A. And the obvious equivalence of 2A and 2A′, and of 2B and 2B′, has already been noted, as was the slightly less obvious equivalence between 3A and 3A′.

Thus, to prove (i), it suffices to consider an f which is Riemann-integrable in sense 1A.

IX.2.4.3. LEMMA. Suppose f is Riemann-integrable on J in sense 1A. Then f is bounded on J.

PROOF: Let R be the Riemann integral of f, and let $\mathcal{P} = \{J_1, \ldots, J_n\}$ be a partition such that for any choice of tags $t_k \in J_k$, we have

$$\left| R - \sum_{k=1}^{n} f(t_k)\ell(J_k) \right| < 1 \ .$$

If f is not bounded above, then there is a k_0 such that f is not bounded above on J_{k_0}. Let $t_1, \ldots, t_n$ be a choice of tags, and choose $s_{k_0} \in J_{k_0}$ with $f(s_{k_0}) > f(t_{k_0}) + \frac{2}{\ell(J_{k_0})}$. Set $s_k = t_k$ for $k \neq k_0$. We then have

$$\left| R - \sum_{k=1}^{n} f(t_k)\ell(J_k) \right| < 1$$

$$\left| R - \sum_{k=1}^{n} f(s_k)\ell(J_k) \right| < 1$$

so we have

$$[f(s_{k_0}) - f(t_{k_0})]\ell(J_{k_0}) = \left| \sum_{k=1}^{n} f(s_k)\ell(J_k) - \sum_{k=1}^{n} f(t_k)\ell(J_k) \right|$$

$$\leq \left| \sum_{k=1}^{n} f(s_k)\ell(J_k) - R \right| + \left| R - \sum_{k=1}^{n} f(t_k)\ell(J_k) \right| < 2$$

which is a contradiction. The proof that f is bounded below is almost identical. 👍

Next we compare definitions 1A and 2A, and 1B and 2B.

IX.2.4.4. LEMMA. Suppose f is a bounded function on J, and $\mathcal{P} = \{J_1, \ldots, J_n\}$ is a partition of J. Then $\underline{\mathcal{S}}(f, \mathcal{P})$ and $\overline{\mathcal{S}}(f, \mathcal{P})$ are the infimum and supremum, respectively, of the Riemann sums corresponding to all choices of tags for $\mathcal{P}$.

PROOF: It is clear that if $t_1, \ldots, t_n$ is any choice of tags for $\mathcal{P}$, then

$$\underline{\mathcal{S}}(f, \mathcal{P}) \leq \sum_{k=1}^{n} f(t_k)\ell(J_k) \leq \overline{\mathcal{S}}(f, \mathcal{P}) \ .$$

Fix $\epsilon > 0$. For each k, choose $t_k \in J_k$ with $f(t_k) > M_k - \frac{\epsilon}{\ell(J)}$. Then

$$\sum_{k=1}^{n} f(t_k)\ell(J_k) > \sum_{k=1}^{n} M_k\ell(J_k) - \frac{\epsilon}{\ell(J)} \sum_{k=1}^{n} \ell(J_k) = \overline{\mathcal{S}}(f, \mathcal{P}) - \epsilon \ .$$

The argument for the infimum is almost identical. 👍

IX.2.4.5. Now let f be a (necessarily bounded) function which is Riemann-integrable on J in sense 1A with Riemann integral R, $\epsilon > 0$, and $\mathcal{P} = \{J_1, \ldots, J_k\}$ a partition of J such that

$$\left| R - \sum_{k=1}^{n} f(t_k)\ell(J_k) \right| < \frac{\epsilon}{3}$$

for any choice of tags $t_k \in J_k$. Then

$$\left|R - \overline{\mathcal{S}}(f, \mathcal{P})\right| \leq \frac{\epsilon}{3} \quad \text{and} \quad |R - \underline{\mathcal{S}}(f, \mathcal{P})| \leq \frac{\epsilon}{3}$$

so we have

$$\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) \leq |\overline{\mathcal{S}}(f, \mathcal{P}) - R| + |R - \underline{\mathcal{S}}(f, \mathcal{P})| \leq \frac{2\epsilon}{3} < \epsilon$$

and so f is Riemann-integrable in sense 2A. An almost identical argument shows that if f is Riemann-integrable in sense 1B, then it is also Riemann-integrable in sense 2B.

For the other direction, note that there is no explicitly given value for the Riemann integral in definitions 2A or 2B, or in 2A′ or 2B′ (or in 3A, but there is one in the equivalent definition 3A′.) The essential point of (ii) is to provide a value for the Riemann integral in case 2A or 2B. To prove (ii), we first make an important observation:

IX.2.4.6. LEMMA. If f is a bounded function on J, and $\mathcal{P}$ and $\mathcal{Q}$ are any partitions of J, then

$$\underline{\mathcal{S}}(f, \mathcal{P}) \leq \overline{\mathcal{S}}(f, \mathcal{Q}) \ .$$

PROOF: The key is that there is a partition $\mathcal{R}$ which is a refinement of both $\mathcal{P}$ and $\mathcal{Q}$ (IX.2.1.2.). Then, by IX.2.1.6.,

$$\underline{\mathcal{S}}(f, \mathcal{P}) \leq \underline{\mathcal{S}}(f, \mathcal{R}) \leq \overline{\mathcal{S}}(f, \mathcal{R}) \leq \overline{\mathcal{S}}(f, \mathcal{Q}) \ .$$

👍

We then conclude from () that

IX.2.4.7. COROLLARY. Let f be a bounded function on J. Then

$$\sup_{\mathcal{P}}\{\underline{\mathcal{S}}(f,\mathcal{P})\} \le \inf_{\mathcal{P}}\{\overline{\mathcal{S}}(f,\mathcal{P})\}$$

where the supremum and infimum are taken over all partitions of J.

IX.2.4.8. Thus, if f is Riemann-integrable in sense 2A, then we have

$$\sup_{\mathcal{P}}\{\underline{\mathcal{S}}(f,\mathcal{P})\} = \inf_{\mathcal{P}}\{\overline{\mathcal{S}}(f,\mathcal{P})\}\ .$$

Call this number I. Then we have

$$\underline{\mathcal{S}}(f,\mathcal{P}) \le I \le \overline{\mathcal{S}}(f,\mathcal{P})$$

for any partition $\mathcal{P}$, and I is clearly the only number with this property, proving (ii). Also, if $\epsilon > 0$ and $\mathcal{P} = \{J_1,\dots,J_n\}$ is a partition with $\overline{\mathcal{S}}(f,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P}) < \epsilon$, then for any choice of tags $t_1,\dots,t_n$ with $t_k \in J_k$ we have

$$\left| I - \sum_{k=1}^{n} f(t_k)\ell(J_k) \right| < \epsilon$$

so f is Riemann-integrable in sense 1A with Riemann integral I. If f is Riemann-integrable in sense 2B, it follows almost identically that f is Riemann-integrable in sense 1B.

IX.2.4.9. Now let us compare definitions 2A and 3A. Let f be a bounded function on J. If $\mathcal{P} = \{J_1,\dots,J_n\}$ is a partition of J, and M_k and m_k the supremum and infimum of f on J_k as usual, define step functions ϕ and ψ by setting $\phi(x) = m_k$ and $\psi(x) = M_k$ for $x \in J_k$. If $x = x_k$ is the common endpoint of J_k and J_{k+1}, set $\phi(x)$ to be the smaller of m_k and m_{k+1}, and $\psi(x)$ the larger of M_k and M_{k+1}. Then ϕ and ψ are step functions, $\phi \le f \le \psi$, and $\int_a^b \phi = \underline{\mathcal{S}}(f,\mathcal{P})$, $\int_a^b \psi = \overline{\mathcal{S}}(f,\mathcal{P})$.

IX.2.4.10. So if f is Riemann-integrable in sense 2A, and $\epsilon > 0$, let $\mathcal{P}$ be a partition for which $\overline{\mathcal{S}}(f,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P}) < \epsilon$. Define ϕ and ψ as above for this $\mathcal{P}$; then

$$\int_a^b \psi - \int_a^b \phi = \overline{\mathcal{S}}(f,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P}) < \epsilon$$

so f is Riemann-integrable in sense 3A. We also conclude that the number I from (ii) is exactly $\int_a^b f(x)\,dx$.

IX.2.4.11. Going from 3A to 2A is equally straightforward except for an annoying technicality. Suppose f is a bounded function on $J = [a,b]$, which we may assume to be nonnegative by IX.2.3.16.. We want to show that

$$\inf_{\mathcal{P}}\{\overline{\mathcal{S}}(f,\mathcal{P})\} = \inf\left\{ \int_a^b \psi : \psi \text{ step}, f \le \psi \right\}$$

and similarly for lower estimates. (We have $\ge$ by the previous argument.) Let ψ be a step function for which $f \le \psi$, and $\mathcal{P}$ the partition consisting of $x_0 = a$, $x_n = b$, and the jump discontinuities $x_1,\dots,x_{n-1}$

of ψ in (a,b). Let c_k be the value of ψ on (x_{k-1},x_k), and $d_k = f(x_k)$. If M_k is the supremum of f on $J_k = [x_{k-1},x_k]$, we have $M_k \le \max(c_k,d_k,d_{k-1})$. So if $d_{k-1},d_k \le c_k$ for all k, we have

$$\overline{\mathcal{S}}(f,\mathcal{P}) \le \int_a^b \psi$$

but this fails if $d_k > c_k$ or $d_{k-1} > c_k$ for some k. To get around this problem, use definition 2A″ instead: we have

$$\overline{\mathcal{S}}_\circ(f,\mathcal{P}) \le \int_a^b \psi$$

for any such ψ. We thus get the desired inequality

$$\inf_{\mathcal{P}}\{\overline{\mathcal{S}}(f,\mathcal{P})\} = \inf_{\mathcal{P}}\{\overline{\mathcal{S}}_\circ(f,\mathcal{P})\} \le \inf\left\{\int_a^b \psi : \psi \text{ step}, \ f \le \psi\right\}$$

where the first equality is from IX.2.1.11.. By a similar argument we obtain

$$\sup_{\mathcal{P}}\{\underline{\mathcal{S}}(f,\mathcal{P})\} \ge \sup\left\{\int_a^b \phi : \phi \text{ step}, \ \phi \le f\right\}$$

(the opposite inequality follows from the previous argument).

IX.2.4.12. Thus, if f is Riemann-integrable in sense 3A, it is Riemann-integrable in sense 2A, and the number I of (ii) is exactly $\int_a^b f(x)\,dx$.

IX.2.4.13. We have now shown that definitions 1A, 2A, and 3A (hence also 2A′ and 3A′) define the same class of Riemann-integrable functions, and the Riemann integrals coincide (using (ii) in case 2A). We will say that a function Riemann-integrable in any of these senses is Riemann-integrable in sense A. Similarly, definitions 1B and 2B (hence also 2B′) define the same class of Riemann-integrable functions, with a well-defined Riemann integral, called the functions Riemann-integrable in sense B. Also, a function Riemann-integrable in sense B is Riemann-integrable in sense A, and the Riemann integrals coincide. Thus, finish the proof of the theorem, we need to show that a function which is Riemann-integrable in sense A is Riemann-integrable in sense B. This is the least obvious part of the proof of the theorem.

IX.2.4.14. So suppose f is Riemann-integrable on J in sense A, and let $\epsilon > 0$. Let $\mathcal{P} = \{x_0,x_1,\dots,x_n\}$ be a partition satisfying

$$\overline{\mathcal{S}}(f,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P}) < \frac{\epsilon}{2}\,.$$

Let M be a bound for f on J, and set $g = f + M$, $h = f - M$. Then $0 \le g \le 2M$ and $-2M \le h \le 0$. Let $\delta < \frac{\epsilon}{4(n-1)M}$. If $\mathcal{Q} = \{I_1,\dots,I_m\}$ is a partition of J with $\|\mathcal{Q}\| < \delta$, then $\mathcal{Q}$ might not quite be a refinement of $\mathcal{P}$ since some or all of the x_k might be interior points of intervals of $\mathcal{Q}$. But there are at most $n-1$ of these intervals; let N_1 be the set of r such that I_r is one of these, and let N_2 be the rest of the indices. Then

$$\overline{\mathcal{S}}(g,\mathcal{Q}) = \sum_{r\in N_1} M_r\ell(I_r) + \sum_{r\in N_2} M_r\ell(I_r)$$

where M_r is the supremum of g on I_r. Since for each $r \in N_2$ we have that I_r is contained in one of the J_k, and the supremum of g over I_r is less than or equal to the supremum of g over the corresponding J_k, we have that

$$\sum_{r \in N_2} M_r \ell(I_r) \leq \overline{\mathcal{S}}(g, \mathcal{P}) \ .$$

Since each I_r has length $< \delta$, and N_1 contains at most $n-1$ indices, we have

$$\sum_{r \in N_1} M_r \ell(I_r) \leq 2(n-1)M\delta < \frac{\epsilon}{2} \ .$$

Thus we have

$$\overline{\mathcal{S}}(g, \mathcal{Q}) < \overline{\mathcal{S}}(g, \mathcal{P}) + \frac{\epsilon}{2} \ .$$

By translation (IX.2.3.16.), we thus have

$$\overline{\mathcal{S}}(f, \mathcal{Q}) < \overline{\mathcal{S}}(f, \mathcal{P}) + \frac{\epsilon}{2} \ .$$

Applying a similar argument to h using infima, we have

$$\underline{\mathcal{S}}(h, \mathcal{Q}) > \underline{\mathcal{S}}(h, \mathcal{P}) - \frac{\epsilon}{2}$$

and so by linearity we have

$$\underline{\mathcal{S}}(f, \mathcal{Q}) > \underline{\mathcal{S}}(f, \mathcal{P}) - \frac{\epsilon}{2} \ .$$

Combining the two estimates, we get

$$\overline{\mathcal{S}}(f, \mathcal{Q}) - \underline{\mathcal{S}}(f, \mathcal{Q}) < \epsilon$$

and so f is Riemann-integrable in sense 2B (i.e. in sense B).

IX.2.4.15. There is one last loose end. We must show that the number R in Definition 1A is unique. Let f be Riemann-integrable in sense 1A, and R a number satisfying Definition 1A. Fix $\epsilon > 0$, and let $\mathcal{P}$ be a partition such that

$$\left| R - \sum_{k=1}^{n} f(t_k)\ell(J_k) \right| < \epsilon$$

for any choice of tags. Then by IX.2.4.4., we have

$$|R - \underline{\mathcal{S}}(f, \mathcal{P})| \leq \epsilon$$

$$|R - \overline{\mathcal{S}}(f, \mathcal{P})| \leq \epsilon$$

and so, since the number I of (ii) is between $\underline{\mathcal{S}}(f, \mathcal{P})$ and $\overline{\mathcal{S}}(f, \mathcal{P})$, we have

$$|R - I| \leq \epsilon \ .$$

since $\epsilon > 0$ is arbitrary, we have $R = I$ and R is uniquely determined.

This completes the proof of IX.2.4.1..

In light of Theorem IX.2.4.1., we can make the definitive definition of Riemann-integrability:

IX.2.4.16. Definition. Let f be a bounded function on a closed bounded interval $J = [a, b]$. Then f is *Riemann-integrable* on J if it is Riemann-integrable in any of the senses 1A-B, 2A-B, 2A′-B′, or 3A-A′. The Riemann integral of f over J, or from a to b, denoted $\int_a^b f(x)\,dx$, is the number defined in IX.2.3.1.–IX.2.3.14..

Additivity of the Riemann Integral

A simple consequence of the equivalence of the definitions is:

IX.2.4.17. Proposition. Let f be a function on $[a, b]$, and let $\{\mathcal{P}_n\}$ be a sequence of partitions of $[a, b]$ with $\lim_{n\to\infty} \|\mathcal{P}_n\| = 0$. Then f is Riemann-integrable on $[a, b]$ if and only if

$$\lim_{n\to\infty} \left[\overline{\mathcal{S}}(f, \mathcal{P}_n) - \underline{\mathcal{S}}(f, \mathcal{P}_n)\right] = 0 \ .$$

If f is Riemann-integrable, then, for any taggings $\dot{\mathcal{P}}_n$ of the $\mathcal{P}_n$, we have

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \overline{\mathcal{S}}(f, \mathcal{P}_n) = \lim_{n\to\infty} \underline{\mathcal{S}}(f, \mathcal{P}_n) = \lim_{n\to\infty} \mathcal{R}(f, \dot{\mathcal{P}}_n) \ .$$

IX.2.4.18. This is useful in many ways. For example, it allows Riemann-integrability to be checked and the Riemann integral to be computed using only partitions of $[a, b]$ into subintervals of equal length $\frac{b-a}{n}$ for any sequence of n's going to infinity (e.g. powers of 2). Another important consequence is:

IX.2.4.19. Corollary. Let f be a function on $[a, b]$, and $\mathcal{P}_0$ a fixed partition of $[a, b]$. In checking Riemann-integrability of f on $[a, b]$ and in computing the integral, it suffices to consider only partitions of $[a, b]$ refining $\mathcal{P}_0$.

IX.2.4.20. Corollary. [Additivity of the Riemann Integral] Let f be a function on $[a, b]$, and $\mathcal{P}_0 = \{a = c_0, \ldots, c_n = b\}$ a partition of $[a, b]$. Then f is Riemann-integrable on $[a, b]$ if and only if it is Riemann-integrable on $[c_{k-1}, c_k]$ for $1 \le k \le n$, and

$$\int_a^b f(x)\,dx = \sum_{k=1}^{n} \int_{c_{k-1}}^{c_k} f(x)\,dx \ .$$

In particular, if $a < c < b$, then f is Riemann-integrable on $[a, b]$ if and only if it is Riemann-integrable on both $[a, c]$ and $[c, b]$, and

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx \ .$$

IX.2.4.21. This property of the Riemann integral is called "additivity" and is not the same as the property in IX.2.6.1.(i) which is called "linearity." (The two are related: additivity is the case of linearity using $f\chi_{[a,b]} = f\chi_{[a,c]} + f\chi_{(c,b]}$.)

IX.2.5. Integrals of Continuous and Monotone Functions

Let $\chi_{\mathbb{Q}}$ be the indicator function of $\mathbb{Q}$.

IX.2.6. Properties of the Riemann Integral

The following result summarizes some of the most important properties of the Riemann integral:

IX.2.6.1. THEOREM. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$, and f and g real-valued functions on $[a,b]$, and $c \in \mathbb{R}$. Then

(i) (*Linearity*) If f and g are Riemann-integrable on $[a,b]$, then $f+g$ and cf are Riemann-integrable on $[a,b]$, and

$$\int_a^b [f(x)+g(x)]\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx$$
$$\int_a^b (cf)(x)\,dx = c\int_a^b f(x)\,dx\ .$$

(ii) (*Monotonicity*) If f and g are Riemann-integrable on $[a,b]$, and if $f(x) \le g(x)$ for all $x \in [a,b]$, then

$$\int_a^b f(x)\,dx \le \int_a^b g(x)\,dx\ .$$

(iii) The constant function c is Riemann-integrable on $[a,b]$, and

$$\int_a^b c\,dx = c(b-a)\ .$$

PROOF: For (i), note that for any tagged partition $\dot{\mathcal{P}}$ of $[a,b]$, we have

$$\mathcal{R}(f+g,\dot{\mathcal{P}}) = \mathcal{R}(f,\dot{\mathcal{P}}) + \mathcal{R}(g,\dot{\mathcal{P}})$$
$$\mathcal{R}(cf,\dot{\mathcal{P}}) = c\mathcal{R}(f,\dot{\mathcal{P}})\ .$$

For (ii), if $f \le g$, for any tagged partition $\dot{\mathcal{P}}$ of $[a,b]$ we have

$$\mathcal{R}(f,\dot{\mathcal{P}}) \le \mathcal{R}(g,\dot{\mathcal{P}})\ .$$

For (iii), for any tagged partition $\dot{\mathcal{P}}$ of $[a,b]$ we have

$$\mathcal{R}(c,\dot{\mathcal{P}}) = c(b-a)\ .$$

Apply Definition 1. 👍

The following crude estimate is surprisingly useful:

IX.2.6.2. Corollary. Let f be Riemann-integrable on $[a,b]$. If $m \le f(x) \le M$ for all $x \in [a,b]$, then

$$m(b-a) \le \int_a^b f(x)\,dx \le M(b-a)\ .$$

In particular, if $|f(x)| \le M$ for all $x \in [a,b]$, then

$$\left| \int_a^b f(x)\,dx \right| \le M(b-a)\ .$$

This Corollary can be proved more easily by simply noting that if $\mathcal{P}$ is the partition of $[a,b]$ consisting of one interval, then

$$m(b-a) \le \underline{\mathcal{S}}(f,\mathcal{P}) \le \int_a^b f(x)\,dx \le \overline{\mathcal{S}}(f,\mathcal{P}) \le M(b-a)\ .$$

Here is a simple but important consequence:

IX.2.6.3. Corollary. Let f be a nonnegative continuous function on $[a,b]$. If

$$\int_a^b f(t)\,dt = 0$$

then f is identically 0 on $[a,b]$.

Proof: Suppose there is an $x_0 \in [a,b]$ with $f(x_0) > 0$. Then by continuity there is a subinterval $[c,d]$ of $[a,b]$ containing x_0 such that $f(x) \ge m := \frac{f(x_0)}{2} > 0$ for all $x \in [c,d]$. Then

$$\int_c^d f(t)\,dt \ge m(d-c)$$

$$0 = \int_a^b f(t)\,dt = \int_a^c f(t)\,dt + \int_c^d f(t)\,dt + \int_d^b f(t)\,dt \ge m(d-c) > 0$$

since the first and third integrals are nonnegative, a contradiction. 👍

See IX.2.7.9. and XVIII.1.3.11. for generalizations.

Another closely related inequality, an integral version of the Triangle Inequality (), is also used regularly:

IX.2.6.4. Corollary. Let f be Riemann-integrable on $[a,b]$. Then $|f|$ is also Riemann-integrable on $[a,b]$, and

$$\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\,dx\ .$$

Proof: We only need to see that $|f|$ is Riemann-integrable. But if $\mathcal{P}$ is any partition of $[a,b]$, then

$$\overline{\mathcal{S}}(|f|,\mathcal{P}) - \underline{\mathcal{S}}(|f|,\mathcal{P}) \leq \overline{\mathcal{S}}(f,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P})$$

since $\omega(|f|,I) \leq \omega(f,I)$ for any interval I. 👍

IX.2.6.5. In the context of Figure IX.9, this is really a triangle inequality: if A_1 and A_3 are the areas of the two blue regions and A_2 the area of the tan region, then $\left|\int_a^b f(x)\,dx\right| = |A_1 - A_2 + A_3|$ and $\int_a^b |f(x)|\,dx = A_1 + A_2 + A_3$.

In connection with this, a product of Riemann integrable functions is Riemann integrable. This follows immediately from the characterization of (), but there is a slick elementary proof.

IX.2.6.6. Lemma. Let f be a bounded function on a set I, with $|f| \leq M$ on I. Then the oscillation

$$\omega(f^2,I) \leq 2M\omega(f,I)\ .$$

Proof: If $s,t \in I$, then

$$|[f(s)]^2 - [f(t)]^2| = |f(s)+f(t)||f(s)-f(t)| \leq 2M|f(s)-f(t)|\ .$$

👍

IX.2.6.7. Lemma. Let f be a Riemann integrable function on an interval $[a,b]$. Then f^2 is also Riemann integrable on $[a,b]$.

Proof: Since f is Riemann integrable on $[a,b]$, it is bounded. Fix M with $|f| \leq M$ on $[a,b]$. If $\mathcal{P}$ is any partition of $[a,b]$, then by IX.2.6.6.

$$\overline{\mathcal{S}}(f^2,\mathcal{P}) - \underline{\mathcal{S}}(f^2,\mathcal{P}) \leq 2M|\overline{\mathcal{S}}(f,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P})|\ .$$

👍

IX.2.6.8. PROPOSITION. Let f and g be Riemann integrable on $[a,b]$. Then fg is Riemann integrable on $[a,b]$.

PROOF: We have that $f+g$ is Riemann integrable on $[a,b]$. Hence f^2, g^2, and $(f+g)^2$ are Riemann integrable on $[a,b]$ by IX.2.6.7., and

$$fg = \frac{1}{2}\left[(f+g)^2 - f^2 - g^2\right] \ .$$



The Mean Value Theorem for Integrals

Here is a version of the Mean Value Theorem for integrals. It is an immediate consequence of the FTC1 and the ordinary MVT, but it has a simple direct proof from the Intermediate Value Theorem.

IX.2.6.9. THEOREM. [MEAN VALUE THEOREM FOR INTEGRALS, VERSION 1] Let f be a continuous function on $[a,b]$. Then f is Riemann integrable on $[a,b]$, and there is a $c \in [a,b]$ for which

$$f(c)(b-a) = \int_a^b f(t)\,dt \ .$$

PROOF: By (), f is Riemann integrable on $[a,b]$. If M and m are the maximum and minimum values f takes on $[a,b]$ (), then

$$m(b-a) \leq \int_a^b f(t)\,dt \leq M(b-a)$$

(), and so by the Intermediate Value Theorem () there is a $c \in [a,b]$ with

$$f(c) = \frac{1}{b-a}\int_a^b f(t)\,dt \ .$$



Here is a generalized version:

IX.2.6.10. THEOREM. [MEAN VALUE THEOREM FOR INTEGRALS, VERSION 2] Let f and g be continuous functions on $[a,b]$, with g nonnegative. Then f, g, and fg are Riemann integrable on $[a,b]$, and there is a $c \in [a,b]$ for which

$$f(c)\int_a^b g(t)\,dt = \int_a^b f(t)g(t)\,dt \ .$$

PROOF: If g is identically 0, then fg is also identically 0 and there is nothing to prove. If g is not identically 0, then $\int_a^b g(t)\,dt > 0$ (IX.2.6.3.). Then, if M and m are the maximum and minimum values of f on $[a,b]$, we have $mg \le fg \le Mg$ on $[a,b]$, so

$$m\int_a^b g(t)\,dt \le \int_a^b f(t)g(t)\,dt \le M\int_a^b g(t)\,dt$$

and thus by the Intermediate Value Theorem there is a $c \in [a,b]$ with

$$f(c) = \frac{\int_a^b f(t)g(t)\,dt}{\int_a^b g(t)\,dt}\ .$$



See IX.2.7.14. and XIX.1.3.2. for more general versions.

Continuity of the Integral

One must be quite careful about interchanging integrals and limits: the Riemann integral is not continuous for all types of limits. In fact, obtaining good limit theorems was one of the main goals of developing modern integration theories.

There is one simple but important situation where Riemann integration is continuous:

IX.2.6.11. PROPOSITION. Let $[a,b]$ be a closed bounded interval, and (f_n) a sequence of functions on $[a,b]$ converging uniformly on $[a,b]$ to a function f. If each f_n is Riemann-integrable on $[a,b]$, then f is Riemann-integrable on $[a,b]$ and

$$\int_a^b f_n(x)\,dx \to \int_a^b f(x)\,dx\ .$$

(Informally, the integral of a uniform limit is the limit of the integrals.)

PROOF: If we know f is Riemann-integrable (e.g. if the f_n are continuous), this follows immediately from IX.2.6.1., IX.2.6.2., and IX.2.6.4.: if $\epsilon > 0$, choose N so that $|f_n(x) - f(x)| < \frac{\epsilon}{b-a}$ for all $n \ge N$ and all $x \in [a,b]$; then

$$\left|\int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx\right| = \left|\int_a^b (f_n(x) - f(x))\,dx\right| \le \int_a^b |f_n(x) - f(x)|\,dx \le \frac{\epsilon}{b-a}(b-a) = \epsilon$$

for all $n \ge N$.

The argument that f is Riemann-integrable is very similar to the proof that a uniform limit of continuous functions is continuous (). Note that if $|f_n(x) - f(x)| < \epsilon$ for all $x \in [a,b]$, then for any partition $\mathcal{P}$ we have

$$|\underline{\mathcal{S}}(f_n,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P})| \le \epsilon(b-a)$$

$$|\overline{\mathcal{S}}(f_n,\mathcal{P}) - \overline{\mathcal{S}}(f,\mathcal{P})| \le \epsilon(b-a)$$

so, if $\epsilon > 0$, and N is such that

$$|f_n(x) - f(x)| < \frac{\epsilon}{3(b-a)}$$

for all $n \geq N$ and $x \in [a,b]$, and $\mathcal{P}$ is a partition such that

$$\overline{\mathcal{S}}(f_N, \mathcal{P}) - \underline{\mathcal{S}}(f_N, \mathcal{P}) < \frac{\epsilon}{3}$$

we have that

$$\begin{aligned}\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) &\leq |\overline{\mathcal{S}}(f, \mathcal{P}) - \overline{\mathcal{S}}(f_N, \mathcal{P})| + [\overline{\mathcal{S}}(f_N, \mathcal{P}) - \underline{\mathcal{S}}(f_N, \mathcal{P})] + |\underline{\mathcal{S}}(f_N, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P})| \\ &< \frac{\epsilon}{3(b-a)}(b-a) + \frac{\epsilon}{3} + \frac{\epsilon}{3(b-a)}(b-a) = \epsilon \ .\end{aligned}$$

Since $\epsilon > 0$ is arbitrary, f is Riemann-integrable by Definition 2A. 👍

Dense Sets and Integrability

IX.2.6.12. PROPOSITION. If f_1 and f_2 are Riemann-integrable on $[a,b]$ and agree on a dense subset D, then

$$\int_a^b f_1(x)\,dx = \int_a^b f_2(x)\,dx \ .$$

PROOF: If $\mathcal{P}$ is any partition of $[a,b]$, then Riemann sums for f_1 and f_2 with respect to $\mathcal{P}$ can be chosen with the tags in D, so these Riemann sums coincide. 👍

We can use this fact to give a definition of Riemann-integrability and the Riemann integral for certain functions which are not defined everywhere.

IX.2.6.13. DEFINITION. Let f be a real-valued function defined on a dense subset D of $[a,b]$. Then f is Riemann-integrable on $[a,b]$ if there is a Riemann-integrable function g on $[a,b]$ whose restriction to A is f. The Riemann integral $\int_a^b f(x)\,dx$ is defined to be $\int_a^b g(x)\,dx$.

IX.2.6.14. The Riemann integral of f is well defined independent of the choice of g by IX.2.6.12., and any two choices for g actually agree almost everywhere (IX.2.12.1.). An f which is Riemann-integrable in this extended sense is necessarily bounded.

The simplest situation is when f is defined except at finitely many points. Such a function can be changed, defined, or left undefined at finitely many points without changing its Riemann-integrability or its Riemann integral:

IX.2.6.15. PROPOSITION. Let f be a function defined at all but finitely many points of $[a,b]$. If f is Riemann-integrable, then any function g defined on all but finitely many points and agreeing with f at all but finitely many points is also Riemann-integrable, with

$$\int_a^b g(x)\,dx = \int_a^b f(x)\,dx\ .$$

PROOF: If $\mathcal{P}$ is a partition fine enough that it includes all the points where f or g is undefined and the points where they have unequal values, we have

$$\overline{\mathcal{S}}_\circ(f,\mathcal{P}) = \overline{\mathcal{S}}_\circ(g,\mathcal{P}) \text{ and } \underline{\mathcal{S}}_\circ(f,\mathcal{P}) = \underline{\mathcal{S}}_\circ(g,\mathcal{P})$$

and we can apply Definition 2A″. 👍

IX.2.6.16. Changing a function on even countably many points can affect its Riemann-integrability, e.g. changing the constant function 0 to $\chi_{\mathbb{Q}}$. (But if the new function is Riemann-integrable, the value of the integral is the same.)

IX.2.7. Null Sets and Almost Everywhere

We need to introduce here an important concept from measure theory. This discussion can serve as a preparatory introduction to measure theory.

IX.2.7.1. DEFINITION. Let A be a subset of $\mathbb{R}$. A has *measure zero*, or is a *null set*, if for every $\epsilon > 0$ there is a sequence (I_k) of finite intervals (which may be degenerate) such that $A \subseteq \cup_{k=1}^{\infty} I_k$ and $\sum_{k=1}^{\infty} \ell(I_k) < \epsilon$.
A has *zero content* if for each $\epsilon > 0$ there is a finite set $\{I_1, \ldots, I_n\}$ of finite intervals such that $A \subseteq \cup_{k=1}^{n} I_k$ and $\sum_{k=1}^{n} \ell(I_k) < \epsilon$.

IX.2.7.2. The term "null set" is often used as a synonym for "empty set", but in analysis it is more useful to use the definition we have given. (Note that the empty set is a null set by the analysis definition, but so are many nonempty sets, such as any finite set; in fact finite sets have zero content.)

Any subset of a null set (resp. a set of zero content) is a null set (resp. a set of zero content).

A set with zero content is obviously a null set (since $\emptyset$ is a degenerate interval); the converse is false:

IX.2.7.3. EXAMPLES. (i) Any countable set is a null set. [If A is countable, say $A = \{x_n : n \in \mathbb{N}\}$, and $\epsilon > 0$, let I_k be the open interval centered at x_k of length $\frac{\epsilon}{2^k}$.] Thus, for example, $\mathbb{Q}$ is a null set. But $\mathbb{Q}$ does not have zero content since a set with zero content is clearly bounded.

(ii) Even a bounded null set need not have zero content. For example, $A = \mathbb{Q} \cap [0, 1]$ is a null set by (i), but if $\{I_1, \ldots, I_n\}$ is a finite set of intervals containing A, then $\cup_{k=1}^{n} I_k$ contains all but finitely many points of $[0, 1]$. It follows that $\sum_{k=1}^{n} \ell(I_k) \geq 1$ (this seems obvious, but is actually somewhat tricky to prove; see (). This argument can be used to show that a true interval cannot be a null set, but this can be otherwise proved; see IX.2.7.12.. Thus the complement of a null set is dense.)

(iii) There are uncountable sets which have measure zero, e.g. the Cantor set K (). In fact, K has zero content.

In testing whether a set is a null set or has zero content, it suffices to use open intervals:

IX.2.7.4. PROPOSITION. Let A be a subset of $\mathbb{R}$. Then A is a null set if and only if there is a sequence (J_k) of bounded *open* intervals with $A \subseteq \cup_{k=1}^{\infty} J_k$ and $\sum_{k=1}^{\infty} \ell(J_k) < \epsilon$. A has zero content if and only if for every $\epsilon > 0$ there is a finite sequence (J_k) of bounded open intervals with the same properties.

PROOF: Suppose A has measure zero, and $\epsilon > 0$. There is a sequence (I_k) of bounded intervals such that $A \subseteq \cup_{k=1}^{\infty} I_k$ and $\sum_{k=1}^{\infty} \ell(I_k) < \frac{\epsilon}{2}$. For each k let J_k be the open interval with the same midpoint as I_k with $\ell(J_k) = 2\ell(I_k)$; then $I_k \subseteq J_k$, and the J_k work. The same argument works for finitely many intervals in the zero content case. The converse implication is trivial. 👍

IX.2.7.5. COROLLARY. A subset A has zero content if and only if it is bounded and its closure has measure zero. In particular, a closed bounded null set has zero content.

PROOF: Suppose A has zero content. Then A is bounded. Let $\epsilon > 0$, and let $\{I_1, \ldots, I_n\}$ be a finite set of bounded intervals with $A \subseteq \cup_{k=1}^n I_k$ and $\sum_{k=1}^n \ell(I_k) < \epsilon$. Replacing each I_k by its closure if necessary, we may assume each I_k is closed. But then $\cup_{k=1}^n I_k$ is closed, so $\bar{A} \subseteq \cup_{k=1}^n I_k$ and $\bar{A}$ is a null set (in fact, has zero content).

Conversely, suppose A is bounded and $\bar{A}$ is a null set. Fix $\epsilon > 0$, and let (I_k) be a sequence of bounded open intervals such that $\bar{A} \subseteq \cup_{k=1}^\infty I_k$ and $\sum_{k=1}^\infty \ell(I_k) < \epsilon$. The I_k cover $\bar{A}$, and since $\bar{A}$ is compact there is a finite subcover, i.e. $\bar{A} \subseteq \cup_{k=1}^n I_k$ for some n. Since $\sum_{k=1}^n \ell(I_k) < \epsilon$, it follows that $\bar{A}$, and hence A, has zero content. 👍

Unions of Null Sets

A finite union of null sets is a null set, and a finite union of sets of zero content has zero content. For null sets, we have a better result:

IX.2.7.6. PROPOSITION. A countable union of null sets is a null set.

PROOF: Let A_n be a null set for each n, and $A = \cup_{n=1}^\infty$. Fix $\epsilon > 0$. For each n, let $\{I_{n,k}\}$ be a sequence of bounded intervals such that $A_n \subseteq \cup_{k=1}^\infty I_{n,k}$ and $\sum_{k=1}^\infty \ell(I_{n,k}) < \frac{\epsilon}{2^n}$. Then the set $\{I_{n,k} : n, k \in \mathbb{N}\}$ is a countable set of bounded intervals with $A \subseteq \cup_{n,k} I_{n,k}$ and $\sum_{n,k} \ell(I_{n,k}) < \epsilon$. Since $\epsilon > 0$ is arbitrary, A is a null set. 👍

IX.2.7.7. This result gives another proof that every countable set is a null set, since it is a countable union of singletons.

It is not true that countable unions of sets of zero content have zero content (e.g. $\mathbb{Q}$). Closure of the null sets under countable unions indicates that the notion of null set is a more flexible one than that of a set of zero content. Good behavior under countable unions turns out to be necessary for a good theory of measure and advanced integration.

Almost Everywhere

IX.2.7.8. It is convenient to use the phrase *almost everywhere*, abbreviated *a.e.*, to mean "except on a null set." Thus, for example, if we say "the functions f and g agree almost everywhere," we mean "the set of all x for which $f(x) = g(x)$ does not hold (i.e. one or both of $f(x)$ and $g(x)$ is undefined, or both are defined but unequal) has measure zero."

Nonnegative Riemann-Integrable Functions

Using the language of null sets and almost everywhere, we can give an important fact about nonnegative Riemann-integrable functions:

IX.2.7.9. PROPOSITION. Let f be a nonnegative function on an interval $[a,b]$. If f is Riemann-integrable on $[a,b]$ and $\int_a^b f(x)\,dx = 0$, then $f = 0$ a.e. on $[a,b]$.

PROOF: For each n, let

$$A_n = \left\{ x \in [a,b] : f(x) \geq \frac{1}{n} \right\} .$$

We will show that each A_n has zero content, hence is a null set. This will suffice, since then

$$\{x \in [a,b] : f(x) \neq 0\} = \bigcup_{n=1}^{\infty} A_n$$

will be a null set by IX.2.7.6..

Fix n, and then fix $\epsilon > 0$. Let $\mathcal{P} = \{x_0, \ldots, x_m\}$ be a partition with $\overline{\mathcal{S}}(f,\mathcal{P}) < \frac{\epsilon}{n}$. Let K be the set of all k such that $A_n \cap [x_{k-1}, x_k]$ is nonempty. Then $A_n \subseteq \cup_{k\in K}[x_{k-1}, x_k]$. If M_k is the supremum of f on $[x_{k-1}, x_k]$, we have

$$\frac{1}{n}\sum_{k\in K}(x_k - x_{k-1}) \leq \sum_{k\in K} M_k(x_k - x_{k-1}) \leq \sum_{k=1}^{m} M_k(x_k - x_{k-1}) = \overline{\mathcal{S}}(f,\mathcal{P}) < \frac{\epsilon}{n}$$

and so $\sum_{k\in K}(x_k - x_{k-1}) < \epsilon$. Since $\epsilon > 0$ is arbitrary, A_n has zero content. 👍

IX.2.7.10. IX.2.6.12. gives a strong converse: if f is Riemann-integrable on $[a,b]$ and equal to zero on a dense subset, then $\int_a^b f(x)\,dx = 0$.

One direction of the next proposition is a corollary of the proof of IX.2.7.9.:

IX.2.7.11. PROPOSITION. Let A be a subset of $[a,b]$. Then A has zero content if and only if the indicator function χ_A is Riemann-integrable on $[a,b]$ and $\int_a^b \chi_A(x)\,dx = 0$.

IX.2.7.12. COROLLARY. No interval is a null set.

PROOF: First let $[a,b]$ be a closed bounded interval (with $a < b$). If $[a,b]$ is a null set, it has zero content by IX.2.7.5.. But $\int_a^b \chi_{[a,b]}(x)\,dx = b - a > 0$, so $[a,b]$ does not have zero content by IX.2.7.11..

Now let I be an arbitrary interval, and let $a, b \in I$, $a < b$. If I is a null set, then its subset $[a,b]$ is also a null set, a contradiction. 👍

Almost the same result as IX.2.6.15. holds if "finite set" is replaced by "set of zero content," but boundedness of the modified function is not automatic and must be assumed:

IX.2.7.13. PROPOSITION. Let f be a function defined on $[a,b]$ except on a set of zero content. If f is Riemann-integrable, then any bounded function g defined except on a set of zero content and agreeing with f except on a set of zero content is also Riemann-integrable, with

$$\int_a^b g(x)\,dx = \int_a^b f(x)\,dx\ .$$

PROOF: Let h be a Riemann-integrable function on $[a,b]$ agreeing with f where f is defined. Extend g in any manner to a bounded function defined everywhere on $[a,b]$; we will from now on assume g is defined everywhere.

Let A be a set of zero content such that f and g are defined and equal on $[a,b]\setminus A$, and let $M>0$ be a bound for both g and h on $[a,b]$. Fix $\epsilon>0$. Let $\{I_1,\dots,I_n\}$ be a set of bounded intervals with $A\subseteq\cup_{k=1}^n I_k$ and $\sum_{k=1}^n \ell(I_k)<\frac{\epsilon}{4M}$. Let $\mathcal{P}$ be a partition of $[a,b]$ such that $\overline{\mathcal{S}}_\circ(h,\mathcal{P})-\underline{\mathcal{S}}_\circ(h,\mathcal{P})<\frac{\epsilon}{2}$ and such that $\mathcal{P}$ includes all endpoints of $I_1,\dots,I_n$ which are in $[a,b]$. Let M_j and m_j be the supremum and infimum of g on (x_{j-1},x_j).

Let J be the set of all indices j such that $(x_{j-1},x_j)\subseteq\cup_{k=1}^n I_k$. Then

$$\overline{\mathcal{S}}_\circ(g,\mathcal{P})-\underline{\mathcal{S}}_\circ(g,\mathcal{P})=\sum_{j=1}^m (M_j-m_j)(x_j-x_{j-1})$$

$$=\sum_{j\in J}(M_j-m_j)(x_j-x_{j-1})+\sum_{j\notin J}(M_j-m_j)(x_j-x_{j-1})<2M\frac{\epsilon}{4M}+\frac{\epsilon}{2}=\epsilon\ .$$

Since $\epsilon>0$ is arbitrary, g is Riemann-integrable on $[a,b]$ by Definition 2A″, and

$$\int_a^b g(x)\,dx=\int_a^b h(x)\,dx\ .$$



The Mean Value Theorem Again

Using IX.2.7.9., we get a slight extension of the Mean Value Theorem for Integrals IX.2.6.10.:

IX.2.7.14. THEOREM. [MEAN VALUE THEOREM FOR INTEGRALS, VERSION 3] Let f be a continuous function on $[a,b]$, and g a nonnegative Riemann integrable function on $[a,b]$. Then fg is Riemann integrable on $[a,b]$, and there is a $c\in[a,b]$ for which

$$f(c)\int_a^b g(t)\,dt=\int_a^b f(t)g(t)\,dt\ .$$

PROOF: The product fg is Riemann integrable on $[a,b]$ by IX.2.6.8.. If $\int_a^b g(t)\,dt=0$, then g is 0 a.e., so fg is also 0 a.e. and there is nothing to prove. Otherwise, the proof of IX.2.6.10. works verbatim. 

Note that g need not be continuous, but f must be continuous for the proof to work, and in fact the result fails in general if f is not continuous (consider the case where f is the indicator function of a proper subinterval).

IX.2.8. The Upper and Lower Envelope of a Riemann-Integrable Function

We continue somewhat in the spirit of IX.2.3.11., in a sense carrying it to its logical conclusion.

IX.2.8.1. DEFINITION. Let f be a bounded function on an interval $[a,b]$. Define

$$\underline{f} = \sup\{\phi : [a,b] \to \mathbb{R} \ : \ \phi \text{ lower semicontinuous, } \phi \le f\}$$

$$\overline{f} = \inf\{\psi : [a,b] \to \mathbb{R} \ : \ \psi \text{ upper semicontinuous, } f \le \psi\} \ .$$

The functions $\overline{f}$ and $\underline{f}$ are called the *upper* and *lower envelopes* of f on $[a,b]$.

We have that $\overline{f}$ and $\underline{f}$ are bounded, and $\underline{f} \le f \le \overline{f}$. By (), $\overline{f}$ is upper semicontinuous and $\underline{f}$ is lower semicontinuous.

IX.2.8.2. PROPOSITION. Let f be a bounded function on $[a,b]$, and $\overline{f}$ and $\underline{f}$ its upper and lower envelopes. If J is any open subinterval of $[a,b]$, then

$$\sup_{x\in J} \overline{f}(x) = \sup_{x\in J} f(x) \quad \text{and} \quad \inf_{x\in J} \underline{f}(x) = \inf_{x\in J} f(x) \ .$$

PROOF: It is clear that $\sup_{x\in J} \overline{f}(x) \ge \sup_{x\in J} f(x)$. Conversely, let $M = \sup_{x\in J} f(x)$. Then there is an upper semicontinuous step function ψ on $[a,b]$, which is constant on J and two other subintervals of $[a,b]$, with $\psi = M$ on J and $f \le \psi$. Thus

$$\sup_{x\in J} \overline{f}(x) \le \sup_{x\in J} \psi(x) = M = \sup_{x\in J} f(x) \ .$$

The argument for $\underline{f}$ is nearly identical. 👍

IX.2.8.3. COROLLARY. Let f be a bounded function on $[a,b]$, and $\overline{f}$ and $\underline{f}$ its upper and lower envelopes. For each $x \in [a,b]$, we have

$$\overline{f}(x) = \max\{f(x), \limsup_{y\to x} f(y)\} \quad \text{and} \quad \underline{f}(x) = \min\{f(x), \liminf_{y\to x} f(y)\} \ .$$

IX.2.8.4. COROLLARY. Let f be a bounded function on $[a,b]$, and $\overline{f}$ and $\underline{f}$ its upper and lower envelopes. If $\mathcal{P}$ is any partition of $[a,b]$, then

$$\overline{\mathcal{S}}_\circ(\overline{f}, \mathcal{P}) = \overline{\mathcal{S}}_\circ(f, \mathcal{P}) \quad \text{and} \quad \underline{\mathcal{S}}_\circ(\underline{f}, \mathcal{P}) = \underline{\mathcal{S}}_\circ(f, \mathcal{P}) \ .$$

Thus we have

$$\overline{\int_a^b} \overline{f}(x)\,dx = \overline{\int_a^b} f(x)\,dx \quad \text{and} \quad \underline{\int_a^b} \underline{f}(x)\,dx = \underline{\int_a^b} f(x)\,dx \ .$$

Using these results, we can prove the fundamental characterization of Riemann integrability:

IX.2.8.5. THEOREM. Let f be a function on an interval $[a,b]$. Then f is Riemann-integrable on $[a,b]$ if and only if it is bounded, and continuous almost everywhere.

PROOF: We have already shown that a Riemann-integrable function must be bounded. Suppose f is Riemann-integrable on $[a,b]$. Let $\overline{f}$ and $\underline{f}$ be its upper and lower envelopes, and set $g = \overline{f} - \underline{f}$. Then $g \geq 0$. Fix $\epsilon > 0$, and choose a partition $\mathcal{P}$ such that $\overline{\mathcal{S}}_\circ(f,\mathcal{P}) - \underline{\mathcal{S}}_\circ(f,\mathcal{P}) < \epsilon$. Then we have

$$\overline{\mathcal{S}}_\circ(g,\mathcal{P}) = \overline{\mathcal{S}}_\circ(\overline{f},\mathcal{P}) - \underline{\mathcal{S}}_\circ(\underline{f},\mathcal{P}) = \overline{\mathcal{S}}_\circ(f,\mathcal{P}) - \underline{\mathcal{S}}_\circ(f,\mathcal{P}) < \epsilon$$

and, since g is nonnegative and hence $0 \leq \underline{\mathcal{S}}_\circ(g,\mathcal{P}) \leq \overline{\mathcal{S}}_\circ(g,\mathcal{P}) < \epsilon$, we have that g is Riemann-integrable on $[a,b]$ and $\int_a^b g(x)\,dx = 0$. Thus $g = 0$ a.e. (IX.2.7.9.), i.e. $\underline{f} = f = \overline{f}$ a.e. If x is such that $\underline{f}(x) = f(x) = \overline{f}(x)$, then f is both upper semicontinuous and lower semicontinuous at x, hence continuous at x (). (The conclusion also follows immediately from IX.2.8.3..)

Conversely, suppose f is bounded and continuous a.e. on $[a,b]$. Let $M > 0$ be a bound for f on $[a,b]$, and A the set of discontinuities of f in $[a,b]$. Fix $\epsilon > 0$. Let (I_k) be a sequence of open intervals with $A \subseteq \cup_{k=1}^\infty I_k$ and $\sum_{k=1}^\infty \ell(I_k) < \frac{\epsilon}{4M}$. For each $x \in [a,b]\setminus A$ there is an open interval J_x around x with $|f(y)-f(z)| < \frac{\epsilon}{2(b-a)}$ for all $y,z \in J_x$, since f is continuous at x. The collection $\{I_k : k \in N\} \cup \{J_x : x \in [a,b] \setminus A\}$ is an open cover of $[a,b]$. By compactness, there is a finite subcover, so there is a partition $\mathcal{P} = \{x_0,\ldots,x_n\}$ of $[a,b]$ such that each interval $[x_{j-1},x_j]$ is contained either in one of the I_k or one of the J_x. Let K be the set of j such that $[x_{j-1},x_j]$ is contained in one of the I_k. Then, if M_j and m_j are the supremum and infimum of f on $[x_{j-1},x_j]$, we have

$$\overline{\mathcal{S}}(f,\mathcal{P}) - \underline{\mathcal{S}}(f,\mathcal{P}) = \sum_{j=1}^n (M_j - m_j)(x_j - x_{j-1})$$

$$= \sum_{j\in K}(M_j - m_j)(x_j - x_{j-1}) + \sum_{j\notin K}(M_j - m_j)(x_j - x_{j-1}) < 2M\frac{\epsilon}{4M} + \frac{\epsilon}{2(b-a)}(b-a) = \epsilon\ .$$

Since $\epsilon > 0$ is arbitrary, f is Riemann-integrable on $[a,b]$. 👍

IX.2.9. Differentiation and the Fundamental Theorem of Calculus

In this section, we establish the fundamental results that integration and differentiation are inverse processes.

IX.2.9.1. Let f be a Riemann-integrable function on $[a, b]$. Then, for any $x \in [a, b]$, f is Riemann-integrable on $[a, x]$ (). For $a \leq x \leq b$, define

$$A(x) = \int_a^x f(t)\, dt \ .$$

$A(x)$ is called the *cumulative area* of f from a to x; this can be regarded as a function A from $[a, b]$ to $\mathbb{R}$, called the *cumulative area function* of f from a to b. (If f is nonnegative, $A(x)$ can be truly regarded as the "area" of the region below the graph of f and above the x-axis, between the vertical lines at a and x. Rigorously, this is backward: the area of this region can be *defined* to be this integral – see (). For a general f, $A(x)$ gives the "signed area" under the graph from a to x, with regions below the x-axis having negative "area.")

IX.2.9.2. Proposition. If f is a Riemann-integrable function on $[a, b]$, then its cumulative area function is Lipschitz () on $[a, b]$, and hence (uniformly) continuous on $[a, b]$.

Proof: Since f is Riemann-integrable on $[a, b]$, it is bounded; let M be a bound for $|f|$ on $[a, b]$. Then, for every $x, y \in [a, b]$, $x < y$, we have

$$|A(y) - A(x)| = \left| \int_a^y f(t)\, dt - \int_a^x f(t)\, dt \right| = \left| \int_x^y f(t)\, dt \right| \leq \int_x^y |f(t)|\, dt \leq M(y - x)$$

by (). 👍

The next theorem is the main result of this section and, as the name suggests, one of the most important theorems in elementary analysis:

IX.2.9.3. Theorem. [Fundamental Theorem of Calculus, Version 1] Let f be Riemann-integrable on $[a, b]$, and A its cumulative area function. If f is continuous at $x_0 \in [a, b]$, then A is differentiable at x_0 and $A'(x_0) = f(x_0)$.

Before giving the proof of this theorem, we give some simple consequences. The first is the same as V.8.5.9.:

IX.2.9.4. Corollary. If f is continuous on $[a, b]$, then there is a function F, unique up to addition of a constant, which is differentiable on $[a, b]$ and for which $F'(x) = f(x)$ for all $x \in [a, b]$.

The uniqueness of the antiderivative up to a constant follows immediately from V.8.2.5..

The next variation gives the common procedure for calculating the value of the integral:

IX.2.9.5. Corollary. If f is continuous on $[a,b]$, and F is any antiderivative for f on $[a,b]$, then

$$\int_a^b f(x)\,dx = F(b) - F(a)\ .$$

Proof: Let A be the cumulative area function of f. Then we have that A is an antiderivative for f by IX.2.9.3., and that

$$\int_a^b f(x)\,dx = A(b)$$

and $A(a) = 0$. Also, by IX.2.9.4., there is a constant C such that $F(x) = A(x) + C$ for all $x \in [a,b]$. We then have

$$F(b) - F(a) = [A(b) + C] - [A(a) + C] = A(b) - A(a) = A(b) = \int_a^b f(x)\,dx\ .$$



IX.2.9.6. Extensive and detailed applications of this version of the Fundamental Theorem of Calculus to evaluating definite integrals can be found in any Calculus text. Since this is not a Calculus text, we omit discussion of these applications.

IX.2.9.7. We now give the proof of IX.2.9.3..
Proof: We have, for $x \in [a,b]$, $x \neq x_0$,

$$A(x) - A(x_0) = \int_a^x f(t)\,dt - \int_a^{x_0} f(t)\,dt = \int_{x_0}^x f(t)\,dt$$

$$\frac{A(x) - A(x_0)}{x - x_0} = \frac{1}{x - x_0}\int_{x_0}^x f(t)\,dt\ .$$

We also have

$$f(x_0) = \frac{1}{x - x_0}\int_{x_0}^x f(x_0)\,dt$$

(since $f(x_0)$ is just a constant), and so

$$\frac{A(x) - A(x_0)}{x - x_0} - f(x_0) = \frac{1}{x - x_0}\int_{x_0}^x [f(t) - f(x_0)]\,dt\ .$$

Let $\epsilon > 0$. Since f is continuous at x_0, there is a $\delta > 0$ such that, whenever $x \in [a,b]$ and $|x - x_0| < \delta$, $|f(x) - f(x_0)| < \epsilon$. If $x \in [a,b]$ and $|x - x_0| < \delta$, then $|t - x_0| < \delta$ for all t between x_0 and x, and hence

$$\left|\int_{x_0}^x [f(t) - f(x_0)]\,dt\right| < \epsilon|x - x_0|$$

by (); so we have

$$\left|\frac{A(x) - A(x_0)}{x - x_0} - f(x_0)\right| = \frac{1}{|x - x_0|}\left|\int_{x_0}^x [f(t) - f(x_0)]\,dt\right| < \epsilon$$

and therefore

$$A'(x_0) = \lim_{x \to x_0} \frac{A(x) - A(x_0)}{x - x_0} = f(x_0) \ .$$



IX.2.9.8. EXAMPLES. If f is Riemann-integrable on $[a, b]$, A is its cumulative area function, and f is not continuous at $x_0 \in [a, b]$, then A may or may not be differentiable at x_0, and if it is we may or may not have $A'(x_0) = f(x_0)$:

(i) Let $f(x) = 0$ for $x < 0$ and $f(x) = 1$ for $x \geq 0$. Then f is Riemann-integrable on $[-1, 1]$, and its cumulative area function A is given by $A(x) = 0$ if $-1 \leq x \leq 0$ and $A(x) = x$ if $0 \leq x \leq 1$. A is not differentiable at 0.

(ii) Let $f(x) = 0$ for $x \neq 0$ and $f(0) = 1$. Then f is Riemann-integrable on $[-1, 1]$, and A is identically 0. We have that A is differentiable at 0, but $A'(0) \neq f(0)$.

(iii) [vRS82, p. 3-4] Let $f(x) = \sin\left(\frac{1}{x}\right)$ for $x \neq 0$ and $f(0) = 0$. Then f is bounded, and continuous except at 0, so f is Riemann-integrable on $[-1, 1]$ (or any other closed bounded interval). If A is its cumulative area function, we will show that A is differentiable at 0 and $A'(0) = 0 = f(0)$ even though f is discontinuous at 0. This can be shown directly by brute force, but we can get it more easily indirectly.

We first show that f has an antiderivative F. Let

$$G(x) = \begin{cases} x^2 \cos\left(\frac{1}{x}\right) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases} \ .$$

Then G is differentiable everywhere, and

$$G'(x) = g(x) = \left\{ \begin{array}{cc} 2x \cos\left(\frac{1}{x}\right) + \sin\left(\frac{1}{x}\right) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{array} \right\} = h(x) + f(x)$$

where

$$h(x) = \begin{cases} 2x \cos\left(\frac{1}{x}\right) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{cases} \ .$$

Since h is continuous, it has an antiderivative H (IX.2.9.4.). Then $f = g - h$ also has an antiderivative $F = G - H$.

We then have that $A - F$ is continuous everywhere, differentiable everywhere except possibly at 0, and $(A - F)'(x) = 0$ for all $x \neq 0$. Thus by V.8.4.1.(iii) $A - F$ is constant, hence differentiable also at 0; so A is differentiable at 0 and $A'(0) = F'(0) = 0 = f(0)$.

More dramatic versions of this example exist: see Exercise V.12.4.3.. For a related example, if f is a Lebesgue integrable function on $[a, b]$ and A is its cumulative area function (i.e. $A(x)$ is the Lebesgue integral of f from a to x), then A is continuous, differentiable almost everywhere, and $A' = f$ almost everywhere (XIX.5.2.15.). Such an f need not be continuous anywhere and need not be equal almost everywhere to a function which is Riemann integrable on any subinterval ().

IX.2.9.9. Version IX.2.9.3. of the Fundamental Theorem of Calculus roughly says that the derivative of the integral is the original function. What about the integral of the derivative? One problem is that the derivative of a differentiable function on an interval need not be Riemann-integrable:

IX.2.9.10. EXAMPLES. (i) Let F be the function

$$F(x) = \left\{ \begin{array}{cc} x^2 \sin\left(\frac{1}{x^2}\right) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{array} \right. .$$

Then F is differentiable everywhere, and

$$F'(x) = f(x) = \left\{ \begin{array}{cc} 2x \sin\left(\frac{1}{x^2}\right) - \frac{2}{x} \cos\left(\frac{1}{x^2}\right) & \text{if} \quad x \neq 0 \\ 0 & \text{if} \quad x = 0 \end{array} \right\}$$

which is unbounded near 0 and hence not Riemann-integrable in any interval containing 0 (in fact, it is not even Lebesgue integrable in any interval containing 0).

(ii) Let F be the function of (). Then F is differentiable on $[0, 1]$, and $F' = f$ is bounded there. Since f is 0 on a dense subset of $[0, 1]$, the Riemann integral of f on any subinterval, if it existed, would have to be 0 (IX.2.6.12.). Since F is not constant on any subinterval, it follows from IX.2.9.11. below that f is not Riemann-integrable on any subinterval. It even turns out that f is not equal almost everywhere to a Riemann-integrable function ().

We do, however, have the following theorem, which extends IX.2.9.5.:

IX.2.9.11. THEOREM. [FUNDAMENTAL THEOREM OF CALCULUS, VERSION 2] Let F be a function which is continuous on $[a, b]$ and differentiable except possibly at finitely many points, with derivative f. If f is Riemann-integrable on $[a, b]$, then for all $x \in [a, b]$ we have

$$F(x) = F(a) + \int_a^x f(t)\, dt$$

and so, in particular,

$$\int_a^b f(t)\, dt = F(b) - F(a) .$$

PROOF: It suffices to prove the second statement; the first statement then results from applying the second to the interval $[a, x]$.

Let $\epsilon > 0$, and let $\mathcal{P} = \{a = x_0, x_1, \ldots, x_n = b\}$ be a partition satisfying Definition 1A for this ϵ, and fine enough that $\mathcal{P}$ contains all points where F is not differentiable (such a partition exists by Definition 1B). By the Mean Value Theorem (), for each k choose $t_k \in (x_{k-1}, x_k)$ so that

$$F(x_k) - F(x_{k-1}) = f(t_k)(x_k - x_{k-1}) .$$

Then, by telescoping, we have

$$F(b) - F(a) = \sum_{k=1}^{n} (F(x_k) - F(x_{k-1})) = \sum_{k=1}^{n} f(t_k)(x_k - x_{k-1})$$

and so, since the right side is a Riemann sum for f with respect to $\mathcal{P}$, we have

$$\left|(F(b) - F(a)) - \int_a^b f(t)\,dt\right| < \epsilon\ .$$

Since $\epsilon > 0$ is arbitrary, the result follows. 👍

This result generalizes to the case in which f is only required to be Lebesgue integrable (XIX.5.2.22.).

The hypothesis can be relaxed to allow F to be nondifferentiable at countably many points (if it is continuous) (XIX.5.2.22.), but not just differentiable almost everywhere ().

More general versions of the Fundamental Theorem of Calculus will be discussed in XVIII.4.2.1., XIX.5.2.15., XIX.5.2.21., ...

The Average Value of a Function

IX.2.9.12. DEFINITION. Let f be Riemann-integrable on $J = [a, b]$. Then the *average value* of f on J is

$$av(f, J) = \frac{1}{b-a}\int_a^b f(x)\,dx\ .$$

(Note that this definition also makes sense if $b < a$, and gives the same value if a and b are interchanged.)

It is not hard to motivate this definition as a reasonable version of the average value of a function on an interval, e.g. by considering Riemann sums corresponding to a subdivision of the interval into a large number of subintervals of equal length. One interpretation which is reasonable motivation is that $av(f, J)$ is the value of a constant function on J which has the same integral over J as f.

IX.2.9.13. Using this notion of average value, we have an interesting relation between average and instantaneous rates of change. If f is differentiable on $[a, b]$ and f' is Riemann integrable on $[a, b]$, then by IX.2.9.11. we have

$$\frac{1}{b-a}\int_a^b f'(t)\,dt = \frac{f(b) - f(a)}{b-a}$$

i.e. the average rate of change of f over $[a, b]$ (the right-hand side) is equal to the average value of the instantaneous rate of change of f over $[a, b]$ (the left-hand side). This can be thought of as a variant of the Mean Value Theorem: If f is differentiable on $[a, b]$ and f' is Riemann integrable on $[a, b]$, the difference quotient of f over $[a, b]$ is the average value of f' over $[a, b]$ (the MVT asserts it is the value of f' at some particular point of $[a, b]$).

We then have the following analog of the Mean Value Theorem (actually more than an analog, in light of IX.2.9.3.).

IX.2.9.14. PROPOSITION. [MEAN VALUE THEOREM FOR INTEGRALS] Let f be a continuous function on a closed bounded interval $J = [a, b]$. Then there is at least one $c \in (a, b)$ with $f(c) = av(f, J)$.

PROOF: Let m and M be the minimum and maximum of f on $[a, b]$ (). Then, if $\mathcal{P} = \{a, b\}$ is the partition with one subinterval, we have $\underline{\mathcal{S}}(f, \mathcal{P}) = m(b - a)$ and $\overline{\mathcal{S}}(f, \mathcal{P}) = M(b - a)$, so we have

$$m(b - a) \leq \int_a^b f(x)\, dx \leq M(b - a)$$

so, dividing through by $b - a$, we have

$$m \leq av(f, J) \leq M \ .$$

If $m = M$, then f is constant on $[a, b]$, and we can take c to be any number in (a, b). If $m < M$, we have $m < av(f, J) < M$ [if $m = av(f, J)$, set $g(x) = f(x) - m$; then $g \geq 0$ is continuous and $\int_a^b g(x)\, dx = 0$, so g is identically 0 (), a contradiction. A similar argument using $h(x) = M - f(x)$ shows that we cannot have $av(f, J) = M$.] Choose $x_1, x_2 \in [a, b]$ with $f(x_1) = m$, $f(x_2) = M$ (). Then by the Intermediate Value Theorem () there is a c in the open interval between x_1 and x_2 with $f(c) = av(f, J)$. 👍

Note that this result can fail if f is not continuous, e.g. if f is a step function taking only the values 0 and 1.

Derivatives and Limits

We can use the FTC plus properties of the Riemann integral to give a simpler proof of a crucial special case of V.8.5.5.:

IX.2.9.15. THEOREM. Let (F_n) and (f_n) be sequences of functions on an interval I. Suppose

(a) Each f_n is continuous on I.

(b) Each F_n is differentiable on I, with $F_n' = f_n$.

(c) The sequence (f_n) converges u.c. on I to a function f.

(d) There is an $x_0 \in I$ such that the sequence $(F_n(x_0))$ converges.

Then:

(i) The sequence (F_n) converges u.c. on I to a function F.

(ii) F is differentiable on I and $F' = f$.

PROOF: Set $L = \lim_{n\to\infty} F_n(x_0)$. By the FTC2, we have, for each n and each $x \in I$,

$$F_n(x) = F_n(x_0) + \int_{x_0}^x f_n(t)\, dt \ .$$

For each fixed $x \in I$, since $f_n \to f$ uniformly on the interval between x_0 and x,

$$F_n(x) = F_n(x_0) + \int_{x_0}^x f_n(t)\, dt \to L + \int_{x_0}^x f(t)\, dt$$

by IX.2.6.11.; so, if we define, for $x \in I$,

$$F(x) = L + \int_{x_0}^{x} f(t)\,dt$$

we have by the FTC1 that F is differentiable on I and $F' = f$; and $F_n \to F$ pointwise on I. It remains to show that $F_n \to F$ u.c. on I. Let $[a,b]$ be a closed bounded subinterval of I containing x_0. Fix $\epsilon > 0$, and choose N so that $|f_n(x) - f(x)| < \frac{\epsilon}{2(b-a)}$ for all $n \geq N$ and all $x \in [a,b]$, and $|F_n(x_0) - L| < \frac{\epsilon}{2}$ for all $n \geq N$. Then, for $n \geq N$ and $x \in [a,b]$,

$$|F_n(x) - F(x)| \leq |F_n(x_0) - L| + \left| \int_{x_0}^{x} [f_n(t) - f(t)]\,dt \right| \leq |F_n(x_0) - L| + \int_{x_0}^{x} |f_n(t) - f(t)|\,dt$$

$$< \frac{\epsilon}{2} + \frac{\epsilon}{2(b-a)}|x - x_0| \leq \epsilon\ .$$



IX.2.10. Integration by Substitution and Integration by Parts

IX.2.10.1. Integration (antidifferentiation) is harder than differentiation. Differentiation of functions composed of standard algebraic, exponential, logarithm, trigonometric, inverse trigonometric, etc., functions can be done more or less mechanically by repeated application of the rules of differentiation. But there is no systematic procedure to find antiderivatives. A number of techniques have been developed to find antiderivatives of certain kinds of functions (a considerable portion of most university calculus sequences is devoted to "techniques of integration"), but no technique or procedure works in general; indeed, there are even functions with standard formulas, such as $f(x) = e^{-x^2}$, which are important in applications but which have no antiderivative formula in terms of standard functions (XII.12.3.18.). Thus there is no systematic technique to evaluate Riemann integrals via the Fundamental Theorem of Calculus in general.

Any differentiation rule can be turned around to give an antiderivative rule (although some care must be exercised to get precise and valid statements in general). Antiderivatives of certain specific functions can be found this way. And the Product Rule and Chain Rule can be turned around to give two of the most basic techniques of integration, Integration by Substitution and Integration by Parts.

Integration by Substitution or Change of Variable

Informally, the Chain Rule says that if F and ϕ are differentiable functions, with $F' = f$, then

$$\frac{d}{dx}\left(F(\phi(x))\right) = f(\phi(x))\phi'(x)$$

so $F \circ \phi$ is an antiderivative for $(f \circ \phi)\phi'$. To make a statement valid for Riemann integrals, we have to be a little careful with hypotheses. We will not state the most general result possible here (see IX.2.12.9.–IX.2.12.10. and XIX.5.5.6. for other results).

IX.2.10.2. THEOREM. [INTEGRATION BY SUBSTITUTION OR CHANGE OF VARIABLE, RIEMANN VERSION] Let I and J be intervals, ϕ a differentiable function on I, $\phi(I) \subseteq J$, and f a continuous function on J. Let $a, b \in I$, $a < b$. If $(f \circ \phi)\phi'$ is Riemann integrable on $[a, b]$, then

$$\int_a^b f(\phi(x))\phi'(x)\,dx = \int_{\phi(a)}^{\phi(b)} f(u)\,du \ .$$

By IX.2.12.7., the hypothesis that $(f \circ \phi)\phi'$ is Riemann integrable on $[a, b]$ is satisfied if ϕ' is Riemann integrable on $[a, b]$, since ϕ is continuous on I and f is continuous on J, so $f \circ \phi$ is continuous on I and thus on $[a, b]$. In particular, the hypothesis is satisfied if ϕ' is continuous on $[a, b]$, i.e. ϕ is $\mathcal{C}^1$ on $[a, b]$.

PROOF: Since f is continuous on J and $\phi(a), \phi(b) \in J$, the Riemann integral

$$\int_{\phi(a)}^{\phi(b)} f(u)\,du$$

is defined (we should not talk about the interval $[\phi(a), \phi(b)]$ since we could have $\phi(b) < \phi(a)$; nonetheless the integral is defined in this case too). Let F be an antiderivative for f on J (); then by the Chain Rule, $F \circ \phi$ is differentiable on I, with derivative $(f \circ \phi)\phi'$. By two applications of the FTC2,

$$\int_{\phi(a)}^{\phi(b)} f(u)\,du = F(\phi(b)) - F(\phi(a)) = \int_a^b f(\phi(x))\phi'(x)\,dx \ .$$



IX.2.10.3. Note that there is no requirement in this result that ϕ is monotone, or even that $\phi([a, b])$ is contained in the interval between $\phi(a)$ and $\phi(b)$ - it just has to be contained in the interval J on which f is defined and continuous. Thus XIX.5.5.6., which has a monotonicity hypothesis on ϕ, is not strictly a generalization of IX.2.10.2. (see IX.2.12.9.–IX.2.12.10. for special cases of XIX.5.5.6. for Riemann integration).

IX.2.10.4. This result is called "Integration by Substitution" or "Change of Variable" since in its applications we usually think of making the "substitution" $u = \phi(x)$, so notationally we have

$$\frac{du}{dx} = \phi'(x)$$

$$du = \phi'(x)\,dx$$

even though both sides of the last equation are "meaningless." We then have

$$f(u)\,du = f(\phi(x))\phi'(x)\,dx$$

so their integrals should be equal. But we must change the limits from the extreme values of x to the corresponding extreme values of u:

$$\int_{u=\phi(a)}^{u=\phi(b)} f(u)\,du = \int_{x=a}^{x=b} f(\phi(x))\phi'(x)\,dx \ .$$

This way of writing the "substitution" highlights the utility of the Leibniz notation for integrals.

IX.2.10.5. Thus we convert an integral in x which is possibly difficult to evaluate directly into an integral in u which may be simpler. The substitution of course depends on the choice of ϕ; there may be more than one way to choose ϕ to rewrite the integrand as $(f \circ \phi)\phi'$ for some f, and the most useful choice of ϕ may not be obvious. Calculus courses extensively discuss the types of integrals for which this technique is successful, and the proper choice of ϕ in various situations. But of course this technique is only applicable to certain kinds of integrals where the integrand can be put into the form $(f \circ \phi)\phi'$ for some choice of f and ϕ.

IX.2.10.6. Example. Although this is not a calculus text and we will not discuss techniques of integration in any detail, we give one simple typical example illustrating the technique. To compute

$$\int_1^4 2x\sqrt{x^2+1}\,dx$$

we make the substitution $u = x^2 + 1$, i.e. $\phi(x) = x^2 + 1$. Then $\phi'(x) = 2x$, i.e. $du = 2x\,dx$. (The hypotheses are satisfied since ϕ is $\mathcal{C}^1$ on all of $\mathbb{R}$ and its range is $[1, +\infty)$, on which $f(u) = \sqrt{u}$ is continuous.) Since $\phi(1) = 2$ and $\phi(4) = 17$, we have

$$\int_1^4 2x\sqrt{x^2+1}\,dx = \int_1^4 \sqrt{x^2+1}\cdot 2x\,dx = \int_2^{17} \sqrt{u}\,du = \frac{2}{3}u^{3/2}\Big|_2^{17} = \frac{2}{3}(17^{3/2} - 2^{3/2})\ .$$

To illustrate the comment in IX.2.10.3., instead compute

$$\int_{-1}^1 2x\sqrt{x^2+1}\,dx$$

in the same way. This time $\phi(-1) = \phi(1) = 2$, so

$$\int_{-1}^1 2x\sqrt{x^2+1}\,dx = \int_2^2 \sqrt{u}\,du = 0\ .$$

This integral could have been evaluated without calculation, since the integrand is an odd function so its integral over any (bounded) interval centered at 0 is 0.

Integration by Parts

IX.2.10.7. If F and G are differentiable on an interval I with $F' = f$ and $G' = g$, then by the Product Rule FG is differentiable on I and

$$\frac{d}{dx}\left(F(x)G(x)\right) = F(x)g(x) + f(x)G(x)$$

so FG is an antiderivative for $Fg + fG$ on I. We thus obtain:

IX.2.10.8. Theorem. [Integration by Parts, Riemann Version] Let F and G be differentiable on $[a, b]$, with $F' = f$ and $G' = g$. If Fg and fG are Riemann integrable on $[a, b]$, then

$$\int_a^b F(x)g(x)\,dx = F(b)G(b) - F(a)G(a) - \int_a^b f(x)G(x)\,dx\ .$$

The hypotheses are satisfied if f and g are Riemann integrable on $[a,b]$, and in particular if they are continuous, i.e. if F and G are $\mathcal{C}^1$ on $[a,b]$.

PROOF: By the FTC2, we have

$$\int_a^b F(x)g(x)\,dx + \int_a^b f(x)G(x)\,dx = \int_a^b [F(x)g(x) + f(x)G(x)]\,dx = F(b)G(b) - F(a)G(a)\ .$$

IX.2.10.9. This result is often efficiently written notationally by setting $u = F(x)$ and $v = G(x)$, so $du = f(x)\,dx$ and $dv = g(x)\,dx$. The statement is then

$$\int_a^b u\,dv = uv\big|_a^b - \int_a^b v\,du\ .$$

IX.2.10.10. This technique is useful in integrating products, which are generally hard to integrate. The integral of Fg, which may be hard to evaluate directly, is converted into the integral of fG, which may be easier. To be successful on a particular problem, the technique requires the proper choice of F and G, which may not always be obvious.

IX.2.10.11. EXAMPLE. We again give a simple typical example. To evaluate

$$\int_0^{\pi/2} x\cos x\,dx$$

we may choose $F(x) = x$ and $g(x) = \cos x$. Thus $f(x) = 1$ and $G(x) = \sin x$ (or, in the alternate notation, $u = x$ and $dv = \cos x\,dx$, so $du = dx$ and $v = \sin x$), so

$$\int_0^{\pi/2} x\cos x\,dx = (x\sin x)\big|_0^{\pi/2} - \int_0^{\pi/2} 1\cdot\sin x\,dx$$

$$= (x\sin x)\big|_0^{\pi/2} - (-\cos x)\big|_0^{\pi/2} = \left(\frac{\pi}{2}\cdot 1 - 0\cdot 0\right) - (0-(-1)) = \frac{\pi}{2} - 1\ .$$

If we had made a bad choice of F and g, say $F(x) = \cos x$ and $g(x) = x$, so $f(x) = -\sin x$ and $G(x) = \frac{1}{2}x^2$, we would obtain the formula

$$\int_0^{\pi/2} x\cos x\,dx = \left(\frac{1}{2}x^2\cos x\right)\Big|_0^{\pi/2} - \int_0^{\pi/2} -\frac{1}{2}x^2\sin x\,dx = \int_0^{\pi/2} \frac{1}{2}x^2\sin x\,dx$$

which is correct but not useful since the new integral is even harder than the original one (in fact, to evaluate it one must do integration by parts twice, first converting it back to the original integral!).

IX.2.11. Improper Riemann Integration

The fact that unbounded functions cannot be Riemann integrable is partially compensated by the various forms of improper Riemann integration.

IX.2.11.1. DEFINITION. If f is a real-valued function defined on $(a, b]$, and Riemann integrable on $[a + \epsilon, b]$ for every $\epsilon > 0$, then f is *improperly Riemann integrable* from a to b if

$$\lim_{\epsilon \to 0+} \left[\int_{a+\epsilon}^{b} f(x)\, dx \right]$$

exists (in the usual sense, as a real number). We then define the *improper Riemann integral* to be

$$\int_a^b f(x)\, dx = \lim_{\epsilon \to 0+} \left[\int_{a+\epsilon}^{b} f(x)\, dx \right] .$$

If f is nonnegative on $(a, b]$, then

$$\lim_{\epsilon \to 0+} \left[\int_{a+\epsilon}^{b} f(x)\, dx \right] = \sup_{\epsilon > 0} \left[\int_{a+\epsilon}^{b} f(x)\, dx \right]$$

always exists (in the regular or extended sense), and in this case, we define

$$\int_a^b f(x)\, dx = \lim_{\epsilon \to 0+} \left[\int_{a+\epsilon}^{b} f(x)\, dx \right] .$$

A nonnegative function is then improperly Riemann integrable from a to b if $\int_a^b f(x)\, dx < \infty$.

IX.2.12. Exercises

IX.2.12.1. (a) Show that a function which is continuous almost everywhere and zero on a dense set is zero almost everywhere. [Show it is zero wherever it is continuous.]

(b) If f and g are functions which are each continuous almost everywhere, and f and g agree on a dense set, then $f = g$ a.e.

IX.2.12.2. Show that a finite union of sets of zero content has zero content. [You can mimic the proof of IX.2.7.6., but there is an easier argument.]

IX.2.12.3. Prove Proposition IX.2.7.11..

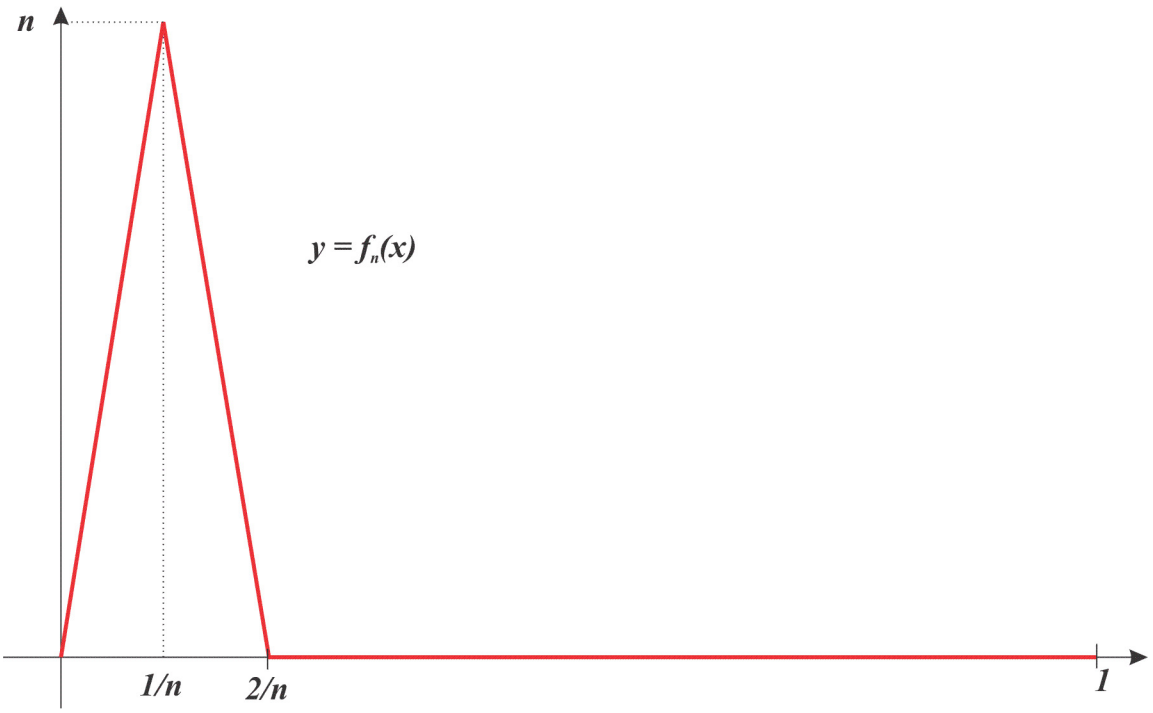


Figure IX.16: Graph of a Function

IX.2.12.4. This problem is a continuation of V.8.6.7..

(a) Show that every regulated function on an interval I is Riemann-integrable on any closed bounded subinterval of I. Then deduce V.8.6.7.(c) from IX.2.9.3..

(b) Let $g(x) = \sin(1/x)$ for $x \neq 0$, and $g(0) = 0$. Let (r_n) be an enumeration of $\mathbb{Q}$, and set

$$f(x) = \sum_{n=1}^{\infty} 2^{-n} g(x - r_n) \ .$$

Show that f is discontinuous at every point of $\mathbb{Q}$ and continuous at every point of $\mathbb{J}$.

(c) Show that f is Riemann-integrable over every closed bounded subinterval of $\mathbb{R}$ (give two proofs, one using IX.2.8.5. and one using ()), but f is not regulated on any interval.

(d) Show using IX.2.9.8.(iii) and V.8.5.5. that f has an antiderivative on $\mathbb{R}$.

IX.2.12.5. Let f be a bounded function on an interval $[a,b]$, and $\overline{f}$ its upper envelope (IX.2.8.1.).

(a) Show that

$$\overline{f} = \inf\{\psi : \psi \geq f, \psi \text{ an upper semicontinuous step function}\} .$$

[Use the argument of the proof of IX.2.8.2..]

(b) Show that, for every $x_0 \in [a,b]$,

$$\overline{f}(x_0) = \max\left(f(x_0), \limsup_{x \to x_0} f(x)\right) .$$

(c) Give similar characterizations of $\underline{f}$.

IX.2.12.6. Let E be a subset of $[a,b]$, and consider the indicator function χ_E.

(a) Show that the upper envelope (IX.2.8.1.) $\overline{\chi_E}$ is $\chi_{\bar{E}}$, and the lower envelope $\underline{\chi_E}$ is χ_{E°.

(b) Show that χ_E is Riemann-integrable on $[a,b]$ if and only if ∂E has zero content (), i.e. E is a Jordan region () in $\mathbb{R}$.

IX.2.12.7. (a) Use Theorem IX.2.8.5. to prove that if f and g are Riemann integrable on $[a,b]$, so is fg.

(b) To obtain a direct low-tech proof of (a), suppose f is a bounded function on an interval I, say $|f(x)| \leq M$ for all $x \in I$. Show that

$$|f(x)^2 - f(y)^2| \leq 2M|f(x) - f(y)|$$

for all $x, y \in I$.

(c) If f is a bounded function on $[a,b]$, say $|f(x)| \leq M$ for all $x \in [a,b]$, and if $\mathcal{P}$ is any partition of $[a,b]$, then

$$\overline{\mathcal{S}}(f^2, \mathcal{P}) - \underline{\mathcal{S}}(f^2, \mathcal{P}) \leq 2M[\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P})] .$$

(d) Conclude from (c) that if f is Riemann integrable on $[a,b]$, so is f^2.

(e) Use the identity

$$4fg = (f+g)^2 - (f-g)^2$$

for functions f, g on $[a,b]$ to prove (a).

(f) Give a similar proof using the triangle inequality that if f is Riemann integrable on $[a,b]$, so is $|f|$. What about the converse? (cf. ())

IX.2.12.8. Let ϕ be a differentiable function on an interval $[a,b]$. Suppose there are constants m, M with $0 < m \le M$ such that $m \le |\phi'(x)| \le M$ for all $x \in [a,b]$.

(a) Show that ϕ is strictly monotone on $[a,b]$, that $\phi([a,b])$ is an interval $[c,d]$, and that $\phi^{-1} : [c,d] \to [a,b]$ is differentiable with

$$\frac{1}{M} \le (\phi^{-1})'(x) \le \frac{1}{m}$$

for all $x \in [c,d]$.

(b) If $A \subseteq [a,b]$, show that A has measure 0 [resp. content 0] if and only if $\phi(A)$ has measure 0 [resp. content 0].

(c) Show that if f is a function on $[c,d]$, then f is Riemann integrable on $[c,d]$ if and only if $f \circ \phi$ is Riemann integrable on $[a,b]$.

The result in (b) can be extended to the case where ϕ is differentiable and ϕ' is never 0 (XIX.5.1.16.).

IX.2.12.9. (a) Extend the result of IX.2.10.2. to the case where f is bounded and has countably many discontinuities, provided ϕ is injective (i.e. strictly monotone).

(b) What is the difficulty in (a) if ϕ is not one-to-one (i.e. strictly monotone)? Is there a weaker restriction on ϕ which suffices?

(c) What is the difficulty in (a) if f is just assumed to be Riemann integrable on $[c,d]$, even if ϕ is strictly monotone? See IX.2.12.8..

IX.2.12.10. Use IX.2.12.8.(c) and IX.2.12.7. to prove the following theorem, which is a special case of XIX.5.5.6.. [Mimic the proof of IX.2.10.2..]

THEOREM. Let ϕ be a differentiable function on an interval $[a,b]$. Suppose there are constants m, M with $0 < m \le M$ such that $m \le |\phi'(x)| \le M$ for all $x \in [a,b]$, and that ϕ' is Riemann integrable on $[a,b]$. Set $[c,d] = \phi([a,b])$. If f is Riemann integrable on $[c,d]$, then $(f \circ \phi)|\phi'|$ is Riemann integrable on $[a,b]$, and

$$\int_a^b f(\phi(x))|\phi'(x)|\,dx = \int_c^d f(u)\,du \ .$$

Note that the hypotheses on ϕ' imply that ϕ is strictly monotone of $[a,b]$, and are automatically satisfied if ϕ' is continuous and nonzero on $[a,b]$. What is the (modest) significance of the absolute value on the ϕ'? [Hint: how are c and d related to $\phi(a)$ and $\phi(b)$?]

IX.2.12.11. The result of this problem may be "geometrically obvious," but requires proof. Let $a > 0$ and f an odd function () on $[-a,a]$ which is Riemann integrable on $[-a,a]$.

(a) Show that if $\mathcal{P}$ is a partition of $[-a,a]$ which is symmetric around 0, then 0 is a Riemann sum for $\int_{-a}^{a} f(x)\,dx$ subordinate to $\mathcal{P}$.

(b) Conclude that $\int_{-a}^{a} f(x)\,dx = 0$.

(c) Use IX.2.12.10. to show that

$$\int_{-a}^{a} f(x)\,dx = -\int_{-a}^{a} f(x)\,dx$$

to obtain an alternate proof of (b).

IX.2.13. Numerical Integration and Simpson's Rule

Usually the best way to compute the exact value of a Riemann integral is via the Fundamental Theorem of Calculus. But sometimes integrals are encountered where this approach is not possible or feasible, and to calculate the integral we need to return to the definition and compute the integral directly as a limit.

We know in theory that if a function is Riemann integrable on an interval, then the Riemann sums for the integral converge to the exact integral as the partitions become sufficiently fine. But how can we choose a Riemann sum, or similar sum, which gives the integral to within a specified accuracy, and how do we know how accurate an approximation we get from a given sum? Numerical integration theory is the study of these problems.

Numerical methods for integration are essentially based on Riemann integration, and do not work satisfactorily unless the function being integrated is well-behaved (piecewise-continuous at the least, and usually with some smoothness). Thus numerical integration is really a part of the theory of Riemann integration and not more advanced integration theories. Numerical integration is frequently used in situations where function values can only be calculated easily at certain points, and even then sometimes only to a certain accuracy, or to functions whose values are known only from data sets.

Numerical integration is an important part of the subject of *numerical analysis*, a part of mathematics (usually regarded as part of "applied mathematics") which seeks to develop efficient ways to numerically calculate or approximate various important mathematical quantities, e.g. integrals, sums of infinite series, zeroes of functions, eigenvalues and eigenvectors, solutions to differential equations. Much of numerical analysis has a very different character than the type of analysis developed in this book, but the word *analysis* is appropriately used for this subject since (among other things) analytic techniques are used in determining the accuracy and efficiency of the numerical algorithms developed. We will only scratch the surface of numerical integration here; books such as [Hil87] or [IK94] have a far more thorough treatment.

Quadrature Procedures

IX.2.13.1. Suppose we have a fixed interval, which we will for simplicity take to be a closed bounded interval $[a, b]$, and we want to have a fixed procedure for estimating integrals

$$\int_a^b f(x)\,dx$$

for continuous functions f. The obvious procedure is to fix points $x_0, \ldots, x_n$ in $[a, b]$ (which we will always take to be in increasing order). These will typically be a partition of $[a, b]$, but not necessarily, i.e. the endpoints need not be among the x_k. We will then fix real constants $\alpha_0, \ldots, \alpha_n$, usually positive (a *nonnegative procedure*, cf. IX.2.13.4.). The procedure $Q = Q(x_0, \ldots, x_n; \alpha_0, \ldots, \alpha_n)$ then gives the approximation

$$\int_a^b f(x)\,dx \approx Q(f) = \sum_{k=0}^{n} \alpha_k f(x_k)\ .$$

IX.2.13.2. The first obvious property we should have is that the approximation should be exact for constant functions. This translates into the restriction

$$\sum_{k=0}^{n} \alpha_k = b - a\ .$$

IX.2.13.3. Another more restrictive condition which is natural is that for any continuous f,

$$\underline{\mathcal{S}}(f,\mathcal{P}) \leq Q(f) \leq \overline{\mathcal{S}}(f,\mathcal{P})$$

for some partition $\mathcal{P}$ of $[a,b]$, where $\underline{\mathcal{S}}(f,\mathcal{P})$ and $\overline{\mathcal{S}}(f,\mathcal{P})$ are the lower and upper Darboux sums (). A procedure satisfying this condition is called *regular* with respect to $\mathcal{P}$. If Q is regular with respect to the partition of $[a,b]$ given by $x_0,\ldots,x_n$ and the endpoints of the interval if not included among the x_k, then Q is called *regular*.

Some quadrature procedures considered in numerical analysis are regular, but many are not, although most are regular with respect to a coarser partition. The condition $\sum \alpha_k = b-a$ insures that all nonnegative procedures are regular with respect to the trivial partition, since any nonnegative procedure is monotone.

IX.2.13.4. In numerical work, subtraction, or addition of numbers of different magnitudes, can lead to a serious loss of precision; thus the most useful quadrature procedures are ones where the α_n are all the same sign (positive) and approximately the same size, although these restrictions are not necessary in principle.

IX.2.13.5. Definition. A *quadrature procedure*, or *numerical integration procedure*, on an interval $[a,b]$ is a choice of points $x_0,\ldots,x_n \in [a,b]$ and real numbers $\alpha_0,\ldots,\alpha_n$ with $\sum_{k=0}^{n} \alpha_k = b-a$. For any real-valued function f, the application of the procedure $Q = Q(x_0,\ldots,x_n;\alpha_0,\ldots,\alpha_n)$ to f is the sum

$$Q(f) = \sum_{k=0}^{n} \alpha_k f(x_k) \ .$$

$Q(f)$ is called the *approximation to the integral of f over $[a,b]$ by the procedure Q*.

Definite integration, especially numerical integration, is often called *quadrature* because of the standard interpretation of integrals as areas.

IX.2.13.6. We could also consider more general quadrature procedures of the form

$$Q(f) = \sum_{k=0}^{n} \alpha_k f(x_k) + \sum_{k=0}^{n} \beta_k f'(x_k) + \sum_{k=0}^{n} \gamma_k f''(x_k) + \cdots$$

which are sometimes used in numerical analysis (cf. IX.2.13.19., IX.2.13.39., IX.2.13.51.).

IX.2.13.7. A *procedure* by this definition must be carefully distinguished from a numerical integration *process* or *scheme*, which is a sequence of quadrature procedures converging to the exact integral for all, or at least nicely behaved, functions. For example, the standard procedures like Simpson's rule begin by partitioning the interval into n subintervals in some specific way. There is thus one procedure for each fixed n, and they form a process as $n \to \infty$.

Note that any sequence (Q_n) of quadrature procedures regular with respect to partitions $\mathcal{P}_n$ with $\|\mathcal{P}_n\| \to 0$ as $n \to \infty$ will converge to the exact integral for all continuous functions (in fact for all Riemann-integrable functions), i.e. will form a numerical integration process. However, for practical reasons the *rate* of convergence is a critical consideration in numerical analysis.

IX.2.13.8. It is easily checked that any quadrature procedure Q is linear, i.e.

$$Q(cf+g) = cQ(f) + Q(g)$$

for any functions f, g and scalar c; a procedure with all $\alpha_k > 0$ is also monotone, i.e. $Q(f) \le Q(g)$ if $f \le g$.

IX.2.13.9. EXAMPLES. (i) The simplest quadrature procedures are the left and right estimates for the Riemann integral (): if $\{x_0, \dots, x_n\}$ is a partition of $[a,b]$, set $\alpha_k = x_{k+1} - x_k$ for $0 \le k < n$ and $\alpha_n = 0$. Then the estimate

$$\sum_{k=0}^{n} \alpha_k f(x_k)$$

is the left estimate L_n for the integral of f. If we instead take $\alpha_0 = 0$ and $\alpha_k = x_k - x_{k-1}$ for $1 \le k \le n$, we get the right estimate R_n.

(ii) More generally, if $\{x_0, \dots, x_n; t_1, \dots, t_n\}$ is a tagged partition of $[a,b]$, set $\alpha_k = \Delta x_k = x_k - x_{k-1}$ for $1 \le k \le n$. Then $Q(t_1, \dots, t_n; \alpha_1, \dots, \alpha_n)$ gives the Riemann sum approximation to the integral for this tagged partition:

$$Q(f) = \sum_{k=1}^{n} f(t_k)\Delta x_k \ .$$

A numerically good procedure is to take the x_k to be evenly spaced and t_k the midpoint of the k'th interval. This procedure is called the *midpoint rule* with n intervals, denoted M_n. Note that in this procedure the points where the function is evaluated do not form a partition of the interval, i.e. the endpoints are not included.

(iii) Here is a quadrature procedure which is not just a Riemann sum. We average the left and right estimates from (i), normally using n subintervals of equal length $h = \frac{b-a}{n}$. In this case $\alpha_0 = \alpha_n = \frac{h}{2}$ and $\alpha_k = h$ for $1 \le k \le n-1$.

This procedure is called the *trapezoid rule* with n intervals, for the following reason. It is also obtained by approximating the function within each subinterval by a linear (affine) function, so that the integral is approximated by a sum of areas of trapezoids (Figure IX.17, from http://tutorial.math.lamar.edu/Classes/CalcII/ApproximatingDefIntegrals.aspx). We call this procedure T_n.

An alternate formula for T_n is

$$T_n(f) = \frac{h}{2}\left[f(x_0) + 2f(x_1) + 2f(x_2) + \cdots + 2f(x_{n-1}) + f(x_n)\right]$$

where $h = \frac{b-a}{n}$ and $x_k = x_0 + kh$ for $0 \le k \le n$.

All these quadrature procedures are regular, as is easily verified.

Simpson's Rule

IX.2.13.10. If the graph of f is concave up on $[a,b]$, it is clear that the trapezoid estimate is too large and the midpoint estimate is too small, for any fixed n, and vice versa if the graph of f is concave down. Thus we would expect the errors to tend to have opposite sign. And experiments with calculable integrals suggest that the absolute error from the midpoint estimate for a given n is typically about half the absolute

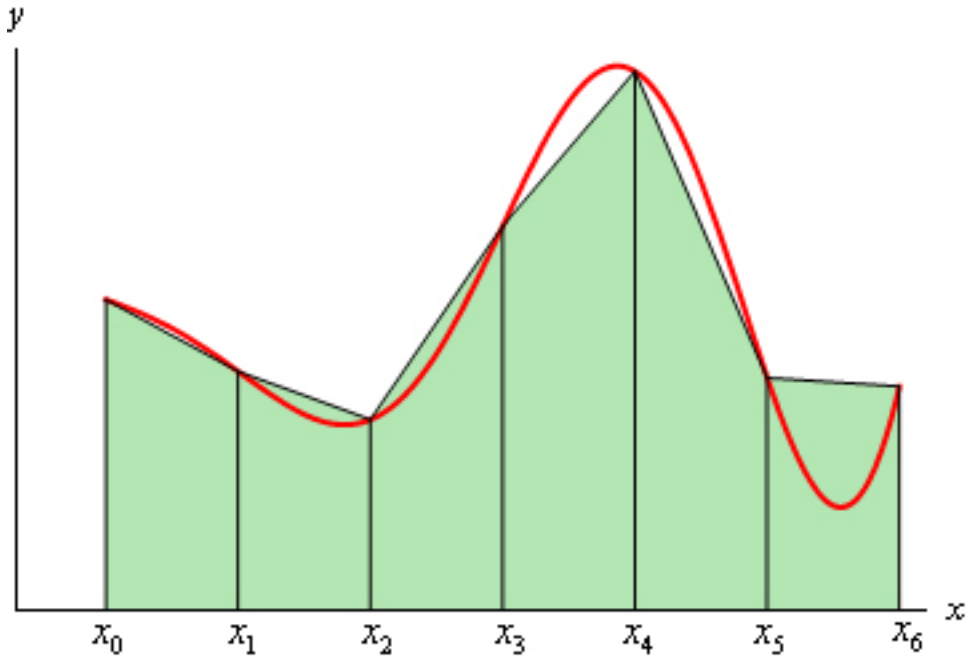


Figure IX.17: The trapezoid rule

error of the trapezoid estimate for the same n – see IX.2.13.44. for a precise statement along these lines. (Some unsophisticated references assert or suggest the absolute midpoint error is exactly half the absolute trapezoid error, but in fact it is rarely exactly half, and in some cases is not even approximately half; cf. Exercise IX.2.14.2..)

IX.2.13.11. These statements suggest that in many cases, twice the midpoint error should largely cancel with the trapezoid error; thus a weighted average of the two should often be much better than either. So we define

$$S_{2n}(f) = \frac{2}{3}M_n(f) + \frac{1}{3}T_n(f)$$

which is *Simpson's rule* with $2n$ intervals.

To put Simpson's rule into the standard form of a quadrature procedure and explain the use of the index $2n$, note that in evaluating $M_n(f)$ and $T_n(f)$, the function f must be evaluated at $2n+1$ points, the $n+1$ points of the partition and the n midpoints. These points form a partition of the interval $[a,b]$ into $2n$ subintervals of equal length.

IX.2.13.12. Changing notation slightly, if n is an even number and $\{x_0,\dots,x_n\}$ is a partition of $[a,b]$ into n subintervals of equal length $h = \frac{b-a}{n}$, we have $\alpha_0 = \alpha_n = \frac{h}{3}$, $\alpha_k = \frac{2h}{3}$ for k even, $0 < k < n$, and

$\alpha_k = \frac{4h}{3}$ for k odd. Thus

$$S_n(f) = \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)\right]$$

where $x_k = x_0 + kh$ for $0 \leq k \leq n$.

IX.2.13.13. The Simpson's rule approximation with $2n$ intervals is indeed typically (but not always!) *much* better than the trapezoid or midpoint approximations with n intervals. The function must, of course, be evaluated at about twice as many points. But $S_{2n}(f)$ is typically even considerably more accurate than $T_{2n}(f)$ or $M_{2n}(f)$. See IX.2.13.36.–IX.2.13.52. for an error analysis.

Simpson's rule S_{2n} is not a regular quadrature procedure, but it is regular with respect to the partition of $[a, b]$ into n equal subintervals; hence $S_{2n}(f) \to \int_a^b f(x)\,dx$ as $n \to \infty$ for any continuous f (even every Riemann-integrable f), as does $T_n(f)$ and $M_n(f)$.

IX.2.13.14. There is a quite different way of motivating and arriving at Simpson's rule; see IX.2.13.24. and IX.2.13.26..

Is There a "Best" Quadrature Procedure?

IX.2.13.15. We would like to find and use the "best" quadrature procedure. However, it appears obvious that we run into a tradeoff between accuracy and complexity, since it stands to reason (and is "generally" correct) that procedures using more points would tend to be more accurate; thus we should not expect a universal procedure which is "best" in all respects.

But there is a more fundamental sense in which there is no "best" procedure. Given *any* two quadrature procedures Q_1 and Q_2, a continuous function f_1, even a polynomial, can be constructed such that $Q_1(f_1)$ is a better approximation to the true integral of f_1 than $Q_2(f_1)$, and another continuous function (polynomial) f_2 such that $Q_2(f_2)$ is better than $Q_1(f_2)$. So even beyond questions of complexity of computation there is no universally "best" quadrature procedure.

IX.2.13.16. We will appear to argue that Simpson's rule is generally the best rule to use for numerical integration, and within the shallow limits of our numerical analysis work here this statement is defensible for many typically encountered integrals, and Simpson's rule is the most widely used quadrature procedure. However, the reader should be aware that it is much too simplistic to tout Simpson's rule as the ultimate quadrature procedure; there are many others (e.g. Romberg integration, cf. IX.2.14.7., or Clenshaw-Curtis Integration, cf. ()) which are better in certain circumstances. See IX.2.13.61. for further comments.

Simple and Composite Quadrature Procedures

One fairly good measure of the accuracy of a quadrature procedure is how well it approximates the exact integral for polynomials of small order.

IX.2.13.17. Definition. Let Q be a quadrature procedure on $[a, b]$. Then Q has *order of precision* m if

$$Q(f) = \int_a^b f(x)\,dx$$

(i.e. Q is exactly correct) whenever f is a polynomial of degree $\leq m$, but not for polynomials of degree $m+1$.

The term *order of precision* is standard, but some references use *order of exactness* which seems more descriptive.

We have built into our definition that a quadrature procedure always has order of precision at least 0.

IX.2.13.18. A quadrature procedure is *simple* if it has high order of precision for the number of nodes used. A *composite* quadrature procedure is one which consists of first dividing the interval $[a,b]$ into a number of smaller intervals and applying a simple quadrature procedure on each subinterval (the simple procedures used on the subintervals are usually, but not always, the same on all subintervals). The order of precision of a composite procedure is at least as great as (usually exactly equal to) the minimum of the order of precision of the simple procedures used on the subintervals. A composite procedure is always regular with respect to the partition into the subintervals on which the simple procedures are applied.

For reasons described in IX.2.13.4., as well as considerations of efficiency, composite procedures using a simple procedure of lower order on each subinterval are often better in practice than simple procedures of higher order.

Simple quadrature procedures are almost always interpolatory or osculatory:

IX.2.13.19. Definition. Let $x_0,\dots,x_n$ be points of $[a,b]$. The *interpolatory quadrature procedure* on $[a,b]$ for the nodes $x_0,\dots,x_n$ is the procedure Q defined by

$$Q(f)=\int_a^b p_{0,\dots,n}(x)\,dx$$

where $p_{0,\dots,n}$ is the interpolating polynomial of f on $[a,b]$ for these nodes (V.11.1.1.).

It follows easily from the Lagrange or Newton form of the interpolating polynomial that an interpolatory quadrature procedure is indeed a quadrature procedure of the form of IX.2.13.5..

This definition can be extended to the situation where the $x_0,\dots,x_n$ are not necessarily distinct by using the osculating polynomial (V.11.6.4.); such a quadrature procedure is called *osculatory*. (An osculatory procedure is of the generalized form of IX.2.13.6., and can only be used for functions which are sufficiently differentiable.)

IX.2.13.20. It is immediate that an interpolatory or osculatory quadrature procedure using $n+1$ nodes has order of precision at least n. (Conversely, a quadrature procedure using $n+1$ nodes of order of precision at least n must by interpolatory or osculatory; but cf. IX.2.14.6..) If the nodes are well chosen, the order of precision can be greater than n (IX.2.13.30., IX.2.13.46.).

IX.2.13.21. There are thus two general strategies for developing simple quadrature procedures:

(1) First pick nodes $x_0,\dots,x_n$ in a convenient way, e.g. evenly spaced, and use the interpolatory or osculatory procedure for these nodes.

(2) Fix n and choose the x_k carefully to maximize the order of precision of the corresponding interpolatory or osculatory procedure.

The Newton-Cotes Formulas

IX.2.13.22. We first examine the procedure obtained by choosing a partition of $[a,b]$ into n subintervals of equal length $h = \frac{b-a}{n}$, i.e. $x_k = a + kh$ for $0 \le k \le n$. The corresponding interpolatory procedure is called the *Newton-Cotes procedure* (or *Newton-Cotes formula*) of order n, denoted NC_n.

Since an explicit formula for the interpolating polynomial can be obtained, a formula for the procedure NC_n can be found by integrating this polynomial. The calculation is straightforward but tedious, and we carry it out only for $n = 1$ and $n = 2$.

IX.2.13.23. First consider the case $n = 1$, i.e. $x_0 = a$, $x_1 = b$. The interpolating polynomial is

$$p_{0,1}(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

so we have

$$NC_1(f) = \int_a^b \left[f(a) + \frac{f(b) - f(a)}{b - a}(x - a) \right] dx$$
$$= f(a)(b - a) + \frac{1}{2}\frac{f(b) - f(a)}{b - a}(b - a)^2 = \frac{h}{2}[f(a) + f(b)]$$

which is just the trapezoid rule for $n = 1$, i.e. $NC_1 = T_1$.

IX.2.13.24. Now consider the case $n = 2$, i.e. $x_0 = a$, $x_1 = \frac{a+b}{2}$, $x_2 = b$. The interpolating polynomial is

$$p_{0,1,2}(x) = f(a) + \frac{f(b) - f(a)}{b - a}(x - a) + \frac{f[b, x_1] - f[a, b]}{x_1 - a}(x - a)(x - b)$$
$$= f(a) + \frac{f(b) - f(a)}{2h}(x - a) + \frac{1}{2h}\left[\frac{f(b) - f(x_1)}{h} - \frac{f(b) - f(a)}{2h} \right](x - a)(x - b)$$
$$= f(a) + \frac{f(b) - f(a)}{2h}(x - a) + \frac{1}{4h^2}\left[f(b) - 2f(x_1) + f(a) \right](x - a)(x - b)$$

so we have

$$NC_2(f) = \int_a^b \left[f(a) + \frac{f(b) - f(a)}{2h}(x - a) + \frac{1}{4h^2}\left[f(b) - 2f(x_1) + f(a) \right](x - a)(x - b) \right] dx \ .$$

We have

$$\int_a^b (x - a)(x - b)\, dx = -\frac{(2h)^3}{6}$$

by making the substitution $x = a + th$; thus

$$NC_2(f) = f(a)(2h) + \frac{f(b) - f(a)}{2h}\left(\frac{1}{2}(2h)^2 \right) - \frac{8h^3}{6}\frac{1}{4h^2}[f(b) - 2f(x_1) + f(a)] = \frac{h}{3}[f(a) + 4f(x_1) + f(b)]$$

which is exactly $S_2(f)$ (IX.2.13.12.), i.e. NC_2 is exactly the simple Simpson's rule for $n = 2$: Simpson's rule is obtained by approximating the graph of f by a parabola through the endpoints and midpoint, cf. Figure IX.18, from http://en.wikipedia.org/wiki/Simpson's_rule. (In fact, this is how SIMPSON came up with the rule in the first place in 1743; the simple Simpson's rule was previously essentially known to others, including NEWTON and KEPLER, and was apparently first described by CAVALIERI in 1639.)

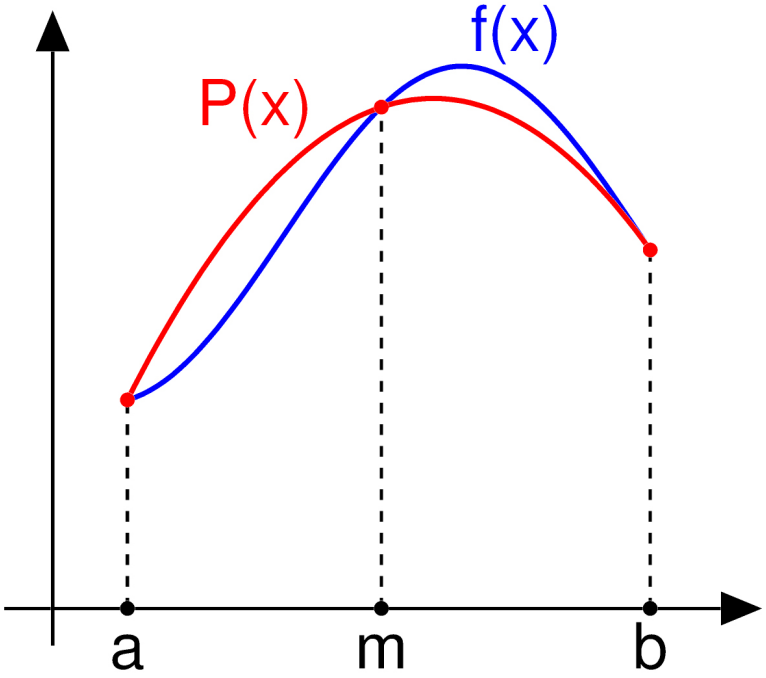


Figure IX.18: The simple Simpson's rule

IX.2.13.25. We list without proof a few more of the Newton-Cotes formulas:

$$NC_3(f) = \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)]$$

$$NC_4(f) = \frac{2h}{45}[7f(x_0) + 32f(x_1) + 12f(x_2) + 32f(x_3) + 7f(x_4)]$$

$$NC_5(f) = \frac{5h}{288}[19f(x_0) + 75f(x_1) + 50f(x_2) + 50f(x_3) + 75f(x_4) + 19f(x_5)]$$

$$NC_8(f) = \frac{4h}{14175}[989f(x_0)+5888f(x_1)-928f(x_2)+10496f(x_3)-4540f(x_4)+10496f(x_5)-928f(x_6)+5888f(x_7)+989f(x_8)]$$

IX.2.13.26. As n increases, the coefficients in NC_n vary more greatly. More problematically, if $n \geq 8$ the coefficients are not all positive. Thus, for reasons described in IX.2.13.4., higher Newton-Cotes formulas are not well suited for numerical work. Composite procedures using Newton-Cotes formulas of smaller degree on the subintervals generally work better numerically.

It turns out that if n is even, NC_n generally gives almost as good an approximation as NC_{n+1} (and sometimes better), for reasons described in IX.2.13.57.. In particular, NC_2 (simple Simpson's rule) is generally nearly as good as NC_3 (*Simpson's three-eighths rule*), which requires more arithmetic. (In fact, the

composite Simpson's rule is generally even better than the composite three-eighths rule for the *same* overall number of subintervals; cf. Exercise IX.2.14.4..)

The ordinary Simpson's rule S_{2n} of IX.2.13.12. is nothing but the composite procedure consisting of dividing $b-a$ into n subintervals of equal length and then applying the simple Simpson's rule S_2 on each subinterval; for this reason, Simpson's rule is sometimes called the *parabolic rule*. In part for the reasons in the last two paragraphs, Simpson's rule is a popular and generally very good procedure.

IX.2.13.27. As a variation, given n we can divide $[a,b]$ into $n+2$ subintervals of equal length $h=\frac{b-a}{n+2}$, and not use the two endpoints but only the $n+1$ interior points $x_1,\dots,x_{n+1}$, where $x_k = a+kh$ for $1\le k\le n+1$. The resulting formulas are the *open Newton-Cotes formulas*, written NC_n^o. The only one of interest to us is NC_0^o, which is just the simple midpoint rule M_1. We list a couple more:

$$NC_1^o(f) = \frac{3h}{2}[f(x_1)+f(x_2)]$$

$$NC_2^o(f) = \frac{4h}{3}[2f(x_1)-f(x_2)+2f(x_3)]$$

Already the coefficients in NC_2^o are not all positive. NC_n^o is again of order of precision at least n.

Gaussian Quadrature

IX.2.13.28. We now change indexing convention slightly, beginning with $k=1$ instead of $k=0$. The quadrature formula

$$Q(f) = \sum_{k=1}^{n} \alpha_k f(x_k)$$

has $2n$ unknowns, the coefficients α_k and the nodes x_k. If we fix the x_k, there is a unique choice of the α_k giving a Q of maximum order of precision, which is necessarily the interpolatory procedure and has order of precision $\ge n-1$ (these α_k can be found by solving a system of linear equations; the technique is called the *method of undetermined coefficients*). The Newton-Cotes approach was of this form. We could alternately fix the α_k and find the x_k for which the procedure has maximum order of precision, or we could vary both the α_k and x_k, keeping only n fixed.

IX.2.13.29. First consider fixing the α_k. For reasons discussed in IX.2.13.4. (a more detailed analysis justifies this further), a good choice is to take all the α_k equal (to $\frac{b-a}{n}$). This problem can be solved and gives a procedure of order of precision $\ge n$ only for $n\le 7$ and $n=9$; if n is even the order of precision is actually $n+1$. The case $n=1$ gives the midpoint rule. If $[a,b]=[-1,1]$ and $n=2$, the optimal notes are at $\pm\frac{1}{\sqrt{3}} = 0.577\cdots$. Tables giving the nodes for other n can be found in [IK94] and other references.

IX.2.13.30. Now consider varying both the α_k and the x_k. It can be shown that the maximum order of precision is obtained by taking the nodes to be the roots of the unique (up to constant multiple) polynomial p_n of degree n which is orthogonal to the subspace of polynomials of degree $\le n-1$ with respect to the usual real inner product

$$\langle f,g\rangle = \int_a^b f(x)g(x)\,dx\ .$$

If $[a,b]=[-1,1]$, the p_n are the *Legendre polynomials* (); other intervals can be handled via affine transformations. The α_k are then chosen so the procedure is interpolatory. The resulting procedure is called *Gaussian quadrature*, or *Legendre-Gauss quadrature*, with n nodes, denoted G_n. The procedure G_n has order $2n-1$. Here are formulas for small $n>1$ for the interval $[-1,1]$ (the case $n=1$ is just the midpoint rule):

$$G_2(f)=f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right)=f(-.577350\cdots)+f(.577350\cdots)$$

$$G_3=\frac{5}{9}f(-\sqrt{0.6})+\frac{8}{9}f(0)+\frac{5}{9}f(\sqrt{0.6})=\frac{1}{9}[5f(-0.774597\cdots)+8f(0)+5f(0.774597\cdots)]$$

$$G_4(f)=(0.347855\cdots)f(-0.861136\cdots)+(0.652145\cdots)f(-0.339981\cdots)$$
$$+(0.652145\cdots)f(0.339981\cdots)+(0.347855\cdots)f(0.861136\cdots)$$

The coefficients and nodes are algebraic numbers, but irrational in general and must be computed by a numerical procedure; see e.g. http://mathworld.wolfram.com/Legendre-GaussQuadrature.html. There are tables available online for at least up to $n=64$.

IX.2.13.31. Advantages of Gaussian quadrature are that it is usually very accurate for modest n if f is sufficiently smooth (cf. IX.2.13.60.), and the coefficients are all positive, although not too close to the same magnitude if n is large (cf. IX.2.13.4.). Disadvantages are that it can be difficult to calculate the nodes and coefficients for n large (although there are tables for reasonable n), and if the process is repeated for a larger n all the coefficients and nodes must be recalculated (unlike, say, Simpson's rule if n is doubled). A method like this also cannot be easily applied to functions arising from data sets.

Weight Functions and Chebyshev Quadrature

IX.2.13.32. It is often important to consider a more general version of quadrature, to estimate integrals of the form

$$I(f)=\int_a^b f(x)w(x)\,dx$$

where w is a fixed weight function on $[a,b]$. Here w can be any nonnegative function which is Riemann-integrable or improperly Riemann-integrable (or, more generally, Lebesgue integrable) on $[a,b]$, but in applications w is usually a simple smooth function. We want a quadrature procedure Q of the form

$$Q(f)=\sum_{k=1}^{n}\alpha_k f(x_k)$$

for constants α_k and nodes x_k. The basic restriction now is

$$\sum_{k=1}^{n}\alpha_k=\int_a^b w(x)\,dx\ .$$

Usual quadrature is the special case where w is the constant function 1.

IX.2.13.33. Gaussian quadrature works similarly in this setting: the optimal nodes are the roots of the polynomial of degree n which is orthogonal to all polynomials of smaller degree with respect to the real inner product

$$\langle f, g\rangle = \int_a^b f(x)g(x)w(x)\,dx$$

and the weights are then found by the method of undetermined coefficients. This works because of the result of Exercise IX.2.14.5..

IX.2.13.34. More generally, a Radon measure μ () can be used on $[a, b]$, and the integrals

$$\int_{[a,b]} f\,d\mu$$

approximated by quadrature procedures as above. The optimal polynomial of degree n is the one orthogonal to the polynomials of smaller degree for the inner product

$$\langle f, g\rangle = \int_{[a,b]} fg\,d\mu\ .$$

It can be shown using the Krein-Milman theorem () that such quadrature procedures must converge for any continuous f as $n \to \infty$. However, this generality is of little or no practical use in numerical analysis.

IX.2.13.35. A particularly important case is the weight function $w(x) = \frac{1}{\sqrt{1-x^2}}$on $[-1, 1]$. Making the substitution $x = \cos\theta$, we have

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} f(\cos(\theta))\,d\theta$$

and the optimal nodes are the Chebyshev nodes (). This case is called *Chebyshev quadrature*.

Error Analysis of the Trapezoid Rule

IX.2.13.36. We begin with an analysis of the accuracy of the simple trapezoid rule

$$T_1(f) = \frac{h}{2}[f(a) + f(b)]$$

where $h = b - a$. The technique in the other error analyses will be similar. We want to calculate the error

$$E_1^T(f) = \int_a^b f(x)\,dx - T_1(f) = \int_a^b [f(x) - p_{0,1}(x)]\,dx = \int_a^b (x-a)(x-b)f[a, b, x]\,dx$$

by ().

Assume f is $\mathcal{C}^2$ on $[a, b]$. Then $g(x) = f[a, b, x]$ is continuous on $[a, b]$ (), and for each $x \in (a, b)$ we can write

$$f[a, b, x] = \frac{f''(\xi(x))}{2}$$

for some $\xi(x) \in (a,b)$ (V.11.5.9.). Since $(x-a)(x-b)$ does not change sign on $[a,b]$, and $f[a,b,x]$ is continuous on $[a,b]$, we can apply the Mean Value Theorem for integrals () to obtain

$$E_1^T(f) = \int_a^b (x-a)(x-b) f[a,b,x]\,dx = f[a,b,c] \int_a^b (x-a)(x-b)\,dx = \frac{f''(\xi)}{2} \int_a^b (x-a)(x-b)\,dx$$

for some $c \in (a,b)$, where $\xi = \xi(c) \in (a,b)$. We calculate

$$\int_a^b (x-a)(x-b) = -\frac{h^3}{6}$$

(cf. the proof of IX.2.13.24.), so we get

$$E_1^T(f) = -\frac{h^3}{12} f''(\xi)$$

for some $\xi \in (a,b)$.

IX.2.13.37. Now suppose we do the composite trapezoid rule with n subintervals, each of length $h = \frac{b-a}{n}$. Using the previous analysis on each subinterval, we have

$$E_n^T(f) = -\frac{h^3}{12} f''(\xi_1) - \frac{h^3}{12} f''(\xi_2) - \cdots - \frac{h^3}{12} = -\frac{(b-a)^3}{12n^3} \left[f''(\xi_1) + \cdots + f''(\xi_n)\right]$$

for some $\xi_k \in (x_{k-1}, x_k)$. If f is $\mathcal{C}^2$, we can apply the Intermediate Value Theorem to f'' to obtain a $\xi \in (a,b)$ with

$$f''(\xi_1) + \cdots + f''(\xi_n) = n f''(\xi)$$

so we obtain a simplified form of the error:

$$E_n^T(f) = -\frac{(b-a)^3}{12n^2} f''(\xi)$$

for some $\xi \in (a,b)$. If $|f''|$ is bounded by K on $[a,b]$, we have

$$|E_n^T(f)| \leq \frac{(b-a)^3}{12n^2} K\ .$$

IX.2.13.38. Thus, as n increases, the error in the trapezoid rule is $O(n^{-2})$, so decreases fairly rapidly. The error bound at the end of IX.2.13.37. is actually conservative in many cases: the sum

$$|f''(\xi_1) + \cdots + f''(\xi_n)|$$

is often much less than $n \cdot \max_{[a,b]} |f''|$, especially if f'' varies significantly or changes sign on $[a,b]$; in fact,

$$f''(\xi_1) + \cdots + f''(\xi_n)$$

is close to n times the *average* value of f'' on $[a,b]$, i.e.

$$n \frac{f'(b) - f'(a)}{b-a}$$

for n large. Thus

$$E_n^T(f) \approx -\frac{(b-a)^2}{12n^2} [f'(b) - f'(a)]$$

for n large.

IX.2.13.39. We can thus add a correction term to the trapezoid rule: set

$$\tilde{T}_n(f) = T_n(f) + \frac{(b-a)^2}{12n^2}[f'(b) - f'(a)] \ .$$

The error for this quadrature rule is $o(n^{-2})$, and $O(n^{-4})$ if f is $\mathcal{C}^4$. (It is not always practical to compute $f'(a)$ and $f'(b)$, e.g. if f is given only by a data set. There are approximations possible using higher-order differences which can be used as a substitute.)

IX.2.13.40. If $f'(a) = f'(b)$, e.g. if f is periodic and the integration is over an entire period, then the correction term is zero and we have that the error $E_n^T(f)$ is $o(n^{-2})$ if f is $\mathcal{C}^2$, and $O(n^{-4})$ if f is $\mathcal{C}^4$. Thus the trapezoid rule is especially accurate for periodic functions. In fact, using the Euler-Maclaurin formula () it can be shown that if f is $\mathcal{C}^r$ and periodic with period $b-a$, then there is a constant C_r depending only on r such that

$$E_n^T(f) = C_r \frac{(b-a)^{r+1}}{n^r} f^{(r)}(\xi)$$

for some $\xi \in (a,b)$, and thus $E_n^T(f)$ is $O(n^{-r})$. See [IK94] for details.

Error Analysis of the Midpoint Rule

IX.2.13.41. Let us examine the error in the simple midpoint rule

$$M_1(f) = f(x_1)h$$

where $h = b-a$ and x_1 is the midpoint of $[a,b]$. The key trick in getting a good error formula is to note that not only is M_1 the interpolatory quadrature rule for the set $\{x_1\}$, but, if f is $\mathcal{C}^1$, it is also the osculatory quadrature rule for the nodes (x_1, x_1): the osculating polynomial is $p(x) = f(x_1) + f'(x_1)(x - x_1)$, and

$$\int_a^b [f(x_1) + f'(x_1)(x-x_1)]\,dx = f(x_1)(b-a) + \frac{1}{2}f'(x_1)[(b-x_1)^2 - (a-x_1)^2]$$

and the second term is zero. Thus the error is

$$E_1^M(f) = \int_a^b f(x)\,dx - M_1(f) = \int_a^b (x-x_1)^2 f[x_1, x_1, x]\,dx \ .$$

Assume f is $\mathcal{C}^2$ on $[a,b]$. Then $g(x) = f[x_1, x_1, x]$ is continuous on $[a,b]$ so by the MVT for integrals, there is a $c \in (a,b)$ such that

$$\int_a^b (x-x_1)^2 f[x_1, x_1, x]\,dx = f[x_1, x_1, c] \int_a^b (x-x_1)^2\,dx \ .$$

We also have by () that

$$f[x_1, x_1, c] = \frac{f''(\xi)}{2}$$

for some $\xi \in (a,b)$. We calculate

$$\int_a^b (x-x_1)^2\,dx = \frac{1}{3}[(b-x_1)^3 - (a-x_1)^3] = \frac{1}{3} \cdot 2\left(\frac{h}{2}\right)^3 = \frac{h^3}{12}$$

so we have

$$E_1^M(f) = \frac{h^3}{24} f''(\xi)$$

for some $\xi \in (a, b)$.

IX.2.13.42. Now consider the composite midpoint rule with n subintervals of length $h = \frac{b-a}{n}$. If f is $\mathcal{C}^2$ on $[a, b]$, we have

$$E_n^M(f) = \frac{h^3}{24} f''(\xi_1) + \cdots + \frac{h^3}{24} f''(\xi_n) = \frac{(b-a)^3}{24n^3}[f''(\xi_1) + \cdots + f''(\xi_n)]$$

for some $\xi_k \in (x_{k-1}, x_k)$. By the Intermediate Value Theorem there is a $\xi \in (a, b)$ with $f''(\xi_1)+\cdots+f''(\xi_n) = nf''(\xi)$, so we get the clean formula

$$E_n^M(f) = \frac{(b-a)^3}{24n^2} f''(\xi)$$

for some $\xi \in (a, b)$. If $|f''|$ is bounded by K on $[a, b]$, we get

$$|E_n^M(f)| \leq \frac{(b-a)^3}{24n^2} K \ .$$

IX.2.13.43. As with the trapezoid rule, the absolute error estimate for the midpoint rule is conservative, and we may add a correction term: set

$$\tilde{M}_n(f) = M_n(f) - \frac{(b-a)^2}{24n^2}[f'(b) - f'(a)] \ .$$

This has error $O(n^{-4})$ if f is $\mathcal{C}^4$. If $f'(a) = f'(b)$, e.g. if f is periodic and the integration is over an entire period, then the correction term is zero and we have that the error $E_n^M(f)$ is $O(n^{-4})$ if f is $\mathcal{C}^4$. (In the periodic case, the midpoint rule, the trapezoid rule, and the left and right approximations are effectively all the same procedure.)

IX.2.13.44. Comparing IX.2.13.41. with IX.2.13.36. and IX.2.13.42. with IX.2.13.37., we see that if f is $\mathcal{C}^2$, the error in the midpoint rule is indeed approximately of opposite sign and half the absolute value of the error in the trapezoid rule for the same n (cf. IX.2.13.10.). The statement is not exactly true since the ξ's are not the same in general; but if n is chosen large enough that f'' varies little over each subinterval, the $f''(\xi_k)$'s are almost equal for each k, so the statement is very nearly true. Thus for such an n we have that Simpson's rule S_{2n} is an extremely good approximation to the integral. See IX.2.13.49..

IX.2.13.45. Another consequence of the error formula (which is pretty obvious just from the definition) is that even though the simple midpoint rule is an interpolatory procedure with only one node, it has order of precision 1, not 0: it gives the exact answer for linear functions, not just constant functions. The simple trapezoid rule also has order of precision 1, as expected since it is an interpolatory procedure with two nodes. The composite procedures also have order of precision 1.

Error Analysis of Simpson's Rule

IX.2.13.46. Consider the simple Simpson's rule S_2, where $h = \frac{b-a}{2}$. We first observe that S_2 actually has order of precision ≥ 3 instead of the expected 2. To show this, by linearity it suffices to show that $S_2(f) = \int_a^b f(x)\,dx$ for $f(x) = x^3$, which is a simple calculation:

$$S_2(f) = \frac{h}{3}[a^3 + 4x_1^3 + b^3] = \frac{b-a}{6}\left[a^3 + 4\left(\frac{a+b}{2}\right)^3 + b^3\right]$$

$$= \frac{b-a}{6}\left[a^3 + \frac{1}{2}(a+b)^3 + b^3\right] = \frac{1}{6}(b-a)\left[a^3 + \frac{1}{2}a^3 + \frac{3}{2}a^2b + \frac{3}{2}ab^2 + \frac{1}{2}b^3 + b^3\right]$$

$$= \frac{1}{4}(b-a)(a^3 + a^2b + ab^2 + b^3) = \frac{1}{4}(b^4 - a^4) = \int_a^b x^3\,dx\ .$$

(It is easily checked the same way that S_2 is not exact on $f(x) = x^4$, so the order of precision of S_2 is exactly 3.)

IX.2.13.47. Now suppose f is $\mathcal{C}^1$, and let p be the osculating polynomial for f with nodes x_0, x_1, x_1, x_2. Then the interpolating polynomial for p for the nodes x_0, x_1, x_2 is the same polynomial q as the interpolating polynomial for f with these same nodes. Since p is cubic, we have

$$\int_a^b p(x)\,dx = S_2(p) = \int_a^b q(x)\,dx = S_2(f)\ .$$

Thus, for $\mathcal{C}^1$ functions, S_2 is the osculating quadrature procedure with nodes x_0, x_1, x_1, x_2.

IX.2.13.48. As a result, if f is $\mathcal{C}^1$ we get a formula for the error in S_2:

$$E_2^S(f) = \int_a^b [f(x) - p(x)]\,dx = \int_a^b (x-x_0)(x-x_1)^2(x-x_2)f[x_0, x_1, x_1, x_2, x]\,dx\ .$$

We proceed as before. Suppose f is $\mathcal{C}^1$ on $[a,b]$. The function $g(x) = f[x_0, x_1, x_1, x_2, x]$ is continuous on $[a,b]$, and $(x-x_0)(x-x_1)^2(x-x_2)$ does not change sign on $[a,b]$, so by the MVT for integrals we have

$$E_2^S(f) = \int_a^b (x-x_0)(x-x_1)^2(x-x_2)f[x_0, x_1, x_1, x_2, x]\,dx = f[x_0, x_1, x_1, x_2, c]\int_a^b (x-x_0)(x-x_1)^2(x-x_2)\,dx$$

for some $c \in (a,b)$.

Next we calculate

$$\int_a^b (x-a)(x-x_1)^2(x-b)\,dx\ .$$

Make the substitution $x = x_0 + th$; then

$$\int_a^b (x-a)(x-x_1)^2(x-b)\,dx = h^5\int_0^2 t(t-1)^2(t-2)\,dt = -\frac{4h^5}{15}\ .$$

Now suppose f is $\mathcal{C}^4$ on $[a,b]$. Then

$$f[x_0,x_1,x_1,x_2,c] = \frac{f^{(4)}(\xi)}{4!}$$

for some $\xi \in (a,b)$. Thus we obtain

$$E_2^S(f) = -\frac{4h^5}{15}\frac{f^{(4)}(\xi)}{24} = -\frac{h^5}{90}f^{(4)}(\xi)$$

for some $\xi \in (a,b)$.

IX.2.13.49. Now consider the composite Simpson's rule with n subintervals of length $h = \frac{b-a}{n}$, where $n = 2m$ is even. If f is $\mathcal{C}^4$ on $[a,b]$, we have, applying S_2 on each pair of successive intervals,

$$E_n^S(f) = -\frac{h^5}{90}f^{(4)}(\xi_1) - \cdots - \frac{h^5}{90}f^{(4)}(\xi_m) = -\frac{(b-a)^5}{90n^5}[f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m)]$$

for some $\xi_k \in (x_{2k-2}, x_{2k})$. By the Intermediate Value Theorem there is a $\xi \in (a,b)$ with $f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m) = mf^{(4)}(\xi)$, so we get the clean formula

$$E_n^M(f) = -\frac{(b-a)^5}{180n^4}f^{(4)}(\xi)$$

for some $\xi \in (a,b)$. If $|f^{(4)}|$ is bounded by K on $[a,b]$, we get

$$|E_n^S(f)| \leq \frac{(b-a)^5}{180n^4}K\ .$$

IX.2.13.50. Thus, as n increases, the error in Simpson's rule is $O(n^{-4})$, so decreases rapidly. The error bound at the end of IX.2.13.49. is actually conservative in many cases: the sum

$$|f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m)|$$

is often much less than $m \cdot \max_{[a,b]} |f^{(4)}|$, especially if $f^{(4)}$ varies significantly or changes sign on $[a,b]$; in fact,

$$f^{(4)}(\xi_1) + \cdots + f^{(4)}(\xi_m)$$

is close to m times the *average* value of $f^{(4)}$ on $[a,b]$, i.e.

$$m\frac{f^{(3)}(b) - f^{(3)}(a)}{b-a}$$

for n large. Thus

$$E_n^S(f) \approx -\frac{(b-a)^4}{180n^4}[f^{(3)}(b) - f^{(3)}(a)]$$

for n large. See also IX.2.13.44..

IX.2.13.51. We can thus add a correction term to Simpson's rule: if n is even, set

$$\tilde{S}_n(f) = S_n(f) + \frac{(b-a)^4}{180n^4}[f^{(3)}(b) - f^{(3)}(a)] \ .$$

The error for this quadrature rule is $o(n^{-4})$ if f is $\mathcal{C}^4$, and $O(n^{-6})$ if f is $\mathcal{C}^6$. (It is not always practical to compute $f^{(3)}(a)$ and $f^{(3)}(b)$, e.g. if f is given only by a data set. There are approximations possible using higher-order differences which can be used as a substitute.)

IX.2.13.52. If $f^{(3)}(a) = f^{(3)}(b)$, e.g. if f is periodic and the integration is over an entire period, then the correction term is zero and we have that the error $E_n^S(f)$ is $O(n^{-6})$ if f is $\mathcal{C}^6$. Thus, like the trapezoid rule, Simpson's rule is especially accurate for periodic functions. But because of IX.2.13.40., there is little practical reason to use Simpson's rule instead of the trapezoid rule for periodic functions.

IX.2.13.53. For functions which are not smooth, the trapezoid rule is actually often a little better than Simpson's rule for the same number of subintervals. See [CUN02] for a detailed analysis of the errors for both methods for classes of nonsmooth functions.

Error Analysis for Newton-Cotes Procedures

IX.2.13.54. Error formulas for the higher-order Newton-Cotes formulas and open Newton-Cotes formulas can be obtained by essentially the same arguments, although the details are more complicated beyond Simpson's rule; see e.g. [IK94]. The result is:

IX.2.13.55. THEOREM. Let f be a function on $[a,b]$, and $n \in \mathbb{N}$. Set $h = \frac{b-a}{n}$.
(i) Suppose n is odd and f is $\mathcal{C}^{n+1}$ on $[a,b]$. Then

$$E_n^{NC}(f) = \int_a^b f(x)\,dx - NC_n(f) = \frac{M_n}{(n+1)!}h^{n+2}f^{(n+1)}(\xi)$$

for some $\xi \in (a,b)$, where

$$M_n = \int_0^n \pi_n(t)\,dt = \int_0^n t(t-1)\cdots(t-n)\,dt < 0 \ .$$

(ii) Suppose n is even and f is $\mathcal{C}^{n+2}$ on $[a,b]$. Then

$$E_n^{NC}(f) = \int_a^b f(x)\,dx - NC_n(f) = \frac{M_n}{(n+2)!}h^{n+3}f^{(n+2)}(\xi)$$

for some $\xi \in (a,b)$, where

$$M_n = \int_0^n t\pi_n(t) = \int_0^n t^2(t-1)\cdots(t-n)\,dt < 0 \ .$$

These formulas are consistent with the cases $n = 1$ (trapezoid rule) and $n = 2$ (Simpson's rule) developed in IX.2.13.36. and IX.2.13.46..

IX.2.13.56. COROLLARY. If n is even, the Newton-Cotes procedure NC_n has order of precision $n+1$; if n is odd, NC_n has order of precision n.

IX.2.13.57. Thus, if n is even, The Newton-Cotes procedure NC_n has essentially the same accuracy as NC_{n+1} (the accuracy of NC_{n+1} is roughly better by a constant factor). Thus Newton-Cotes procedures are "better" for even n. The "explanation" is that if n is even, the procedure NC_n, which is the interpolatory procedure for the evenly spaced nodes $x_0, \ldots, x_n$, turns out to be also the osculatory procedure for the $n+2$ nodes $x_0, \ldots, x_{n/2}, x_{n/2}, \ldots, x_n$, with $x_{n/2}$ repeated, as in the case of the midpoint rule and Simpson's rule.

IX.2.13.58. There are similar results for the open Newton-Cotes formulas NC_n^o, which also gain an extra order of precision for n even (e.g. for the midpoint rule NC_0^o). In this case the constants are positive.

IX.2.13.59. These error formulas are for simple Newton-Cotes procedures. Similar error formulas for compound Newton-Cotes procedures can be obtained in exactly the same manner as for the trapezoid, midpoint, and Simpson's rule cases.

Error Analysis for Gaussian Quadrature

IX.2.13.60. The error for Gaussian quadrature G_n on $[a,b]$ may be calculated in a similar manner: if f is $\mathcal{C}^{2n}$ on $[a,b]$, then

$$E_n^G(f) = \int_a^b f(x)\,dx - G_n(f) = \frac{2}{(2n+1)!}\left[\frac{2^n(n!)^2}{(2n)!}\right]^2 (b-a)^{2n+1} f^{(2n)}(\xi)$$

for some $\xi \in (a,b)$. Using Stirling's formula () to approximate the factorials, we get, for n large,

$$E_n^G(f) \approx \frac{\pi n (b-a)^{2n+1}}{2^{2n-1}(2n+1)!} f^{(2n)}(\xi)$$

for some $\xi \in (a,b)$. Thus if the derivatives of f do not grow too rapidly (e.g. if f is complex-analytic on a large enough region containing $[a,b]$), the Gaussian quadrature approximations converge rapidly as $n \to \infty$. See e.g. [IK94] for details.

IX.2.13.61. In practice, knowing when a procedure gives a sufficiently accurate estimate is not easy. There are procedures which carefully combine results of a standard procedure done twice for different partitions to greatly improve accuracy and make it easier to know when to stop. A theoretically and computationally appealing such improvement of the trapezoid rule is *Romberg integration* (Exercise IX.2.14.7.). A widely-used similar variation of Gaussian quadrature is *Gauss-Kronrod integration*.

There are also *adaptive procedures* for numerical integration, where the interval lengths and procedures are not fixed in advance but the details vary according to properties of the function being integrated. These are generally the best numerical procedures when feasible. Most software packages for numerical integration use some combination or variation of these procedures; further discussion is beyond our scope, but can be found in numerical analysis texts.

IX.2.14. Exercises

IX.2.14.1. Here are some additional relations among standard integration procedures on a fixed interval $[a, b]$.

(a) Show that $T_{2n} = \frac{1}{2}(T_n + M_n)$ for any n.

(b) Show that $S_{2n} = \frac{1}{3}(4T_{2n} - T_n)$ for any n.

IX.2.14.2. (a) Consider

$$\int_{-m}^{m} \frac{1}{x^2+1}\, dx \ .$$

Compute the exact value, and the trapezoid and midpoint estimates using one subinterval. Show that the absolute trapezoid error is less than the absolute midpoint error if $m > 1.44$, and the trapezoid estimate is *much* better than the midpoint estimate for m large.

(b) Consider

$$\int_{0}^{1} x^m\, dx \ .$$

Compute the exact value, and the trapezoid and midpoint estimates using one subinterval. Show that the midpoint estimate is *much* better than the trapezoid estimate for m large.

Actually, none of the estimates are very good in these cases, as may be expected when only one interval is used. In (b), at least the midpoint errors approach 0 as $m \to +\infty$, while the absolute trapezoid error increases as m increases. In (a), both the absolute trapezoid and midpoint errors increase as m increases, but the trapezoid errors are bounded and the midpoint errors unbounded as $m \to +\infty$.

IX.2.14.3. Consider the integral

$$\int_{-1}^{1} |x|\, dx \ .$$

Show that either the trapezoid rule or the midpoint rule for n subintervals gives the exact answer if n is even (even $n = 2$), and not if n is odd. Thus the approximations do not necessarily become monotonically better as n increases (although they do converge to the exact integral as $n \to \infty$).

This function is not differentiable everywhere. But it can be uniformly approximated arbitrarily closely by a polynomial () for which the same statements are approximately true; in particular, approximations using a relatively small even number of intervals, even 2, are better than approximations of the same kind using a larger odd number of subintervals.

It can be shown that if f is a polynomial, then the absolute error $|E_n^T(f)|$ monotonically decreases for sufficiently large n, but this is not always true if f is merely smooth. See http://mathoverflow.net/questions/180542/monotonicity-of-trapezoid-approximations.

IX.2.14.4. **Simpson's Three-Eighths Rule.** Fix an interval $[a, b]$.

(a) Show that if f is a function on $[a, b]$, then

$$NC_3(f) = \frac{3h}{8}[f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)]$$

where $h = \frac{b-a}{3}$.

(b) Show that if f is $\mathcal{C}^4$ on $[a,b]$, then

$$E_3^{NC}(f) = -\frac{3}{80}h^5 f^{(4)}(\xi)$$

for some $\xi \in (a,b)$.

(c) Compare $E_3^{NC}(f)$ with $E_2^{NC}(f) = E_2^S(f)$ (IX.2.13.46.). [Caution: the h's are different in the two cases, and the ξ's are also normally different.]

(d) If $n = 3m$ is divisible by 3, let R_n be the composite quadrature procedure obtained by dividing $[a,b]$ into m equal subintervals and applying NC_3 on each (so $[a,b]$ is divided into n equal subintervals of length $h = \frac{b-a}{n}$). Show that if f is $\mathcal{C}^4$ on $[a,b]$, then

$$E_n^R(f) = \int_a^b f(x)\,dx - R_n(f) = -\frac{3}{80}\frac{(b-a)^5}{n^5}[f^{(4)}(\xi_1) + \cdots + f(\xi_m)] = -\frac{1}{80}\frac{(b-a)^5}{n^4}f^{(4)}(\xi)$$

for some $\xi_k \in (x_{3k-3}, x_{3k})$ and some $\xi \in (a,b)$.

(e) Suppose f is $\mathcal{C}^4$ on $[a,b]$, and n is divisible by 6 and large enough that $f^{(4)}$ is almost constant on each subinterval. Show that $|E_n^S(f)|$ is "typically" smaller than $|E_n^R(f)|$.

IX.2.14.5. Let w be a nonnegative Riemann-integrable function on $[a,b]$ with $\int_a^b w(x)\,dx > 0$. Define a real inner product on the set of polynomials (with real coefficients) by

$$\langle p, q\rangle = \int_a^b p(x)q(x)w(x)\,dx\ .$$

(a) Show that for each $n \in \mathbb{N}$ there is a unique (up to scalar multiple) nonzero polynomial p of degree n which is orthogonal to all polynomials of degree $\leq n-1$ under this inner product.

(b) Suppose p changes sign exactly m times in $[a,b]$, at $x_1, \ldots, x_m$ in (a,b), i.e. $x_1, \ldots, x_m$ are the distinct roots of p of odd order in (a,b). Show that p cannot be orthogonal to

$$q(x) = (x - x_1)(x - x_2)\cdots(x - x_m)$$

($q(x) = 1$ if $m = 0$) since pq does not change sign on $[a,b]$. Thus $m = deg(q) \geq n$.

(c) Conclude that all the roots of p are real and distinct and lie in (a,b).

IX.2.14.6. [Ham86, 17.7] Consider the following quadrature procedure for twice-differentiable functions on $[-1,1]$:

$$Q(f) = \frac{1}{21}[5f(-1) + 32f(0) + 5f(1)] - \frac{1}{315}[f''(-1) + 32f''(0) + f''(1)]\ .$$

(a) Show that Q has order of precision 6.

(b) Show that Q is not an osculating procedure.

IX.2.14.7. Romberg Integration. Fix an interval $[a,b]$. Let T_n^0 be the trapezoid rule T_n for any n. Recursively define

$$T_{2n}^k = \frac{1}{4^k-1}\left(4^k T_{2n}^{k-1} - T_n^{k-1}\right)$$

for all n and k.

(a) Show that $T_{2n}^1 = S_{2n}$ for any n.

(b) Show that T_{4n}^2 is $NC_{4,n}$, the process of dividing $[a,b]$ into n subintervals of equal length and applying NC_4 on each subinterval.

(c) The simple relation of (a) and (b) with the Newton-Cotes formulas does not persist. Show that

$$T_8^3 = \frac{4h}{2835}\left(217f(x_0) + 1024f(x_1) + 352f(x_2) + 1024f(x_3) + 436f(x_4) + 1024f(x_5) + 352f(x_6) + 1024f(x_7) + 217f(x_8)\right)$$

($h = \frac{b-a}{8}$) which is not the same as NC_8 (however, the coefficients of T_8^3 are all positive, which is desirable computationally, and the sizes of the coefficients do not vary greatly).

The Romberg integration procedure is as follows. Fix an n (in practice $n = 1$ is often used). Successively calculate $T_{2^m n}^k$ for $k \le 2^m$. The results can be conveniently collected in a triangular array

$$\begin{array}{cccc} T_n^0 & & & \\ T_{2n}^0 & T_{2n}^1 & & \\ T_{4n}^0 & T_{4n}^1 & T_{4n}^2 & \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

where each entry is computed from the one to the left and the one diagonally above to the left (the entries in the first column are computed from the trapezoid rule using additional function evaluations; note that $T_{2^{m+1}n}^0$ can be calculated from $T_{2^m n}^0$ by only $2^m n$ additional function evaluations). In practice the entries are rounded to the desired accuracy and the process stops when two identical numbers to the desired accuracy are obtained. The process is computationally efficient and frequently produces the desired accuracy with fewer function evaluations than Simpson's Rule. The convergence rate is comparable to the rate using correction terms (IX.2.13.39.), and is more computationally efficient and avoids making a careful estimate of the starting n needed. See books on Numerical Analysis for details and error analyses.

IX.3. Jordan Content and Riemann Integration in $\mathbb{R}^n$

Generalization of Riemann integration to $\mathbb{R}^n$ is in many ways straightforward, but there are some significant technicalities to overcome. One is to nail down what kinds of sets functions can be reasonably integrated over, and the answer is considerably more complicated than in the one-dimensional case where intervals are the obvious answer. There is a natural collection of subsets called *Jordan regions* which work nicely for Riemann integration and which are general enough to cover most cases of practical interest. As a special case of integration, we also have a natural notion of "volume," called *Jordan content*, for Jordan regions.

IX.3.1. Integration over Rectangles

It is most straightforward to integrate over rectangles in $\mathbb{R}^n$ whose edges are parallel to the axes. The following definition will be used throughout this section (a more general definition is used in ()).

IX.3.1.1. Definition. A *rectangle* in $\mathbb{R}^n$ is a set R of the form $I_1 \times \cdots \times I_n$, where each I_k is a closed bounded interval $[a_k, b_k]$ or degenerate interval (if $a_k = b_k$) in $\mathbb{R}$, i.e.

$$R = \{(x_1, \ldots, x_n) : a_k \leq x_k \leq b_k \text{ for all } k\} .$$

The rectangle is *degenerate* if I_k is degenerate for at least one k. The *content* (*n-dimensional content*), or *volume* (*n-dimensional volume*) of R is

$$V(R) = \prod_{k=1}^{n} \ell(I_k) = \prod_{k=1}^{n} (b_k - a_k) .$$

Such rectangles should more properly be called closed coordinate rectangles or closed aligned rectangles, since there are more general sets which can be called rectangles. However, rectangles as defined in IX.3.1.1. are exclusively used in developing Riemann integration and it is pedantic to have to constantly qualify them.

IX.3.1.2. We use the term "rectangle" in any number of dimensions even though it is only geometrically accurate if $n = 2$; in $\mathbb{R}^3$ a rectangle by our definition is what is usually called a rectangular solid, and for $n > 3$ there is not a standard geometric name. Note that the case $n = 1$ is not excluded; a rectangle in $\mathbb{R}$ is just a closed bounded interval or single point. Similarly, "volume" is only strictly appropriate if $n = 3$, which is why we prefer the term "content": if $n = 2$ content is usually called "area" (and for $n = 1$ it is "length"). We will, however, use the notation $V(R)$ which suggests the term "volume" no matter what n is. To emphasize the dimension, we can (but rarely will) write V_n in place of V.

Note that the n-dimensional content of a degenerate rectangle is always 0. A degenerate rectangle is geometrically m-dimensional for some $m < n$ (the number of nondegenerate coordinate intervals) and so has an m-dimensional "content"; thus we must exercise some care in fixing n and interpreting "content" to be n-dimensional content for this n.

It is a matter of taste whether to regard $\emptyset$ as a rectangle; whether we do or not necessarily leads to some notational complications, which are greater if we do. We will not.

The next observation is obvious from the fact that the intersection of two intervals in $\mathbb{R}$ is an interval or degenerate interval.

IX.3.1.3. PROPOSITION. If R and S are rectangles in $\mathbb{R}^n$, and $R \cap S \neq \emptyset$, then $R \cap S$ is also a rectangle.

PROOF: If $R = I_1 \times \cdots \times I_n$ and $S = J_1 \times \cdots \times J_n$, then $R \cap S = (I_1 \cap J_1) \times \cdots \times (I_n \cap J_n)$. 👍

IX.3.1.4. DEFINITION. Let R and S be rectangles in $\mathbb{R}^n$. Then R and S are *nonoverlapping* if either $R \cap S = \emptyset$ or $R \cap S$ is a degenerate rectangle.

Partitions and Grids

IX.3.1.5. DEFINITION. Let R be a rectangle in $\mathbb{R}^n$. A *partition* of R is a set $\mathcal{P} = \{R_1, \ldots, R_m\}$ of pairwise nonoverlapping rectangles in $\mathbb{R}^n$ with $R = \cup_{k=1}^m R_k$.

A partition $\mathcal{Q}$ is a *refinement* of $\mathcal{P}$ if every rectangle in $\mathcal{Q}$ is contained in a rectangle of $\mathcal{P}$.

A *grid* in R is a partition of $R = I_1 \times \cdots \times I_n$ of the form $\mathcal{G} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$, where $\mathcal{P}_k$ is a partition of I_k.

Note that a partition of a rectangle is not necessarily a grid. See Figure IX.19. It is convenient to allow degenerate rectangles in partitions, as our definition does; thus a partition of an interval in $\mathbb{R}$ by this definition is slightly more general than the definition in (), since repeated points are allowed. This will make no practical difference but simplifies some notation. We could develop the whole theory just using grids, but it is convenient to allow partitions which are not grids. Every partition has a refinement which is a grid (see the proof of IX.3.1.6.).

By IX.3.1.3., any two partitions of a rectangle have a common refinement obtained by intersecting the subrectangles. If the partitions are grids, so is the common refinement constructed in this way.

Note that there is no natural ordering of the rectangles in a partition, even a grid, if $n > 1$. We have notationally fixed an ordering, but the ordering is not part of the definition of the partition.

Here is a technicality which may be considered geometrically "obvious" (but see the discussion in I.6.1.3.); however, it needs proof (if $n > 1$). This is one of the fundamental observations which gets Jordan content and Riemann integration theory in $\mathbb{R}^n$ off the ground.

IX.3.1.6. PROPOSITION. Let R be a rectangle in $\mathbb{R}^n$. Suppose R is a union of a finite number of pairwise nonoverlapping subrectangles $R = \cup_{k=1}^m R_k$. Then

$$V(R) = \sum_{k=1}^{m} V(R_k) \ .$$

To prove IX.3.1.6., we will first prove a special case where the subdivision is a grid:

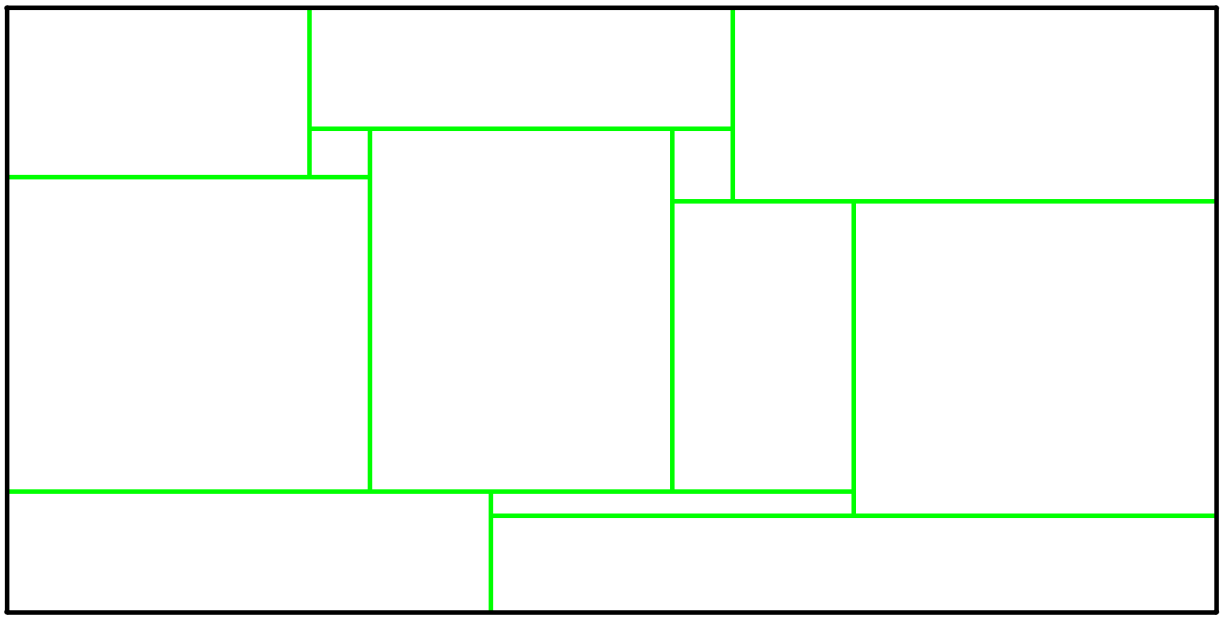


Figure IX.19: Subdivision of a Rectangle

IX.3.1.7. LEMMA. Let $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ be a rectangle in $\mathbb{R}^n$. For each k let $\mathcal{P}_k$ be a partition of $[a_k, b_k]$, and let $\mathcal{G} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_n$ be the corresponding grid. Let $\{R_1, \ldots, R_p\}$ be the subrectangles in the grid. Then

$$V(R) = \sum_{k=1}^{p} V(R_k) \ .$$

PROOF: We write the proof for $n = 2$; the proof for general $n > 1$ is nearly identical but notationally offensive. So we have $R = [a_1, b_1] \times [a_2, b_2]$, $\mathcal{P}_1 = \{c_0, \ldots, c_r\}$, $\mathcal{P}_2 = \{d_0, \ldots, d_s\}$. Set

$$R_{jk} = [c_{j-1}, c_j] \times [d_{k-1}, d_k]$$

so

$$V(R_{jk}) = (c_j - c_{j-1})(d_k - d_{k-1}) \ .$$

We then have

$$\sum_{j,k} V(R_{jk}) = \sum_{j=1}^{r} \left[\sum_{k=1}^{s} (c_j - c_{j-1})(d_k - d_{k-1}) \right]$$

$$= \sum_{j=1}^{r} \left[(c_j - c_{j-1}) \sum_{k=1}^{s} (d_k - d_{k-1}) \right] = \sum_{j=1}^{r} (c_j - c_{j-1})(b_2 - a_2)$$

$$= (b_2 - a_2)\sum_{j=1}^{r}(c_j - c_{j-1}) = (b_2 - a_2)(b_1 - a_1) = V(R) \ .$$



We now give the proof of IX.3.1.6..

PROOF: Let $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ and $R_k = [a_{k,1}, b_{k,1}] \times \cdots \times [a_{k,n}, b_{k,n}]$. For each j, let $\mathcal{P}_j$ be the partition of $[a_j, b_j]$ consisting of all the $a_{k,j}$ and $b_{k,j}$. Let $\mathcal{G}$ be the corresponding grid, and let $\{R_{j_1,\ldots,j_n}\}$ be the rectangles in the grid. Then, by IX.3.1.7.,

$$V(R) = \sum_{j_1,\ldots,j_n} V(R_{j_1,\ldots,j_n}) \ .$$

For each k, set

$$I_k = \{(j_1, \ldots, j_n) : R_{j_1,\ldots,j_n} \subseteq R_k\} \ .$$

Every $(j_1, \ldots, j_n)$ for which $R_{j_1,\ldots,j_n}$ is nondegenerate is in exactly one I_k. For each k, the $R_{j_1,\ldots,j_n}$ for $(j_1, \ldots, j_n) \in I_k$ form a grid subdivision of R_k, so again by IX.3.1.7. we have

$$V(R_k) = \sum_{(j_1,\ldots,j_n)\in I_k} V(R_{j_1,\ldots,j_n})$$

for each k, and thus we have

$$V(R) = \sum_{j_1,\ldots,j_n} V(R_{j_1,\ldots,j_n}) = \sum_{k=1}^{m}\left[\sum_{(j_1,\ldots,j_n)\in I_k} V(R_{j_1,\ldots,j_n})\right] = \sum_{k=1}^{m} V(R_k) \ .$$



See Figure IX.20.

Upper and Lower Integrals over a Rectangle

We now proceed just as we did to define the integral of a bounded function over an interval in $\mathbb{R}$:

IX.3.1.8. DEFINITION. Let R be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a bounded function. If $\mathcal{P} = \{R_1, \ldots, R_m\}$ is a partition of R, for each k set $M_k = \sup\{f(\boldsymbol{x}) : \boldsymbol{x} \in R_k\}$ and $m_k = \inf\{f(\boldsymbol{x}) : \boldsymbol{x} \in R_k\}$. Then the *upper and lower sums for f with respect to $\mathcal{P}$* are

$$\overline{\mathcal{S}}(f, \mathcal{P}) = \sum_{k=1}^{m} M_k \cdot V(R_k)$$

$$\underline{\mathcal{S}}(f, \mathcal{P}) = \sum_{k=1}^{m} m_k \cdot V(R_k) \ .$$

This definition agrees with the one in () if $n = 1$. Note that it does not make sense unless f is bounded.

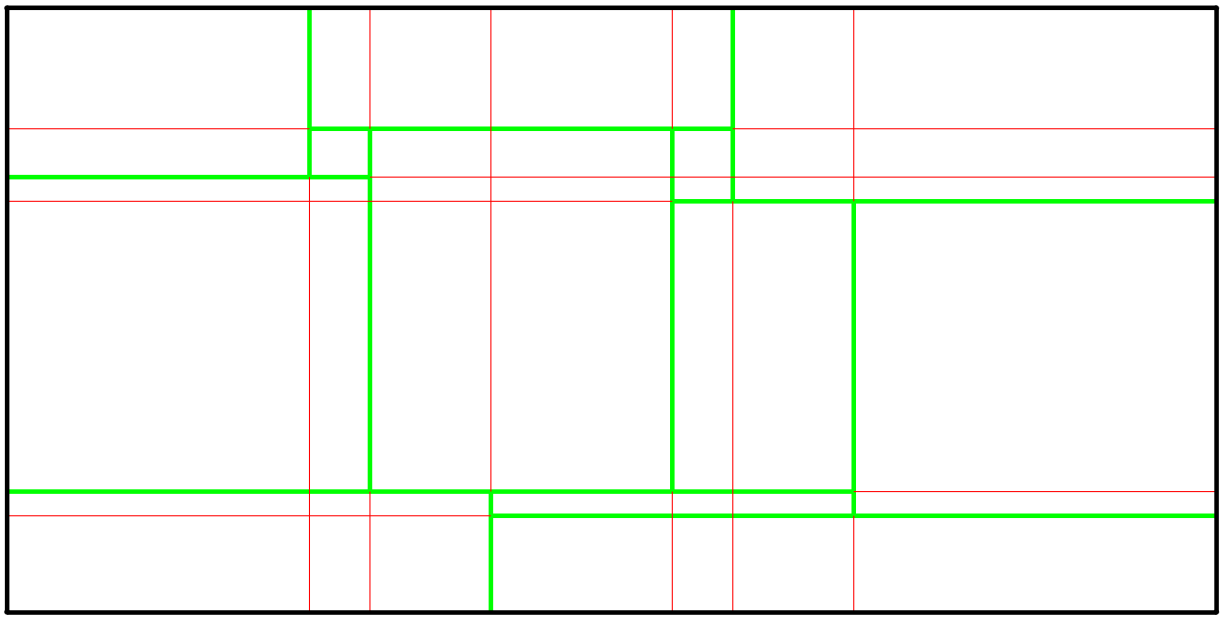


Figure IX.20: Grid Refinement of a Rectangle Subdivision

IX.3.1.9. It is clear that if $\mathcal{Q}$ is a refinement of $\mathcal{P}$, then

$$\underline{\mathcal{S}}(f,\mathcal{P}) \le \underline{\mathcal{S}}(f,\mathcal{Q}) \le \overline{\mathcal{S}}(f,\mathcal{Q}) \le \overline{\mathcal{S}}(f,\mathcal{P})$$

for any bounded function f. Thus the following definition is reasonable:

IX.3.1.10. Definition. Let R be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a bounded function. The *upper and lower integrals* of f over R are

$$\overline{\int_R} f\,dV = \inf_{\mathcal{P}} \overline{\mathcal{S}}(f,\mathcal{P})$$

$$\underline{\int_R} f\,dV = \sup_{\mathcal{P}} \underline{\mathcal{S}}(f,\mathcal{P})\ .$$

The function f is *(Riemann) integrable over* R if $\overline{\int_R} f\,dV = \underline{\int_R} f\,dV$; in this case, the common value is the *integral of* f *over* R, denoted $\int_R f\,dV$.

IX.3.1.11. If $n = 1$, this definition is consistent with (): If $R = [a,b]$ and $f : R \to \mathbb{R}$ is bounded, then f is integrable over R if and only if it is Riemann integrable in the sense of (), and we have

$$\int_R f\,dV = \int_{[a,b]} f\,dV = \int_a^b f(x)\,dx\ .$$

IX.3.1.12. Of course we have $\underline{\int_R} f\,dV \le \overline{\int_R} f\,dV$ for any R and f. As in () we have:

IX.3.1.13. Proposition. Let R be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a bounded function. Then f is integrable on R if and only if, for every $\epsilon > 0$, there is a partition $\mathcal{P}$ of R such that

$$\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) < \epsilon \ .$$

Reasonable functions turn out to be Riemann integrable. Most importantly, we have the following higher-dimensional version of (), which is proved almost identically. We omit the proof since we will obtain a more general result in IX.3.2.21. with almost the same proof.

IX.3.1.14. Theorem. Let R be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a continuous function. Then f is integrable over R.

However, functions which are too discontinuous are not integrable:

IX.3.1.15. Example. Let A be a subset of $\mathbb{R}^n$ which is dense and has dense complement, e.g. the set of points with rational coordinates. If R is any rectangle in $\mathbb{R}^n$ and $\mathcal{P}$ is any partition of R, then $\overline{\mathcal{S}}(\chi_A, \mathcal{P}) = V(R)$ and $\underline{\mathcal{S}}(\chi_A, \mathcal{P}) = 0$. Thus $\overline{\int_R} \chi_A\,dV = V(R)$ and $\underline{\int_R} \chi_A\,dV = 0$, so χ_A is not integrable on R unless R is degenerate.

Additivity of Upper and Lower Integrals

IX.3.1.16. Proposition. Let R be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a bounded function. If $\mathcal{P} = \{R_1, \ldots, R_m\}$ is a partition of R, then we have

$$\overline{\int_R} f\,dV = \sum_{k=1}^{m} \overline{\int_{R_k}} f\,dV \text{ and } \underline{\int_R} f\,dV = \sum_{k=1}^{m} \underline{\int_{R_k}} f\,dV \ .$$

Proof: In calculating $\overline{\int_R} f\,dV$, only partitions of R refining $\mathcal{P}$ need be considered. Each such partition $\mathcal{Q}$ splits into partitions $\mathcal{Q}_k$ of R_k for each k, so we obtain

$$\overline{\mathcal{S}}(f, \mathcal{Q}) = \sum_{k=1}^{m} \overline{\mathcal{S}}(f, \mathcal{Q}_k) \ .$$

Conversely, if $\mathcal{Q}_k$ is a partition of R_k for each k, the union of the $\mathcal{Q}_k$ is a partition of R refining $\mathcal{P}$. So the infimum over partitions can be taken separately in each term. The argument for $\underline{\int_R} f\,dV$ is essentially identical. 👍

IX.3.1.17. Also note that if R is a degenerate rectangle and f is any bounded function on R, then f is integrable on R and $\int_R f\,dV = 0$.

Linearity of the Integral

IX.3.1.18. PROPOSITION. Let R be a rectangle in $\mathbb{R}^n$, $f : R \to \mathbb{R}$ a bounded function, and $\alpha \in \mathbb{R}$. Then

$$\overline{\int_R} \alpha f\,dV = \begin{cases} \alpha \overline{\int_R} f\,dV & \text{if} \quad \alpha \geq 0 \\ \alpha \underline{\int_R} f\,dV & \text{if} \quad \alpha < 0 \end{cases}$$

$$\underline{\int_R} \alpha f\,dV = \begin{cases} \alpha \underline{\int_R} f\,dV & \text{if} \quad \alpha \geq 0 \\ \alpha \overline{\int_R} f\,dV & \text{if} \quad \alpha < 0 \end{cases}$$

In particular, if f is integrable over R, then so is αf, and

$$\int_R \alpha f\,dV = \alpha \int_R f\,dV\ .$$

PROOF: This is obvious from the fact that if $\mathcal{P} = \{R_1, \ldots, R_m\}$ is any partition of R, then, for $1 \leq k \leq m$, $M_k(\alpha f) = \alpha M_k(f)$ if $\alpha \geq 0$ and $M_k(\alpha f) = \alpha m_k(f)$ if $\alpha < 0$, so that $\overline{\mathcal{S}}(\alpha f, \mathcal{P}) = \alpha \overline{\mathcal{S}}(f, \mathcal{P})$ if $\alpha \geq 0$ and $\overline{\mathcal{S}}(\alpha f, \mathcal{P}) = \alpha \underline{\mathcal{S}}(f, \mathcal{P})$ if $\alpha < 0$, and similarly for the other. 👍

IX.3.1.19. PROPOSITION. Let R be a rectangle in $\mathbb{R}^n$, and f and g bounded functions on R. Then

$$\overline{\int_R} (f+g)\,dV \leq \overline{\int_R} f\,dV + \overline{\int_R} g\,dV$$

$$\underline{\int_R} (f+g)\,dV \geq \underline{\int_R} f\,dV + \underline{\int_R} g\,dV\ .$$

In particular, if f and g are integrable over R, then so is $f + g$, and

$$\int_R (f+g)\,dV = \int_R f\,dV + \int_R g\,dV\ .$$

PROOF: This is similar to the previous proposition, noting that if $\mathcal{P} = \{R_1, \ldots, R_m\}$ is any partition of R, then, for $1 \leq k \leq m$, $M_k(f+g) \leq M_k(f) + M_k(g)$ and $m_k(f+g) \geq m_k(f) + m_k(g)$ (). 👍

IX.3.1.20. Thus the integral over a fixed rectangle is linear: if f and g are integrable over a rectangle R, and $\alpha, \beta \in \mathbb{R}$, then $\alpha f + \beta g$ is integrable over R and

$$\int_R (\alpha f + \beta g)\, dV = \alpha \int_R f\, dV + \beta \int_R g\, dV \ .$$

However, the upper and lower integrals are not linear for functions which are not integrable. For example, A is a dense set with dense complement, then for any nondegenerate R we have

$$V(R) = \overline{\int_R} 1\, dV = \overline{\int_R} (\chi_A + \chi_{A^c})\, dV < \overline{\int_R} \chi_A\, dV + \overline{\int_R} \chi_{A^c}\, dV = 2V(R)$$

and similarly for the lower integral for $-\chi_A - \chi_{A^c}$.

Order Properties of the Integral

The order properties of the integral are obvious (consider partitions):

IX.3.1.21. Proposition. Let R be a rectangle in $\mathbb{R}^n$, and f and g bounded functions on R. If $f \le g$ on R, then $\overline{\int_R} f\, dV \le \overline{\int_R} g\, dV$ and $\underline{\int_R} f\, dV \le \underline{\int_R} g\, dV$.

IX.3.1.22. Corollary. Let R be a rectangle in $\mathbb{R}^n$, and f a bounded function on R. If $m \le f(\boldsymbol{x}) \le M$ for all $\boldsymbol{x} \in R$, then

$$m \cdot V(R) \le \underline{\int_R} f\, dV \le \overline{\int_R} f\, dV \le M \cdot V(R) \ .$$

In particular, if $f \ge 0$ on R, then $\underline{\int_R} f\, dV \ge 0$.

Integrability of Products

Although the result is a special case of the general characterization of integrable functions in (), we can give a simple proof that a product of integrable functions is integrable.

IX.3.1.23. Lemma. Let R be a rectangle in $\mathbb{R}^n$, and f a bounded function on R. If $|f(\boldsymbol{x})| \le B$ for all $\boldsymbol{x} \in R$, then, for any partition $\mathcal{P} = \{R_1, \ldots, R_m\}$ of R,

$$\overline{\mathcal{S}}(f^2, \mathcal{P}) - \underline{\mathcal{S}}(f^2, \mathcal{P}) \le 2B[\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P})] \ .$$

Proof: For $1 \le k \le m$, we have, for any $\boldsymbol{x}, \boldsymbol{y} \in R_k$,

$$f(\boldsymbol{x})^2 - f(\boldsymbol{y})^2 = (f(\boldsymbol{x}) + f(\boldsymbol{y}))(f(\boldsymbol{x}) - f(\boldsymbol{y}))$$

$$|f(\boldsymbol{x})^2 - f(\boldsymbol{y})^2| = |f(\boldsymbol{x}) + f(\boldsymbol{y})||f(\boldsymbol{x}) - f(\boldsymbol{y})| \le 2B|f(\boldsymbol{x}) - f(\boldsymbol{y})|$$

and therefore $M_k(f^2) - m_k(f^2) \le 2B(M_k(f) - m_k(f))$, and the result follows by summing over k.

IX.3.1.24. Corollary. Let R be a rectangle in $\mathbb{R}^n$, and f a bounded function on R. If $|f(\boldsymbol{x})| \le B$ for all $\boldsymbol{x} \in R$, then

$$\overline{\int_R} f^2\,dV - \underline{\int_R} f^2\,dV \le 2B\left[\overline{\int_R} f\,dV - \underline{\int_R} f\,dV\right] .$$

In particular, if f is integrable over $\mathbb{R}$, then so is f^2.

IX.3.1.25. Proposition. Let R be a rectangle in $\mathbb{R}^n$, and f and g bounded functions on R. If f and g are integrable over R, then so is fg.

Proof: We have $f+g$ integrable over R by IX.3.1.19., and thus f^2, g^2, and $(f+g)^2$ are integrable over R by IX.3.1.24.. Since

$$fg = \frac{1}{2}\left[(f+g)^2 - f^2 - g^2\right]$$

we have that fg is integrable over R by IX.3.1.20.. 👍

Expanding the Rectangle

We now obtain a technical result which will be important when extending the notion of Riemann integral from rectangles to more general regions. If R is a rectangle, S is a subrectangle, and f is integrable on S, extend f to $\mathbb{R}$ by setting $f(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in R \setminus S$. Then the integral of the extended function over R should be the same as the integral of f over S. It is:

IX.3.1.26. Theorem. Let R be a rectangle in $\mathbb{R}^n$, and S a subrectangle of R. Let $f : S \to \mathbb{R}$ be a bounded function. Define $g : R \to \mathbb{R}$ by

$$g(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & \text{if} \quad \boldsymbol{x} \in S \\ 0 & \text{if} \quad \boldsymbol{x} \in R \setminus S \end{cases} .$$

Then

$$\overline{\int_R} g\,dV = \overline{\int_S} f\,dV \text{ and } \underline{\int_R} g\,dV = \underline{\int_S} f\,dV .$$

In particular, g is integrable over R if and only if f is integrable over S, and in this case

$$\int_R g\,dV = \int_S f\,dV .$$

There would seem to be a simple proof: consider partitions of R consisting of a partition of S and some additional rectangles on which f is zero. But the problem is that the rectangles in a partition are only nonoverlapping, not disjoint, so some of the additional rectangles will contain points of ∂S at which f may well be nonzero. Thus we must work harder to prove the theorem, taking partitions for which all the rectangles covering ∂S but not in S have a small total content.

IX.3.1.27. The following construction will be used in the proof. Let $S = [a_1, b_1] \times \cdots \times [a_n, b_n]$ be a rectangle in $\mathbb{R}^n$. If $\delta > 0$, set

$$S_\delta = [a_1 - \delta, b_1 + \delta] \times \cdots \times [a_n - \delta, b_n + \delta] \ .$$

Then S_δ is a rectangle in $\mathbb{R}^n$ containing S in its interior. $V(S_\delta)$ is a polynomial in δ, so by continuity, for any $\epsilon > 0$ there is a $\delta > 0$ such that $V(S) < V(S_\delta) < V(S) + \epsilon$.

We now give the proof of IX.3.1.26..

PROOF: We may assume R contains S in its interior, since if not two applications of the theorem for a larger rectangle containing both R and S in its interior give the result. Then $S_\delta \subseteq R$ for all sufficiently small $\delta > 0$. Note that if $S_\delta \subseteq R$, any partition of S can be extended to a partition of R in which S_δ as well as S is a union of rectangles in the partition.

Fix $\epsilon > 0$. Let B be a bound for $|f|$ on S. Choose $\delta > 0$ so that $S_\delta \subseteq R$ and $V(S_\delta) < V(S) + \frac{\epsilon}{3B}$. If $\mathcal{P}$ is any partition of S, extend $\mathcal{P}$ to a partition $\mathcal{Q}$ of R for which S_δ is a union of rectangles in $\mathcal{Q}$. Let $\mathcal{Q}_1$ be the set of rectangles in $\mathcal{Q}$ contained in S_δ but not in S, and $\mathcal{Q}_2$ the set of rectangles in $\mathcal{Q}$ not contained in S_δ. Then

$$\overline{\mathcal{S}}(g, \mathcal{Q}) = \sum_{k=1}^{m} M_k \cdot V(R_k) = \sum_{R_k \in \mathcal{P}} M_k \cdot V(R_k) + \sum_{R_k \in \mathcal{Q}_1} M_k \cdot V(R_k) + \sum_{R_k \in \mathcal{Q}_2} M_k \cdot V(R_k) \ .$$

The first sum is exactly $\overline{\mathcal{S}}(f, \mathcal{P})$, and the last sum is 0 since g is identically 0 on the rectangles in $\mathcal{Q}_2$. If $R_k \in \mathcal{Q}_2$, $|M_k| \leq B$, so the second sum in absolute value is dominated by

$$\sum_{R_k \in \mathcal{Q}_2} B \cdot V(R_k) = B(V(S_\delta) - V(S)) < \frac{\epsilon}{3}$$

and therefore

$$|\overline{\mathcal{S}}(f, \mathcal{P}) - \overline{\mathcal{S}}(g, \mathcal{Q})| < \frac{\epsilon}{3}$$

i.e. for every partition $\mathcal{P}$ of S there is a partition $\mathcal{Q}$ of R for which this holds. On the other hand, if $\tilde{\mathcal{Q}}$ is any refinement of $\mathcal{Q}$, then $\tilde{\mathcal{Q}}$ is of this form for some refinement $\tilde{\mathcal{P}}$ of $\mathcal{P}$. We can thus choose $\mathcal{P}$ and $\mathcal{Q}$ so that

$$\overline{\mathcal{S}}(f, \mathcal{P}) - \overline{\int_S} f \, dV < \frac{\epsilon}{3}$$

$$\overline{\mathcal{S}}(g, \mathcal{Q}) - \overline{\int_R} g \, dV < \frac{\epsilon}{3}$$

from which it follows that

$$\left| \overline{\int_S} f \, dV - \overline{\int_S} g \, dV \right| < \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, the upper integrals are equal. The argument for the lower integrals is almost identical. 👍

See IX.3.12.2. for an alternate version of the proof.

IX.3.2. Content and Jordan Regions

Using the integral over rectangles, we can say what the "volume" is of more general subsets of Euclidean space.

IX.3.2.1. DEFINITION. Let E be a bounded subset of $\mathbb{R}^n$. The *upper n-dimensional (Jordan) content* of E is

$$\overline{V_n}(E) = \overline{\int_R \chi_E \, dV}$$

where R is a rectangle in $\mathbb{R}^n$ containing E.
The *lower n-dimensional (Jordan) content* of E is

$$\underline{V_n}(E) = \underline{\int_R} \chi_E \, dV$$

where R is a rectangle in $\mathbb{R}^n$ containing E.
E is a *Jordan region* if $\overline{V_n}(E) = \underline{V_n}(E)$; the *(Jordan) content* $V_n(E)$ is the common value.

IX.3.2.2. Note that the upper and lower content are well defined and independent of the containing rectangle R by IX.3.1.26.. We usually suppress the words "n-dimensional" and the subscript n, but the dimension must be carefully kept in mind. Some authors use the term "volume" instead of (Jordan) content. See IX.3.1.2..

IX.3.2.3. A bounded subset E of $\mathbb{R}^n$ is a Jordan region if and only if χ_E is integrable over some (hence every by IX.3.1.26.) rectangle R containing E, and $V(E) = \int_R \chi_E \, dV$.

IX.3.2.4. PROPOSITION. If R is a rectangle in $\mathbb{R}^n$, then R is a Jordan region and its Jordan content is $V(R)$, as defined in IX.3.1.1..

PROOF: We can take the containing rectangle to be R itself. Then χ_R is the constant function 1 on R. By IX.3.1.22. (or by simple direct calculation), χ_R is integrable on R and $\int_R \chi_R \, dV = \int_R 1 \, dV = V(R)$. 👍

Thus our notation and terminology is consistent.

IX.3.2.5. A subset of $\mathbb{R}^n$ is a Jordan region if and only if it boundary has zero content (IX.3.2.24.). "Geometrically reasonable" subsets of $\mathbb{R}^n$ are thus Jordan regions. But a set like the intersection of the set A of IX.3.1.15. with a rectangle is not a Jordan region. Note that Jordan regions are by definition bounded.

IX.3.2.6. If A and B are bounded subsets of $\mathbb{R}^n$ with $A \subseteq B$, then $\overline{V}(A) \leq \overline{V}(B)$ and $\underline{V}(A) \leq \underline{V}(B)$ (cf. IX.3.1.22.).

The next observation will be considerably improved in IX.3.2.25., but this form is needed to logically develop the theory and characterization of Jordan regions:

IX.3.2.7. LEMMA. Let A and B be Jordan regions in $\mathbb{R}^n$. Then $A \cup B$ is a Jordan region if and only if $A \cap B$ is a Jordan region. In this case, we have

$$V(A \cup B) = V(A) + V(B) - V(A \cap B) \ .$$

PROOF: We have $\chi_{A\cup B} = \chi(A) + \chi(B) - \chi(A \cap B)$. Let R be a rectangle containing both A and B. By assumption, $\chi(A)$ and $\chi(B)$ are integrable over R, so it follows from IX.3.1.20. that $\chi_{A\cup B}$ is integrable over R if and only if $\chi_{A\cap B}$ is. The volume formula also follows from IX.3.1.20.. 👍

IX.3.2.8. COROLLARY. If $\{R_1, \ldots, R_m\}$ is a finite set of rectangles in $\mathbb{R}^n$, not necessarily nonoverlapping, then $\cup_{k=1}^m R_k$ is a Jordan region and

$$V\left(\bigcup_{k=1}^{m} R_k\right) \leq \sum_{k=1}^{m} V(R_k)$$

with equality if and only if the R_k are nonoverlapping.

PROOF: For $m = 2$ the result follows immediately from IX.3.2.4. and IX.3.2.7. since the intersection of two rectangles is a rectangle. The result for general m follows by induction. 👍

Along the same lines, the next observation is needed for the logical development although it will be improved in IX.3.2.25.:

IX.3.2.9. PROPOSITION. Let E be a Jordan region, and R a rectangle containing E. Then $R \setminus E$ is also a Jordan region, and $V(R \setminus E) = V(R) - V(E)$.

PROOF: We have that $\chi_{R\setminus E} = \chi_R - \chi_E$; χ_E is integrable over R by assumption, and χ_R is integrable over R by IX.3.2.4., so $\chi_{R\setminus E}$ is integrable over R and the content formula holds by IX.3.1.20.. 👍

An Alternate Description of Content

We can give an illuminating direct description of content by explicitly describing upper and lower sums for the integral of an indicator function.

IX.3.2.10. Let E be a bounded subset of $\mathbb{R}^n$. If R is a rectangle in $\mathbb{R}^n$ containing E and $\mathcal{P} = \{R_1, \ldots, R_m\}$ is a partition of R, we have

$$\overline{\mathcal{S}}(\chi_E, \mathcal{P}) = \sum_{R_k \cap E \neq \emptyset} V(R_k)$$

$$\underline{\mathcal{S}}(\chi_E, \mathcal{P}) = \sum_{R_k \subseteq E} V(R_k)$$

i.e. $\overline{V}(E)$ is estimated as the content of the union of the rectangles in $\mathcal{P}$ containing points of E, and $\underline{V}(E)$ is estimated as the content of the union of the rectangles in $\mathcal{P}$ entirely contained in E; in fact, they are the infimum and supremum of such contents over all partitions of R.

Combining with IX.3.2.8., we obtain:

IX.3.2.11. PROPOSITION. If E is a bounded subset of $\mathbb{R}^n$, then

$$\underline{V}(E) = \sup\left(\sum_{k=1}^{m} V(R_k)\right)$$

where the supremum is over all finite sets $\{R_1, \ldots, R_m\}$ of pairwise nonoverlapping rectangles contained in E.

For upper sums, since any finite set of rectangles has a subdivision into nonoverlapping rectangles, we can relax the requirement that the rectangles be nonoverlapping:

IX.3.2.12. PROPOSITION. If E is a bounded subset of $\mathbb{R}^n$, then

$$\overline{V}(E) = \inf\left(\sum_{k=1}^{m} V(R_k)\right)$$

where the infimum is over all finite sets $\{R_1, \ldots, R_m\}$ of rectangles covering E.

Sets of Zero Content

IX.3.2.13. DEFINITION. A subset E of $\mathbb{R}^n$ has *zero content* if $\overline{V}(E) = 0$.
Two subsets A and B of $\mathbb{R}^n$ are *nonoverlapping* if $A \cap B$ has zero content.

IX.3.2.14. If E has zero content, it is a Jordan region since $0 \leq \underline{V}(E) \leq \overline{V}(E) = 0$. Any subset of a set of zero content also has zero content. A rectangle has zero content if and only if it is degenerate. Thus two rectangles are nonoverlapping in the sense of IX.3.2.13. if and only if they are nonoverlapping in the sense of IX.3.1.4..

The next result is then an immediate consequence of IX.3.2.7.:

IX.3.2.15. PROPOSITION. Let $E_1, \ldots, E_m$ be pairwise nonoverlapping Jordan regions. Then $\cup_{k=1}^{m} E_k$ is a Jordan region and

$$V\left(\bigcup_{k=1}^{m} E_k\right) = \sum_{k=1}^{m} V(E_k) \ .$$

IX.3.2.16. Corollary. A finite union of sets of zero content has zero content.

The next result gives a large supply of sets of zero content:

IX.3.2.17. Proposition. Let E be a compact Jordan region in $\mathbb{R}^{n-1}$, $\phi : E \to \mathbb{R}$ a continuous function, and G the graph of ϕ, i.e.

$$G = \{(x_1, \ldots, x_n) : (x_1, \ldots, x_{n-1}) \in E, x_n = \phi(x_1, \ldots, x_{n-1})\} \subseteq \mathbb{R}^n \ .$$

Then G is a set of zero content in $\mathbb{R}^n$.

Proof: Let R be a rectangle in $\mathbb{R}^n$ containing E. Let $\epsilon > 0$. Since E is compact, ϕ is uniformly continuous on E, so there is a $\delta > 0$ such that $|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})| < \frac{\epsilon}{V(R)}$ whenever $\boldsymbol{x}, \boldsymbol{y} \in E$ and $|\boldsymbol{x} - \boldsymbol{y}| < \delta$. There is a partition $\mathcal{P} = \{R_1, \ldots, R_m\}$ of R such that the diameter of each R_k is less than δ; thus if M_k and m_k are the supremum and infimum of ϕ over $R_k \cap E$ for each R_k containing points of E, we have $M_k - m_k \leq \frac{\epsilon}{V(R)}$ for all such k. Then if

$$\tilde{R}_k = R_k \times [m_k, M_k]$$

for each such k, then the $\tilde{R}_k$ are nonoverlapping rectangles covering G, and

$$\sum_k V(\tilde{R}_k) \leq \sum_k V(R_k) \cdot \frac{\epsilon}{V(R)} \leq \epsilon$$

(the sum is over all k for which $R_k \cap E \neq \emptyset$). Thus $\overline{V_n}(G) \leq \epsilon$. Since $\epsilon > 0$ is arbitrary, $\overline{V_n}(G) = 0$. 👍

IX.3.2.18. If E is a compact set in $\mathbb{R}^n$ and every point of E has a neighborhood in E which is a set of zero content in $\mathbb{R}^n$, then a simple compactness argument and IX.3.2.16. shows that E is a set of zero content. In particular, if E is compact and locally the graph of a continuous function of some $n-1$ of the coordinates, then E is a set of zero content. Many sets can be shown to be of this form by the Implicit Function Theorem.

IX.3.2.19. It is crucial to specify the n when saying that a set has zero content in $\mathbb{R}^n$. If we regard $\mathbb{R}^n$ as a subset of $\mathbb{R}^{n+1}$ by identifying it with an affine subspace, every bounded subset of $\mathbb{R}^n$ has zero content in $\mathbb{R}^{n+1}$. In fact, the notion of Jordan region in $\mathbb{R}^n$ depends on n: any bounded subset of $\mathbb{R}^n$ is thus a Jordan region in $\mathbb{R}^{n+1}$. So the property of E being a Jordan region is not intrinsic to E, but depends on the choice of Euclidean space containing E (and also how E is embedded in that space). ✋

IX.3.2.20. Proposition. Let A and B be bounded regions in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. If A has zero content in $\mathbb{R}^p$, then $A \times B$ has zero content in $\mathbb{R}^{p+q}$.

Proof: Let S be a bounded rectangle in $\mathbb{R}^q$ containing B. Let $\epsilon > 0$. There is a finite set $\{R_1, \ldots, R_n\}$ in $\mathbb{R}^p$ covering A with

$$\sum_{k=1}^{n} V(R_k) < \frac{\epsilon}{V(S) + 1} \ .$$

Then $\{R_1 \times S, \ldots, R_n \times S\}$ is a finite set of rectangles covering $A \times B$, and $V(R_k \times S) = V(R_k)V(S)$, so

$$\overline{V}(A \times B) \leq \sum_{k=1}^{n} V(R_k \times S) = V(S) \sum_{k=1}^{n} V(R_k) < \epsilon \ .$$



The next theorem is a key result about integrability, generalizing IX.3.1.14.. This is, however, not the ultimate characterization of integrability, which will be given in ().

IX.3.2.21. THEOREM. Let R be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a bounded function. If f is continuous except on a set of zero content, then f is integrable over R.

PROOF: Suppose f is continuous except on a set A with $V(A) = 0$. We will use the following notation. If $\mathcal{P} = \{R_1, \ldots, R_m\}$ is a partition of R, set $J = \{j : R_j \cap A \neq \emptyset\}$ and $K = \{k : R_k \cap A = \emptyset\}$.

Let M be a bound for $|f|$ on R. Let $\epsilon > 0$. Then, since A has volume 0, there is a partition $\mathcal{P}$ of R such that

$$\sum_{j \in J} V(R_j) < \frac{\epsilon}{4M} \ .$$

Let $B = \cup_{k \in K} R_k$. Then B is compact and f is continuous on B, hence uniformly continuous, so there is a $\delta > 0$ such that $|f(\boldsymbol{x}) - f(\boldsymbol{y})| < \frac{\epsilon}{2V(R)}$ whenever $\boldsymbol{x}, \boldsymbol{y} \in B$ and $|\boldsymbol{x} - \boldsymbol{y}| < \delta$. By subdividing the R_k for $k \in K$ if necessary, we may assume that the diameter of each R_k is less than δ. Thus if M_k and m_k are the supremum and infimum of f over R_k, we have $M_k - m_k \leq \frac{\epsilon}{2V(R)}$. Furthermore, for each $j \in J$ we have $M_j - m_j \leq 2M$. Thus

$$\overline{\mathcal{S}}(f, \mathcal{P}) - \underline{\mathcal{S}}(f, \mathcal{P}) = \sum_{j \in J} (M_j - m_j)V(R_j) + \sum_{k \in K} (M_k - m_k)V(R_k)$$

$$\leq 2M \sum_{j \in J} V(R_j) + \frac{\epsilon}{2V(R)} \sum_{k \in K} V(R_k) \leq 2M \frac{\epsilon}{4M} + \frac{\epsilon}{2V(R)} V(R) = \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, f is integrable over R. 

Characterization of Jordan Regions

The first result is a corollary of IX.3.2.21.

IX.3.2.22. COROLLARY. Let E be a bounded set in $\mathbb{R}^n$. If ∂E has zero content, then E is a Jordan region.

PROOF: The indicator function χ_E is discontinuous exactly on ∂E. 

IX.3.2.23. PROPOSITION. If E is a Jordan region in $\mathbb{R}^n$, then $\bar{E}$ and E^o are also Jordan regions, and

$$V(\bar{E}) = V(E^o) = V(E) \ .$$

PROOF: Let R be a rectangle containing E. Let $\epsilon > 0$, and let $\mathcal{P} = \{R_1, \ldots, R_m\}$ be a partition of R such that

$$\sum_{j \in J} V(R_j) < V(E) + \epsilon$$

where J is the set of j for which $R_j \cap E \neq \emptyset$. Let $B = \cup_{j \in J} R_j$. Then B is a closed set and $E \subseteq B$, so $\bar{E} \subseteq B$, and B is a Jordan region by IX.3.2.8.. We have

$$V(E) = \underline{V}(E) \leq \underline{V}(\bar{E}) \leq \overline{V}(\bar{E}) \leq \overline{V}(B) = V(B) = \sum_{j \in J} V(R_j) < V(E) + \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, we have

$$V(E) = \underline{V}(E) = \underline{V}(\bar{E}) = \overline{V}(\bar{E})$$

so $\bar{E}$ is a Jordan region and $V(\bar{E}) = V(E)$.

By IX.3.2.9., $R \setminus E$ is a Jordan region, so $\overline{R \setminus E} = R \setminus E^o$ is a Jordan region, and hence E^o is a Jordan region; and we have

$$V(R) - V(E^o) = V(R \setminus E^o) = V(\overline{R \setminus E}) = V(R \setminus E) = V(R) - V(E)$$

and it follows that $V(E^o) = V(E)$. 👍

We now come to the main theorem characterizing Jordan regions.

IX.3.2.24. THEOREM. Let E be a bounded subset of $\mathbb{R}^n$. Then E is a Jordan region in $\mathbb{R}^n$ if and only if ∂E has zero content.

PROOF: If ∂E has zero content, then E is a Jordan region by IX.3.2.22.. Conversely, suppose E is a Jordan region. Then $\bar{E}$ and E^o are also Jordan regions, and $V(\bar{E}) = V(E^o) = V(E)$ by IX.3.2.23.. Let R be a rectangle containing E. If $\mathcal{P} = \{R_1, \ldots, R_m\}$ is a partition of R, we can partition $\{1, \ldots, m\}$ into three sets I, J, K, where I is the set of i such that $R_i \subseteq E^o$, J is the set of j such that $R_j \subseteq R \setminus \bar{E}$ (i.e. such that $R_j \cap \bar{E} = \emptyset$), and K is the set of k such that $R_k \cap \partial E \neq \emptyset$.

Let $\epsilon > 0$. If $\mathcal{P}$ is sufficiently fine, then by IX.3.2.11. and IX.3.2.9. we have

$$\sum_{i \in I} V(R_i) > \underline{V}(E^o) - \frac{\epsilon}{2} = V(E) - \frac{\epsilon}{2}$$

$$\sum_{j \in J} V(R_j) > \underline{V}(R \setminus \bar{E}) - \frac{\epsilon}{2} = V(R) - V(E) - \frac{\epsilon}{2} \ .$$

$$\sum_{i \in I} V(R_i) + \sum_{j \in J} V(R_j) > \left[V(E) - \frac{\epsilon}{2}\right] + \left[V(R) - V(E) - \frac{\epsilon}{2}\right] = V(R) - \epsilon \ .$$

By IX.3.1.6., we have

$$V(R) = \sum_{i\in I} V(R_i) + \sum_{j\in J} V(R_j) + \sum_{k\in K} V(R_k)$$

so we conclude

$$\sum_{k\in K} V(R_k) < \epsilon \ .$$

If we let $B = \cup_{k\in K} R_k$, then $\partial E \subseteq B$, so by IX.3.2.8.

$$\overline{V}(\partial E) \leq \overline{V}(B) = V(B) = \sum_{k\in K} V(R_k) < \epsilon$$

and since $\epsilon > 0$ is arbitrary, $\overline{V}(\partial E) = 0$. 👍

IX.3.2.25. Corollary. Let A and B be Jordan regions in $\mathbb{R}^n$. Then $A \cup B$, $A \cap B$, and $A \setminus B$ are Jordan regions, and

$$V(A \cup B) = V(A) + V(B) - V(A \cap B)$$
$$V(A \setminus B) = V(A) - V(A \cap B) \ .$$

Proof: By IX.3.2.24., ∂A and ∂B have zero content. We have $\partial(A \cup B) \subseteq \partial A \cup \partial B$, and $\partial A \cup \partial B$ has zero content by IX.3.2.16.; thus $\partial(A \cup B)$ has zero content and $A \cup B$ is a Jordan region by IX.3.2.24.. It then follows from IX.3.2.7. that $A \cap B$ is a Jordan region and the first content formula holds. If R is a rectangle containing A and B, then $R \setminus B$ is a Jordan region by IX.3.2.9., so $A \setminus B = A \cap (R \setminus B)$ is a Jordan region. We have

$$V(A) = V(A \cap B) + V(A \setminus B)$$

since $A \cap B$ and $A \setminus B$ are disjoint. 👍

IX.3.2.26. Corollary. Let A and B be Jordan regions in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. Then $A \times B$ is a Jordan region in $\mathbb{R}^{p+q}$.

Proof: We have $\partial(A \times B) = (\partial A \times B) \cup (A \times \partial B)$, which is a set of zero content by IX.3.2.20.. 👍

IX.3.3. Integration over Jordan Regions

It is now a simple matter to extend the integral to general Jordan regions.

IX.3.3.1. Definition. Let E be a Jordan region in $\mathbb{R}^n$, and $f : E \to \mathbb{R}$. Denote by $f\chi_E$ the function from $\mathbb{R}^n$ to $\mathbb{R}$ defined by

$$f\chi_E(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & \text{if} \quad \boldsymbol{x} \in E \\ 0 & \text{if} \quad \boldsymbol{x} \notin E \end{cases} \ .$$

This definition is a slight abuse of notation since f is only defined on E, so $f\chi_E$ is technically not a product; but this ambiguity will cause no trouble.

IX.3.3.2. DEFINITION. Let E be a Jordan region in $\mathbb{R}^n$, and $f : E \to \mathbb{R}$ a function. Define

$$\underline{\int_E} f\,dV = \underline{\int_R} f\chi_E\,dV \quad \text{and} \quad \overline{\int_E} f\,dV = \overline{\int_R} f\chi_E\,dV$$

where R is any rectangle containing E (cf. IX.3.1.26.). The function f is *(Riemann) integrable over* E if $f\chi_E$ is integrable over some (hence every) rectangle containing E, i.e. if

$$\underline{\int_E} f\,dV = \overline{\int_E} f\,dV\ .$$

If f is integrable over E, define

$$\int_E f\,dV = \underline{\int_E} f\,dV = \overline{\int_E} f\,dV\ .$$

Note that an integrable function is necessarily bounded.

IX.3.3.3. Comparing definitions, the constant function 1 is integrable over E, and $\int_E 1\,dV = V(E)$. (This example indicates that it is not reasonable to try to similarly define the integral of a function over a set which is not a Jordan region, since even constant functions will not then be integrable).

IX.3.3.4. PROPOSITION. Let A and B be Jordan regions in $\mathbb{R}^n$, with $B \subseteq A$. If $f : A \to \mathbb{R}$ is integrable on A, then $f|_B$ is integrable on B.

PROOF: We have $f\chi_B = (f\chi_A)\chi_B$ (this is an honest product). If R is a rectangle containing A, then $f\chi_A$ and χ_B are integrable over R by assumption, so $(f\chi_A)\chi_B$ is integrable over R by IX.3.1.25.. 👍

The next extension of IX.3.2.21. gives a large supply of integrable functions including almost all functions of interest in applications:

IX.3.3.5. THEOREM. Let E be a Jordan region in $\mathbb{R}^n$, and $f : E \to \mathbb{R}$ a bounded function. If f is continuous except on a set of zero content, then f is integrable over E.

PROOF: Suppose f is continuous on E except on a subset A of zero content. If R is a rectangle containing E, then $f\chi_E$ is continuous on R except (possibly) on $A \cup \partial E$, which is a set of zero content (IX.3.2.16.). Thus $f\chi_E$ is integrable over R by IX.3.2.21.. 👍

Note that even if f is continuous on E, IX.3.2.21. is needed to show that f is integrable on E; IX.3.1.14. is not enough.

Properties of the Integral

The standard properties of the integral over Jordan regions then follow immediately from the properties of integrals over rectangles (note e.g. that $(f+g)\chi_E = f\chi_E + g\chi_E$, etc.)

IX.3.3.6. PROPOSITION. The integral over a fixed Jordan region is linear: if f and g are integrable over a Jordan region E, and $\alpha, \beta \in \mathbb{R}$, then $\alpha f + \beta g$ is integrable over E and

$$\int_E (\alpha f + \beta g)\, dV = \alpha \int_E f\, dV + \beta \int_E g\, dV \ .$$

IX.3.3.7. PROPOSITION. Let E be a Jordan region in $\mathbb{R}^n$, and f and g bounded functions on E. If $f \leq g$ on E, then

$$\int_E f\, dV \leq \int_E g\, dV \ .$$

IX.3.3.8. COROLLARY. Let E be a Jordan region in $\mathbb{R}^n$, and f a bounded function on E. If $m \leq f(\boldsymbol{x}) \leq M$ for all $\boldsymbol{x} \in E$, then

$$m \cdot V(E) \leq \int_E f\, dV \leq M \cdot V(E) \ .$$

In particular, if $f \geq 0$ on E, then $\int_E f\, dV \geq 0$.

IX.3.3.9. PROPOSITION. Let E be a set of zero content, and $f : E \to \mathbb{R}$ a bounded function. Then f is integrable over E, and $\int_E f\, dV = 0$.

PROOF: Suppose $m \leq f(\boldsymbol{x}) \leq M$ for all $\boldsymbol{x} \in E$. If R is a rectangle containing E, then $m\chi_E \leq f\chi_E \leq M\chi_E$ on R, so

$$0 = mV(E) = \int_R m\chi_E\, dV \leq \underline{\int_R} f\chi_E\, dV \leq \overline{\int_R} f\chi_E\, dV \leq \int_R M\chi_E\, dV = 0 \ .$$

👍

IX.3.3.10. PROPOSITION. Let A and B be Jordan regions in $\mathbb{R}^n$, and $f : A \cup B \to \mathbb{R}$ an integrable function. Then f is integrable over A, B, and $A \cap B$, and

$$\int_{A \cup B} f\, dV = \int_A f\, dV + \int_B f\, dV - \int_{A \cap B} f\, dV \ .$$

PROOF: We have f integrable over A, B, $A \cap B$ by IX.3.3.4., and the formula then follows by linearity. 👍

IX.3.3.11. Corollary. Let $E_1, \dots, E_m$ be pairwise nonoverlapping Jordan regions in $\mathbb{R}^n$, and $E = E_1 \cup \cdots \cup E_n$. If $f : E \to \mathbb{R}$ is a function, then f is integrable over E if and only if $f|_{E_k}$ is integrable over E_k for each k. If this is the case, then

$$\int_E f\,dV = \sum_{k=1}^{m} \int_{E_k} f\,dV\ .$$

IX.3.4. Fubini's Theorem

IX.3.4.1. The definition of the integral works more or less the same for $\mathbb{R}^n$, $n > 1$, as it does for $\mathbb{R}$. But there is one dramatic difference: so far we have no effective way of calculating integrals in the multivariable case, i.e. there is no Fundamental Theorem of Calculus in the vector case (actually there is, Stokes' Theorem (), but this is usually not useful for computing the integral).

IX.3.4.2. Many authors use different notation for multivariable integrals, e.g. writing

$$\iint_R f\,dA \quad (n = 2)$$

$$\iiint_R f\,dV \quad (n = 3)$$

$$\int \cdots \int_R f\,dV \quad (n \text{ integral signs})$$

in general for $\int_R f\,dV$. I prefer the single integral sign notation, since the integral is a single integral over a multidimensional set; to emphasize the dimension, we can write $\int_R f\,dV_n$. But the multiple integral notation is suggestive of a possible way to compute the integral: as an n-fold repeated one-dimensional integral, each of which can (potentially) be calculated using the Fundamental Theorem of Calculus. Fubini's Theorem (the Riemann integral form) says that this can be done under reasonable hypotheses.

Iterated Integrals

IX.3.4.3. If $A \subseteq \mathbb{R}^p$ and $B \subseteq \mathbb{R}^q$, we can regard $A \times B$ as a subset of $\mathbb{R}^{p+q} \cong \mathbb{R}^p \times \mathbb{R}^q$. If f is a bounded real-valued function on $A \times B$, for each $\boldsymbol{x} \in A$ we have a function $f_{[\boldsymbol{x}]} : B \to \mathbb{R}$ defined by $f_{[\boldsymbol{x}]}(\boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y})$. If A and B are Jordan regions and we set

$$g(\boldsymbol{x}) = \overline{\int_B} f_{[\boldsymbol{x}]}\,dV$$

then g is a bounded function from A to $\mathbb{R}$; we symbolically write

$$\overline{\int_A} g\,dV = \overline{\int_A} \left[\overline{\int_B} f(\boldsymbol{x}, \boldsymbol{y})\,dV(\boldsymbol{y})\right] dV(\boldsymbol{x}) = \overline{\int_A \int_B} f(\boldsymbol{x}, \boldsymbol{y})\,dV(\boldsymbol{y})\,dV(\boldsymbol{x})$$

(the brackets are typically omitted), and similarly

$$h(\boldsymbol{x}) = \underline{\int_B} f_{[\boldsymbol{x}]}\, dV$$

$$\underline{\int_A} h\, dV = \underline{\int_A \int_B} f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) \ .$$

In the same way, for each $\boldsymbol{y} \in B$ we have a function $f^{[\boldsymbol{y}]} : A \to \mathbb{R}$ defined by $f^{[\boldsymbol{y}]}(\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{y})$, and if

$$g(\boldsymbol{y}) = \overline{\int_A} f^{[\boldsymbol{y}]}\, dV$$

we write

$$\overline{\int_B} g\, dV = \overline{\int_B \int_A} f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{x})\, dV(\boldsymbol{y})$$

and if

$$h(\boldsymbol{y}) = \underline{\int_A} f^{[\boldsymbol{y}]}\, dV$$

$$\underline{\int_B} h\, dV = \underline{\int_B \int_A} f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{x})\, dV(\boldsymbol{y}) \ .$$

These are called *iterated upper and lower integrals* (we could in principle also define $\overline{\int_A}\underline{\int_B} f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x})$, etc., but these are rarely if ever used). If all the functions involved are integrable, we can write

$$\int_A \int_B f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) \text{ or } \int_B \int_A f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{x})\, dV(\boldsymbol{y}) \ .$$

These are *iterated integrals*.

Fubini's Theorem for Rectangles

IX.3.4.4. THEOREM. [FUBINI'S THEOREM, UPPER AND LOWER FORM] Let R be a rectangle in $\mathbb{R}^p$ and S a rectangle in $\mathbb{R}^q$, so that $R \times S$ is a rectangle in $\mathbb{R}^{p+q}$. Let $f : R \times S \to \mathbb{R}$ be a bounded function. Then

$$\underline{\int_{R\times S}} f\, dV \le \underline{\int_R \int_S} f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) \le \overline{\int_R \int_S} f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) \le \overline{\int_{R\times S}} f\, dV \ .$$

To emphasize the dimensions of the integrals, we can write

$$\underline{\int_{R\times S}} f\, dV_{p+q} \le \underline{\int_R \int_S} f(\boldsymbol{x},\boldsymbol{y})\, dV_q(\boldsymbol{y})\, dV_p(\boldsymbol{x}) \le \overline{\int_R \int_S} f(\boldsymbol{x},\boldsymbol{y})\, dV_q(\boldsymbol{y})\, dV_p(\boldsymbol{x}) \le \overline{\int_{R\times S}} f\, dV_{p+q} \ .$$

IX.3.4.5. Note that the order of R and S is arbitrary: we equally obtain

$$\underline{\int_{R\times S}} f\,dV \le \underline{\int_S\int_R} f(\boldsymbol{x},\boldsymbol{y})\,dV(\boldsymbol{x})\,dV(\boldsymbol{y}) \le \overline{\int_S\int_R} f(\boldsymbol{x},\boldsymbol{y})\,dV(\boldsymbol{x})\,dV(\boldsymbol{y}) \le \overline{\int_{R\times S}} f\,dV$$

by interchanging R and S.

Proof: We prove the third inequality; the proof of the first is almost identical, and the second inequality is trivial. So, for $\boldsymbol{x}\in R$, set

$$g(\boldsymbol{x}) = \overline{\int_S} f_{[\boldsymbol{x}]}\,dV = \overline{\int_S} f(\boldsymbol{x},\boldsymbol{y})\,dV(\boldsymbol{y})\ .$$

We must prove

$$\overline{\int_R} g\,dV \le \overline{\int_{R\times S}} f\,dV\ .$$

Since every grid partition of $R\times S$ is of the form $\mathcal{P}\times\mathcal{Q}$ for some partitions $\mathcal{P}$ of R and $\mathcal{Q}$ of S, it suffices to show that

$$\overline{\mathcal{S}}(g,\mathcal{P}) \le \overline{\mathcal{S}}(f,\mathcal{P}\times\mathcal{Q})$$

for any such $\mathcal{P}$ and $\mathcal{Q}$, since then the desired inequality is obtained by taking the infimum over all $\mathcal{P}$ and $\mathcal{Q}$.

So let $\mathcal{P}=\{R_1,\dots,R_m\}$ and $\mathcal{Q}=\{S_1,\dots,S_n\}$ be partitions of R and S. For $1\le j\le m$ and $1\le k\le n$, let $M_{jk}(f)$ be the supremum of f over $R_j\times S_k$, and let $M_j(g)$ be the supremum of g over R_j. We then have

$$\overline{\mathcal{S}}(f,\mathcal{P}\times\mathcal{Q}) = \sum_{j=1}^m\sum_{k=1}^n M_{jk}(f)V(R_j)V(S_k)$$

$$\overline{\mathcal{S}}(g,\mathcal{P}) = \sum_{j=1}^m M_j(g)V(R_j)$$

so it suffices to show that for each j,

$$M_j(g) \le \sum_{k=1}^n M_{jk}(f)V(S_k)\ .$$

Fix j. For $\boldsymbol{x}\in R_j$,

$$g(\boldsymbol{x}) = \overline{\int_S} f_{[\boldsymbol{x}]}\,dV \le \overline{\mathcal{S}}(f_{[\boldsymbol{x}]},\mathcal{Q}) = \sum_{k=1}^n M_k(f_{[\boldsymbol{x}]})V(S_k)$$

where

$$M_k(f_{[\boldsymbol{x}]}) = \sup_{\boldsymbol{y}\in S_k} f(\boldsymbol{x},\boldsymbol{y}) \le M_{jk}(f)$$

since $\boldsymbol{x}\in R_j$. Thus

$$g(\boldsymbol{x}) \le \sum_{k=1}^n M_{jk}(f)V(S_k)$$

and, taking the supremum $M_j(g)$ over $\boldsymbol{x}\in R_j$, the desired inequality follows. 👍

The next result, an immediate corollary of IX.3.4.4., is the case of most interest:

IX.3.4.6. Corollary. [Fubini's Theorem, Riemann Integral Version] Let R be a rectangle in $\mathbb{R}^p$ and S a rectangle in $\mathbb{R}^q$, so that $R \times S$ is a rectangle in $\mathbb{R}^{p+q}$. Let $f : R \times S \to \mathbb{R}$ be a bounded function.
(i) Suppose f is integrable over $R \times S$ and, for each $\boldsymbol{x} \in R$, $f_{[\boldsymbol{x}]}$ is integrable over S. If

$$g(\boldsymbol{x}) = \int_S f_{[\boldsymbol{x}]}\, dV$$

for $\boldsymbol{x} \in R$, then g is integrable over R, and

$$\int_{R\times S} f\, dV = \int_R g\, dV = \int_R \int_S f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) \ .$$

(ii) Suppose f is integrable over $R \times S$ and, for each $\boldsymbol{y} \in S$, $f^{[\boldsymbol{y}]}$ is integrable over R. If

$$h(\boldsymbol{y}) = \int_R f^{[\boldsymbol{y}]}\, dV$$

for $\boldsymbol{y} \in S$, then h is integrable over S, and

$$\int_{R\times S} f\, dV = \int_S h\, dV = \int_S \int_R f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{x})\, dV(\boldsymbol{y}) \ .$$

(iii) If both (i) and (ii) hold (i.e. if f and each $f_{[\boldsymbol{x}]}$ and $f^{[\boldsymbol{y}]}$ are integrable), then

$$\int_R \int_S f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) = \int_S \int_R f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{x})\, dV(\boldsymbol{y}) \ \left(= \int_{R\times S} f\, dV \right) \ .$$

Proof: (i): From the Theorem, we have

$$\int_{R\times S} f\, dV \le \underline{\int_R} \int_S f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) \le \overline{\int_R} \int_S f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) \le \int_{R\times S} f\, dV$$

so the inequalities are equalities, and the middle equality exactly says that g is integrable over R. The proof of (ii) is almost identical, and (iii) is just a combination of (i) and (ii). 👍

A clean and especially important special case (cf. IX.3.1.14., XX.10.3.5.(c)) is:

IX.3.4.7. Corollary. [Fubini's Theorem, Continuous Version] Let R be a rectangle in $\mathbb{R}^p$ and S a rectangle in $\mathbb{R}^q$, so that $R \times S$ is a rectangle in $\mathbb{R}^{p+q}$. Let $f : R \times S \to \mathbb{R}$ be a continuous function. Then

$$\int_{R\times S} f\, dV = \int_R \int_S f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{y})\, dV(\boldsymbol{x}) = \int_S \int_R f(\boldsymbol{x},\boldsymbol{y})\, dV(\boldsymbol{x})\, dV(\boldsymbol{y}) \ .$$

In particular, the two iterated integrals are equal.

IX.3.4.8. This result (IX.3.4.6.) can be interpreted, and used, in two different ways. On the one hand, it says that (under certain conditions) the integral over $R \times S$ can be computed by an iterated integral. On the other hand, (iii) says that (under certain conditions) the order of integration can be reversed in an iterated integral. Both interpretations are important in different contexts; for example, it is sometimes possible to evaluate an iterated integral in one order but not the other using the FTC (cf. IX.3.12.5.).

IX.3.4.9. Actually these versions of "Fubini's Theorem" may have been "known" before FUBINI, who proved a general version for Lebesgue integration (XVIII.3.1.). All theorems involving changing the order of iterated summation or integration are now generically called "Fubini's Theorem" since they are special cases of XVIII.3.1.14.. See IX.3.4.34. for further comments.

If $p = q = 1$, IX.3.4.7. can be rephrased, changing notation slightly:

IX.3.4.10. THEOREM. [FUBINI'S THEOREM IN $\mathbb{R}^2$, RECTANGLE CONTINUOUS VERSION] Let $R = [a, b] \times [c, d]$ be a rectangle in $\mathbb{R}^2$, and $f : R \to \mathbb{R}$ a continuous function. Then

$$\int_R f \, dV = \iint_R f \, dA = \int_a^b \int_c^d f(x, y) \, dy \, dx = \int_c^d \int_a^b f(x, y) \, dx \, dy \ .$$

In particular, the two iterated integrals are equal.

IX.3.4.11. In calculating the inner integral (e.g. by the FTC), the outer variable is treated as a constant. Thus the inner integral can be regarded as a "partial antiderivative."

IX.3.4.12. If p or q is greater than 1, the iterated integrals can be further broken into iterated integrals of smaller dimension by repeatedly applying the various versions of Fubini's Theorem. It is cumbersome to explicitly state the appropriate hypotheses for the higher iterated version of IX.3.4.6., but IX.3.4.7. can be cleanly stated. We only state the ultimate reduced version as in IX.3.4.10.:

IX.3.4.13. THEOREM. [FUBINI'S THEOREM IN $\mathbb{R}^n$, RECTANGLE CONTINUOUS VERSION] Let $R = [a_1, b_1] \times \cdots \times [a_n, b_n]$ be a rectangle in $\mathbb{R}^n$, and $f : R \to \mathbb{R}$ a continuous function. Then

$$\int_R f \, dV = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n) \, dx_n \cdots dx_2 \, dx_1 \ .$$

The iterated integrals can be done in any order.

Iterated Integrals over Other Regions

In applications we frequently want to integrate over regions more general than rectangles. In good cases such integrals can also be calculated as iterated integrals. Note that even to do this when integrating a continuous function requires the version of Fubini's Theorem in IX.3.4.6., i.e. IX.3.4.7. is not enough.

We will not try to do this in the utmost generality, which requires cumbersome notation and hypotheses; instead, we will restrict to clean forms which are adequate for almost all applications. The first version is straightforward:

IX.3.4.14. THEOREM. Let A and B be Jordan regions in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively (so $A \times B$ is a Jordan region in $\mathbb{R}^{p+q}$). If $f : A \times B \to \mathbb{R}$ is bounded and continuous, then

$$\int_{A\times B} f\,dV = \int_A \int_B f(\boldsymbol{x},\boldsymbol{y})\,dV(\boldsymbol{y})\,dV(\boldsymbol{x}) = \int_B \int_A f(\boldsymbol{x},\boldsymbol{y})\,dV(\boldsymbol{x})\,dV(\boldsymbol{y})$$

(in particular, the iterated integrals exist and are equal).

PROOF: $A \times B$ is a Jordan region (IX.3.2.26.). Let R and S be rectangles containing A and B respectively, and apply IX.3.4.6. to $f\chi_{A\times B}$ on $R \times S$ (it is easily checked that the hypotheses apply). ☝

IX.3.4.15. COROLLARY. Let A and B be Jordan regions in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively (so $A \times B$ is a Jordan region in $\mathbb{R}^{p+q}$). Then $V(A \times B) = V(A)V(B)$.

To handle regions more general than products, we begin with subsets of $\mathbb{R}^2$:

IX.3.4.16. DEFINITION. A subset E of $\mathbb{R}^2$ is *projectable* (with respect to the order x, y) if there are intervals $[a, b]$ and continuous functions $\phi, \psi : [a, b] \to \mathbb{R}$ with $\phi \le \psi$ and

$$E = \{(x, y) : a \le x \le b,\ \phi(x) \le y \le \psi(x)\}\ .$$

IX.3.4.17. See Figure IX.21 (adapted from http://tutorial.math.lamar.edu/Classes/CalcIII/DIGeneralRegion.aspx). We can also define projectability in the order y, x: E is projectable in the order y, x if there is an interval $[c, d]$ and functions $\phi, \psi : [c, d] \to \mathbb{R}$ such that

$$E = \{(x, y) : c \le y \le d,\ \phi(y) \le x \le \psi(y)\}\ .$$

A set E is *projectable* if it is projectable in at least one order; some simple regions are projectable in either order, and others projectable in one order but not the other. Most regions in $\mathbb{R}^2$ one would want to integrate over are projectable, or at least a finite union of nonoverlapping projectable regions. Note that a projectable region is compact.

IX.3.4.18. The boundary of a projectable region

$$E = \{(x, y) : a \le x \le b,\ \phi(x) \le y \le \psi(x)\}$$

consists of the graphs of ϕ and ψ, and the vertical line segments

$$\{(a, y) : \phi(a) \le y \le \psi(a)\} \text{ and } \{(b, y) : \phi(b) \le y \le \psi(b)\}$$

(one or both vertical segments can be degenerate), all of which have volume 0 in $\mathbb{R}^2$ by IX.3.2.17.. Thus E is a Jordan region in $\mathbb{R}^2$.

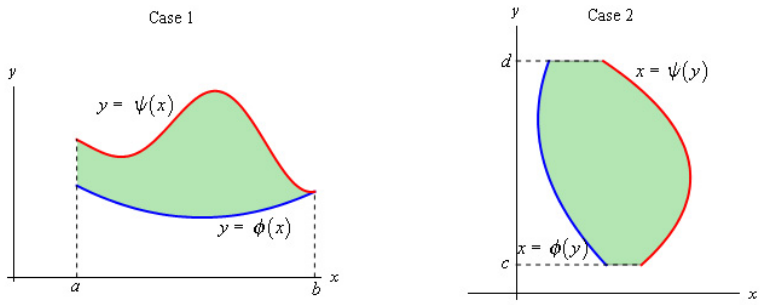


Figure IX.21: Projectable Regions.

IX.3.4.19. If E is the projectable region above, fix c and d with $c \le \phi(x) \le \psi(x) \le d$ for all $x \in [a,b]$. If $f : E \to \mathbb{R}$ is a function, then f can be considered a function on $R = [a,b] \times [c,d]$ in the usual way by setting it 0 on $R \setminus E$. If f is continuous on E, then for each $x \in [a,b]$ we have that $f_{[x]}$ is piecewise-continuous on $[c,d]$, supported on $[\phi(x), \psi(x)]$, hence integrable on $[c,d]$. Fubini's Theorem (IX.3.4.6.(i) form) applies to give

$$\int_E f\,dV = \int_R f\,dV = \int_a^b \int_c^d f(x,y)\,dy\,dx = \int_a^b \int_{\phi(x)}^{\psi(x)} f(x,y)\,dy\,dx\ .$$

Thus we obtain:

IX.3.4.20. Theorem. [Fubini's Theorem for Projectable Sets in $\mathbb{R}^2$] Let

$$E = \{(x,y) : a \le x \le b,\ \phi(x) \le y \le \psi(x)\}$$

be a projectable region (with respect to the order x, y) in $\mathbb{R}^2$, and $f : E \to \mathbb{R}$ a continuous function. Then

$$\int_E f\,dV = \int_a^b \int_{\phi(x)}^{\psi(x)} f(x,y)\,dy\,dx$$

(in particular, the iterated integral exists).

IX.3.4.21. Of course, the same result holds with x and y interchanged: if

$$E = \{(x,y) : c \le y \le d,\ \phi(y) \le x \le \psi(y)\}$$

and f is continuous on E, then

$$\int_E f\,dV = \int_c^d \int_{\phi(y)}^{\psi(y)} f(x,y)\,dx\,dy\ .$$

If E is projectable in either order, we thus get a result about interchanging order of integration:

IX.3.4.22. Corollary. [Fubini's Theorem for Iterated Integrals on Projectable Regions] Let

$$E = \{(x,y) : a \le x \le b,\ \phi_1(x) \le y \le \psi_1(x)\} = \{(x,y) : c \le y \le d,\ \phi_2(y) \le x \le \psi_2(y)\}$$

be a region in $\mathbb{R}^2$ which is projectable in either order. If $f : E \to \mathbb{R}$ is continuous, then

$$\int_a^b \int_{\phi_1(x)}^{\psi_1(x)} f(x,y)\,dy\,dx = \int_c^d \int_{\phi_2(y)}^{\psi_2(y)} f(x,y)\,dx\,dy$$

(in particular, the iterated integrals exist).

IX.3.4.23. However, note that in general $[c,d] \neq [a,b]$, $\phi_2 \neq \phi_1$, $\psi_2 \neq \psi_1$, so the limits of integration must be carefully modified. It is frequently easy to do an iterated integral in one order and difficult or even impossible in the other order; cf. IX.3.12.5..

IX.3.4.24. We can do the same thing in higher dimensions, although the notation becomes cumbersome. Fix an ordering $x_{j_1}, \ldots, x_{j_n}$ of the coordinates of $\mathbb{R}^n$. To streamline notation, we will write things out with $j_k = k$ for all k; the masochistic reader can easily adapt to the general situation. Inductively define $E_k \subseteq \mathbb{R}^k$ as follows. Let $E_1 = [a,b]$ be an interval in $\mathbb{R}$, and $\phi_1, \psi_1 : E_1 \to \mathbb{R}$ continuous functions with $\phi_1 \le \psi_1$ on E_1. Suppose $E_1, \ldots, E_k$ have been defined, and $\phi_k, \psi_k : E_k \to \mathbb{R}$ continuous functions with $\phi_k \le \psi_k$ on E_k. Set

$$E_{k+1} = \{(x_1, \ldots, x_{k+1}) \in \mathbb{R}^{k+1} : (x_1, \ldots, x_k) \in E_k, \phi_k(x_1, \ldots, x_k) \le x_{k+1} \le \psi_k(x_1, \ldots, x_k)\}\ .$$

By inductively applying IX.3.2.17. as in IX.3.4.18., we obtain that E_k is a Jordan region in $\mathbb{R}^k$ for each k.

IX.3.4.25. Definition. A set $E \subseteq \mathbb{R}^n$ is called a *projectable region* with respect to the order $x_1, \ldots, x_n$ if $E = E_n$ for some set of continuous functions as in IX.3.4.24..

Let E be a projectable region (for the order $x_1, \ldots, x_n$), and $f : E \to \mathbb{R}$ a continuous function. Use the notation of IX.3.4.24.. By successively applying the argument of IX.3.4.19., we obtain (note that all the integrals exist as Riemann integrals):

IX.3.4.26. Theorem. [Fubini's Theorem for Projectable Sets in $\mathbb{R}^n$] Let E be a projectable region (for the order $x_1, \ldots, x_n$), and $f : E \to \mathbb{R}$ a continuous function. Then

$$\int_E f\,dV = \int_a^b \int_{\phi_1(x_1)}^{\psi_1(x_1)} \cdots \int_{\phi_{n-1}(x_1,\ldots,x_{n-1})}^{\psi_{n-1}(x_1,\ldots,x_{n-1})} f(x_1, \ldots, x_n)\,dx_n \cdots dx_2\,dx_1\ .$$

IX.3.4.27. If E is also projectable in another ordering of the coordinates, the integral $\int_E f\,dV$ can be computed by an iterated integral in that order, although the ϕ_k and ψ_k will change in general when the ordering of the coordinates is changed. Thus we obtain as a corollary that the iterated integrals in the two orders are equal. The precise statement along the lines of IX.3.4.22. is left to the reader.

The case $n = 3$ is important enough for applications that it is worth stating explicitly:

IX.3.4.28. THEOREM. [FUBINI'S THEOREM FOR PROJECTABLE SETS IN $\mathbb{R}^3$] Let $[a,b]$ be an interval in $\mathbb{R}$, ϕ_1, ψ_1 continuous functions from $[a,b]$ to $\mathbb{R}$ with $\phi_1 \le \psi_1$ on $[a,b]$, $E_2 = \{(x,y) : a \le x \le b, \phi_1(x) \le y \le \psi_1(x)\}$, ϕ_2, ψ_2 continuous functions from E_2 to $\mathbb{R}$ with $\phi_2 \le \psi_2$, and

$$E = \{(x,y,z) : a \le x \le b, \phi_1(x) \le y \le \psi_1(x), \phi_2(x,y) \le z \le \psi_2(x,y)\} \ .$$

Let $f : E \to \mathbb{R}$ be a continuous function. Then

$$\int_E f \, dV = \int_a^b \int_{\phi_1(x)}^{\psi_1(x)} \int_{\phi_2(x,y)}^{\psi_2(x,y)} f(x,y,z) \, dz \, dy \, dx \ .$$

If E can be similarly written with respect to a different ordering of x, y, z, $\int_E f \, dV$ can be also calculated as an iterated integral in that order, and in particular the iterated integrals in the two orders are equal.

Most calculus texts have extensive examples of computing areas, volumes, and multiple integrals.

Counterexamples

We now give some examples showing that the hypotheses of Fubini's Theorem are necessary.

IX.3.4.29. EXAMPLES. Existence and equality of the iterated integrals does not imply integrability of f:
(i) Let $R = [0,1]^2 \subseteq \mathbb{R}^2$. Define $f : R \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} 2^n & \text{if} \quad 2^{-n} < x, y \le 2^{-n+1} \\ 0 & \text{otherwise} \end{cases} \ .$$

Then $f_{[x]}$ and $f^{[y]}$ are integrable for all x and y, and all have integral 1, so the iterated integrals exist and

$$\int_0^1 \int_0^1 f(x,y) \, dy \, dx = \int_0^1 \int_0^1 f(x,y) \, dx \, dy = 1 \ .$$

But f is not integrable over R since it is unbounded.
(ii) A bounded example is more complicated. Let $R = [0,1]^2 \subseteq \mathbb{R}^2$. Define $f : R \to \mathbb{R}$ by setting $f(x,y) = 1$ if x and y are rational numbers with the same denominator when written in lowest terms, and $f(x,y) = 0$ otherwise. Then for fixed x, $f(x,y) = 1$ for only finitely many y, hence $f_{[x]}$ is integrable and has integral 0. Similarly, any $f^{[y]}$ is integrable and has integral 0. Thus the iterated integrals both exist and are 0. However, if $\mathcal{P}$ is any partition of R, it is easily seen that $\overline{\mathcal{S}}(f,\mathcal{P}) = 1$ and $\underline{\mathcal{S}}(f,\mathcal{P}) = 0$, so f is not integrable over R. In fact, f is discontinuous everywhere on R.

IX.3.4.30. EXAMPLE. Existence of the iterated integrals does not imply they are equal if f is not integrable: let $R = [0,1]^2 \subseteq \mathbb{R}^2$. Define $f : R \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} 2^{2n} & \text{if} \quad 2^{-n} < x, y \le 2^{-n+1} \\ -2^{2n+1} & \text{if} \quad 2^{-n} < x \le 2^{-n+1}, \ 2^{-n-1} < y \le 2^{-n} \\ 0 & \text{otherwise} \end{cases} \ .$$

Then $f_{[x]}$ is integrable on $[0,1]$ and $\int_0^1 f_{[x]}(y)\,dy = 0$ for all x; thus

$$\int_0^1 \int_0^1 f(x,y)\,dy\,dx \text{ exists and equals } 0\ .$$

Also, $f^{[y]}$ is integrable on $[0,1]$ for all y; $\int_0^1 f^{[y]}(x)\,dx = 0$ if $y \le \frac{1}{2}$ but $\int_0^1 f^{[y]}(x)\,dx = 2$ if $y > \frac{1}{2}$; thus

$$\int_0^1 \int_0^1 f(x,y)\,dy\,dx \text{ exists and equals } 1\ .$$

So the iterated integrals are not equal. Note that f is unbounded and hence not integrable on R.
This example is an adaptation of the double sum example IV.2.10.6..

IX.3.4.31. EXAMPLE. Integrability of f does not imply integrability of $f_{[\boldsymbol{x}]}$ or $f^{[\boldsymbol{y}]}$ for all $\boldsymbol{x}$, $\boldsymbol{y}$: let $R = [0,1]^2 \subseteq \mathbb{R}^2$, B any finite subset of $[0,1]$, and set

$$f(x,y) = \begin{cases} 1 & \text{if} \quad x \in B,\ y \in \mathbb{Q} \\ 0 & \text{otherwise} \end{cases}\ .$$

Then f is bounded, and continuous except on $B \times [0,1]$, a set of content 0, so f is integrable on R. However, for $x \in B$, $f_{[x]}$ is not integrable on $[0,1]$. The example can be easily modified to also make $f^{[y]}$ nonintegrable for finitely many y. Replacing B with the Cantor set () gives examples where $f_{[x]}$ and/or $f^{[y]}$ are nonintegrable for infinitely many x, y.

IX.3.4.32. The full measure-theoretic form of Fubini's Theorem implies that if f is Riemann integrable on $R \times S$, hence continuous except on a set E of measure 0, then the vertical cross sections $E_{\boldsymbol{x}}$ have measure 0 for almost all $\boldsymbol{x}$, and hence $f_{[\boldsymbol{x}]}$ is Riemann integrable on S for almost all $\boldsymbol{x}$, and similarly $f^{[\boldsymbol{y}]}$ is Riemann integrable on R for almost all $\boldsymbol{y}$. Thus IX.3.4.31. is nearly as bad as the situation can get. But $f_{[\boldsymbol{x}]}$ and/or $f^{[\boldsymbol{y}]}$ can fail to be Riemann integrable for a dense set of $\boldsymbol{x}$ and/or $\boldsymbol{y}$ (Exercise IX.3.12.4.).

IX.3.4.33. The example IX.3.4.30. of an f where the iterated integrals exist but are unequal has an unbounded f. In fact, a bounded example is impossible. For the full Fubini Theorem () implies that if $f : R \times S \to \mathbb{R}$ is bounded and Lebesgue measurable, a very mild restriction satisfied by any function which can be explicitly described, the iterated integrals (which automatically exist as Lebesgue integrals) are equal. Whether or not f is Lebesgue measurable (but bounded), if $f_{[\boldsymbol{x}]}$ and $f^{[\boldsymbol{y}]}$ are Riemann integrable for all $\boldsymbol{x}$ and $\boldsymbol{y}$, the iterated integrals automatically exist as Riemann integrals and are equal ([**?**], [**?**]; cf. [Bog07, 3.10.57]). Tonelli's Theorem () also rules out an example with an unbounded nonnegative (measurable) f, hence an example with (measurable) f bounded either above or below.

IX.3.4.34. Some authors have objected to the use of the name "Fubini's Theorem" for results like IX.3.4.10. or even IX.3.4.6., on the grounds that they were "known" to people like EULER, long before FUBINI. But the multidimensional Riemann integral was not even defined until the 1870s (by JORDAN), so IX.3.4.10. as stated could not have been known to EULER. The part of this result about interchanging

order of integration in an iterated integral (the "Weak Fubini Theorem" IX.3.5.2.), or its equivalent IX.3.5.1., might have been a folklore result in Euler's time, but it is highly doubtful it had been proved (for one thing, at some point in the proof compactness or uniform continuity is needed, concepts not understood until the nineteenth century). Even justifying interchanging order of summation in a double infinite sum (IV.2.10.8.) requires careful analysis not nailed down until the nineteenth century. Some results which can be viewed as special cases of Fubini's Theorem, or more accurately of Cavalieri's Principle, such as computing the volume of a solid of revolution, had been proved (by the standards of the time), but such results only occasionally involved iterated integrals. I have yet to see clear evidence that anyone proved IX.3.4.10. or any other true version of Fubini's Theorem before Fubini. Thus I see no reason not to continue to use the term "Fubini's Theorem" generically for these results.

IX.3.5. Interchanging Differentiation and Integration

From Fubini's Theorem we can obtain an important result on interchanging the order of differentiation and integration: for a function of two or more variables, differentiation with respect to one variable commutes with integration with respect to another variable. We will obtain a more general version in XVIII.2.1.1., but from IX.3.4.10. we can easily get the continuous version, which is adequate for many (perhaps most) applications.

IX.3.5.1. Theorem. Let I be an open interval in $\mathbb{R}$, $[c,d]$ a closed bounded interval in $\mathbb{R}$, and $R = I \times [c,d]$ the corresponding rectangle in $\mathbb{R}^2$. Let $f : R \to \mathbb{R}$ be a function such that f and $\frac{\partial f}{\partial x}$ are continuous on R. For $x \in I$ define

$$g(x) = \int_c^d f(x,y)\,dy \ .$$

Then g is $\mathcal{C}^1$ on I and, for $x \in I$,

$$g'(x) = \int_c^d \frac{\partial f}{\partial x}(x,y)\,dy \ .$$

Slightly informally,

$$\frac{d}{dx}\left(\int_c^d f(x,y)\,dy\right) = \int_c^d \frac{\partial f}{\partial x}(x,y)\,dy$$

(often called "differentiation under the integral sign.")

In fact, for the proof we only need the part of Fubini's Theorem asserting that iterated integrals are equal:

IX.3.5.2. Theorem. [Weak Fubini Theorem] Let $R = [a,b] \times [c,d]$ be a rectangle in $\mathbb{R}^2$, and $f : R \to \mathbb{R}$ a continuous function. Then

$$\int_a^b \int_c^d f(x,y)\,dy\,dx = \int_c^d \int_a^b f(x,y)\,dx\,dy \ .$$

In particular, the two iterated integrals exist as Riemann integrals.

IX.3.5.3. Actually, Theorem IX.3.5.1. is equivalent to the Weak Fubini Theorem in the sense that each is an easy corollary of the other.

For the proofs, we need the following uniform continuity result, which is a special case of ():

IX.3.5.4. PROPOSITION. Let I be an interval and $[c,d]$ a closed bounded interval in $\mathbb{R}$, and let $f : I \times [c,d] \to \mathbb{R}$ be a continuous function. Then for every $x_0 \in I$ and $\epsilon > 0$, there is a $\delta > 0$ such that $|f(x,y) - f(x_0,y)| < \epsilon$ for any $x \in I$, $|x - x_0| < \delta$, and every $y \in [c,d]$.

IX.3.5.5. COROLLARY. Let I be an interval and $[c,d]$ a closed bounded interval in $\mathbb{R}$, and let $f : I \times [c,d] \to \mathbb{R}$ be a continuous function. For $x \in I$, define

$$\phi(x) = \int_c^d f(x,y)\,dy \ .$$

Then ϕ is a continuous function from I to $\mathbb{R}$.

PROOF: Let $x_0 \in I$. If $\epsilon > 0$, choose $\delta > 0$ such that $|f(x,y) - f(x_0,y)| < \frac{\epsilon}{d-c}$ for any $x \in I$, $|x - x_0| < \delta$, and every $y \in [c,d]$. Then, for $x \in I$, $|x - x_0| < \delta$,

$$|\phi(x) - \phi(x_0)| = \left| \int_c^d [f(x,y) - f(x_0,y)]\,dy \right| \leq \int_c^d |f(x,y) - f(x_0,y)|\,dy < \frac{\epsilon}{d-c}(d-c) = \epsilon \ .$$



IX.3.5.6. We now give the proof of IX.3.5.1. from IX.3.5.2.:

PROOF: For $x \in I$, define

$$\phi(x) = \int_c^d \frac{\partial f}{\partial x}(x,y)\,dy \ .$$

Then ϕ is a continuous function from I to $\mathbb{R}$ (IX.3.5.5.). Fix $x_0 \in I$, and fix $a \in I$, $a < x_0$. For $x \in I$, $x > a$, define

$$h(x) = \int_a^x \phi(t)\,dt = \int_a^x \int_c^d \frac{\partial f}{\partial x}(t,y)\,dy\,dt \ .$$

Then, by the Fundamental Theorem of Calculus (FTC1), h is differentiable and $h'(x) = \phi(x)$ for $x \in I$, $x > a$, and in particular $h'(x_0) = \phi(x_0)$. On the other hand, by IX.3.5.2. and the FTC2, for $x \in I$, $x > a$ we have

$$h(x) = \int_c^d \int_a^x \frac{\partial f}{\partial x}(t,y)\,dt\,dy = \int_c^d [f(x,y) - f(a,y)]\,dy$$
$$= \int_c^d f(x,y)\,dy - \int_c^d f(a,y)\,dy = g(x) - g(a)$$

so g is differentiable and $g'(x) = h'(x) = \phi(x)$ for $x \in I$, $x > a$, and in particular g is differentiable at x_0 and $g'(x_0) = \phi(x_0)$.



IX.3.5.7. Conversely, we can easily deduce IX.3.5.2. from IX.3.5.1.:

PROOF: Let $B > b$ and extend f to a continuous function from $[a, B] \times [c, d]$ to $\mathbb{R}$ by setting $f(x, y) = f(b, y)$ for $x \in [b, B]$, $y \in [c, d]$. For $x \in [a, B]$, define

$$g(x) = \int_c^d \int_a^x f(t, y)\, dt\, dy \ .$$

Then $g(a) = 0$, and by IX.3.5.1. and the FTC1 g is differentiable on (a, B) and

$$g'(x) = \int_c^d \frac{\partial}{\partial x} \left[\int_a^x f(t, y)\, dt \right] dy = \int_c^d f(x, y)\, dy$$

for $x \in (a, B)$. Since g' is continuous, by the FTC2 we have, for $x \in (a, B)$,

$$g(x) = g(a) + \int_a^x g'(t)\, dt = \int_a^x \int_c^d f(t, y)\, dy\, dt \ .$$

Then IX.3.5.2. follows by setting $x = b$. 👍

IX.3.5.8. We can give a direct proof of Theorem IX.3.5.1., which by IX.3.5.7. gives an alternate proof of IX.3.5.2.:

PROOF: Fix $x_0 \in I$. Then, for h small enough that $x_0 + h \in I$, we have

$$\frac{g(x_0 + h) - g(x_0)}{h} - \int_c^d \frac{\partial f}{\partial x}(x_0, y)\, dy = \int_c^d \left[\frac{f(x_0 + h, y) - f(x_0, y)}{h} - \frac{\partial f}{\partial x}(x_0, y) \right] dy \ .$$

By the Mean Value Theorem, if the interval between x_0 and $x_0 + h$ is contained in I, then for each $y \in [c, d]$ there is a u_y between x_0 and $x_0 + h$ (depending on y and h) with

$$\frac{f(x_0 + h, y) - f(x_0, y)}{h} = \frac{\partial f}{\partial x}(u_y, y) \ .$$

Let $\epsilon > 0$. Since $\frac{\partial f}{\partial x}$ is continuous, by IX.3.5.4. there is a $\delta > 0$ such that

$$\left| \frac{\partial f}{\partial x}(x, y) - \frac{\partial f}{\partial x}(x_0, y) \right| < \frac{\epsilon}{d - c}$$

whenever $x \in I$, $|x - x_0| < \delta$, and $y \in [c, d]$. Thus, if h is sufficiently small,

$$\left| \frac{g(x_0 + h) - g(x_0)}{h} - \int_c^d \frac{\partial f}{\partial x}(x_0, y)\, dy \right| = \left| \int_c^d \left[\frac{\partial f}{\partial x}(u_y, y) - \frac{\partial f}{\partial x}(x_0, y) \right] dy \right|$$

$$\leq \int_c^d \left| \frac{\partial f}{\partial x}(u_y, y) - \frac{\partial f}{\partial x}(x_0, y) \right| dy < \frac{\epsilon}{d - c}(d - c) = \epsilon \ .$$

👍

Note, however, that this does not give a proof of the full Fubini Theorem, even IX.3.4.10., without additional argument. ✋

Using the Chain Rule, we can extend IX.3.5.1. to integration over more general regions:

IX.3.5.9. THEOREM. Let U be an open set in $\mathbb{R}^2$ containing the closed set

$$S = \{(x,y) \in \mathbb{R}^2 : a \le x \le b, u(x) \le y \le v(x)\}$$

where u and v are differentiable functions from $I = [a,b]$ to $\mathbb{R}$ with $u \le v$. Let $f : U \to \mathbb{R}$ be a function such that f and $\frac{\partial f}{\partial x}$ are continuous on U. For $x \in I$ define

$$g(x) = \int_{u(x)}^{v(x)} f(x,y)\,dy \ .$$

Then g is differentiable on I and, for $x \in I$,

$$g'(x) = \int_{u(x)}^{v(x)} \frac{\partial f}{\partial x}(x,y)\,dy + f(x,v(x))v'(x) - f(x,u(x))u'(x) \ .$$

If u and v are $\mathcal{C}^1$, so is g.

PROOF: Set $\phi(x,u,v) = \int_u^v f(x,y)\,dy$. Then ϕ is defined on a suitable neighborhood W of

$$\{(x,u,v) : a \le x \le b, u(x) \le u \le v \le v(x)\}$$

in $\mathbb{R}^3$, and $g(x) = \phi(x,u(x),v(x))$. By IX.3.5.1.,

$$\frac{\partial \phi}{\partial x}(x,u,v) = \int_u^v \frac{\partial f}{\partial x}(x,y)\,dy \ .$$

In particular, $\frac{\partial \phi}{\partial x}$ exists and is continuous on W. By the FTC, $\frac{\partial \phi}{\partial v}(x,u,v) = f(x,v)$ and $\frac{\partial \phi}{\partial u}(x,u,v) = -f(x,u)$ and are thus continuous on W. Therefore ϕ is differentiable ($\mathcal{C}^1$) on W. By the Chain Rule, g is diffrentiable on I and, for $x \in I$,

$$\begin{aligned} g'(x) &= \frac{\partial \phi}{\partial x}(x,u(x),v(x)) \cdot 1 + \frac{\partial \phi}{\partial u}(x,u(x),v(x))u'(x) + \frac{\partial \phi}{\partial v}(x,u(x),v(x))v'(x) \\ &= \int_{u(x)}^{v(x)} \frac{\partial f}{\partial x}(x,y)\,dy - f(x,u(x))u'(x) + f(x,v(x))v'(x) \ . \end{aligned}$$



Multidimensional versions can be obtained similarly, such as (cf. Exercise IX.3.12.7.):

IX.3.5.10. THEOREM. Let R be an open rectangle in $\mathbb{R}^p$, S a closed bounded rectangle in $\mathbb{R}^q$, and $f(x_1,\ldots,x_p,y_1,\ldots,y_q) : R \times S \to \mathbb{R}$ a function such that f and $\frac{\partial f}{\partial x_i}$ are continuous on $R \times S$ for some i, $1 \le i \le p$. For $\boldsymbol{x} = (x_1,\ldots,x_p) \in R$ define

$$g(\boldsymbol{x}) = \int_S f(\boldsymbol{x},\boldsymbol{y})\,dV(\boldsymbol{y}) \ .$$

Then $\frac{\partial g}{\partial x_i}$ exists and is continuous on R and, for $\boldsymbol{x} \in R$,

$$\frac{\partial g}{\partial x_i}(\boldsymbol{x}) = \int_S \frac{\partial f}{\partial x_i}(\boldsymbol{x},\boldsymbol{y})\,dV(\boldsymbol{y}) \ .$$

If $\frac{\partial f}{\partial x_i}$ exists and is continuous on $R \times S$ for all i, $1 \le i \le p$ (and f is continuous), then g is $\mathcal{C}^1$ on R.

IX.3.6. Change of Variable

IX.3.6.1. EXAMPLE. Here is a standard elementary example ("polar coordinates") of a change of variable for which the hypotheses of () are not satisfied. Let $U = V = \mathbb{R}^2$. Denote the coordinates in U by (r, θ) and those of V by (x, y). Let $\boldsymbol{f} : U \to V$ be

$$\boldsymbol{f}(r, \theta) = (r \cos\theta, r \sin\theta) \ .$$

Fix $b > 0$ and let $R = [0, b] \times [0, 2\pi] \subseteq U$ and S the closed disk of radius b around $\mathbf{0}$ in V. Then $\boldsymbol{f}$ maps R onto S, and R° homeomorphically into $S^{\mathbf{0}}$. But $\boldsymbol{f}$ is not a proper map from U to V (the preimage of $\{\mathbf{0}\}$ is the entire θ-axis), and is not a proper map from R° to S° (for one thing, its range is not closed in S°, cf. ()); $\boldsymbol{f}$ is not one-to-one on ∂R, and $\boldsymbol{f}(\partial R)$ is not contained in ∂S.

The function $\boldsymbol{f}$ is smooth, and

$$det(D\boldsymbol{f}(r, \theta)) = det\left(\begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}\right) = r \ .$$

Thus $det(D\boldsymbol{f}) \geq 0$ on R and is strictly positive on R°. Although () does not apply as stated, it is nonetheless true that if $g : S \to \mathbb{R}$ is a continuous function, then

$$\int_S g \, dV = \int_R (g \circ \boldsymbol{f}) det(D\boldsymbol{f}) \, dV = \int_0^{2\pi} \int_0^b g(r\cos\theta, r\sin\theta) \, r \, dr \, d\theta \ .$$

We need a version of () which applies in this and similar examples to give the conclusion.

IX.3.7. Improper Integrals

IX.3.8. Step Functions

Riemann integration can be alternately described via step functions. There are some subtleties which must be addressed in this approach concerning values on boundaries of rectangles; these subtleties are "wasted" in the sense that they have no ultimate effect, but they are necessary for a logically complete development, and they can be handled cleanly if done right.

IX.3.8.1. DEFINITION. A *step function* on $\mathbb{R}^p$ is a finite linear combination of indicator functions of rectangles (degenerate rectangles are allowed).

IX.3.8.2. Any (finite) linear combination of step functions is a step function.

IX.3.8.3. If $\phi = \sum_{k=1}^{n} \alpha_k \phi_{R_k}$ is a step function, there is a rectangle R such that $R_k \subseteq R$ for all k. A partition $\mathcal{P}$ of R is *ϕ-fine* if each R_k is a union of rectangles in $\mathcal{P}$. By (), every step function ϕ has a ϕ-fine partition of any sufficiently large rectangle. Any refinement of a ϕ-fine partition is ϕ-fine.

IX.3.8.4. In the step function approach to Riemann integration, it is very convenient to allow partitions which include degenerate rectangles. In fact, it is most convenient to work with *full partitions* $\mathcal{P}$, in which any nonempty intersection of rectangles in $\mathcal{P}$ is also in $\mathcal{P}$ (all such intersections are, of course, degenerate since the rectangles in $\mathcal{P}$ are nonoverlapping). A grid in which every one-dimensional factor has all transition points included as degenerate intervals is a full partition, so any partition has a full refinement.

If ϕ is a step function and $\mathcal{P}$ is a full ϕ-fine partition, then ϕ can be written as a linear combination of indicator functions of rectangles in $\mathcal{P}$. This is an immediate consequence of the following fact, which has a straightforward but notationally complex proof by induction on n left to the reader:

IX.3.8.5. PROPOSITION. Let X be a set, and $A_1, \ldots, A_n$ subsets of X. Then

$$\chi_{A_1 \cup \cdots \cup A_n} = \sum_{k=1}^{n} \chi_{A_k} - \sum_{j<k} \chi_{A_j \cap A_k} + \sum_{i<j<k} \chi_{A_i \cap A_j \cap A_k} - \cdots + (-1)^{n+1} \chi_{A_1 \cap \cdots \cap A_n}$$

where each sum is over all strictly increasing sets of distinct indices.

IX.3.8.6. Each rectangle R has an "interior" R^o consisting of all points of R not contained in a boundary rectangle of smaller dimension. If R is a singleton, set $R^o = R$. Note that R^o is not the topological interior of R in $\mathbb{R}^p$ if R is degenerate, but is rather the topological interior relative to the affine subspace of $\mathbb{R}^p$ generated by R. R^o is a relatively open subset of R. If $\mathcal{P}$ is a full partition of a rectangle R, for each $R_k \in \mathcal{P}$ we have that R_k^o is the set of points of R_k not contained in any $R_j \in \mathcal{P}$ properly contained in R_k. Then R is the disjoint union of the R_k^o. Any linear combination ϕ of the indicator functions of the R_k^o is a step function (use IX.3.8.5.) for which $\mathcal{P}$ is ϕ-fine, and any step function ψ for which $\mathcal{P}$ is ψ-fine can be written uniquely as a linear combination of these indicator functions.

IX.3.8.7. It follows immediately that any finite maximum or minimum of step functions is a step function. [Consider a full partition which is fine for all the step functions.]

Integrals of Step Functions

IX.3.8.8. If $\phi = \sum_{k=1}^{m} \alpha_k \chi_{R_k}$ is a step function, then ϕ is Riemann integrable over any rectangle R containing all the R_k and

$$\int_R \phi \, dV = \sum_{k=1}^{m} \alpha_k V(R_k) \ .$$

It is fairly easily verified using IX.3.8.5. that the sum on the right depends only on ϕ and not on how it is written as a linear combination of indicator functions; thus it can be taken as the *definition* of the integral of ϕ over R. The integral of step functions so defined is easily verified to be linear, additive, and monotone. It also has a continuity property (IX.3.8.20.).

IX.3.8.9. The Riemann integral can be alternately defined and constructed using this integral for step functions, without using the upper and lower sum definitions previously used ().

Maximum and Minimum Step Functions

IX.3.8.10. Let R be a rectangle and f a bounded function on R. If $\mathcal{P}$ is a full partition of R, then there is a unique largest step function ϕ for which $\mathcal{P}$ is ϕ-fine and $\phi \le f$. If $\mathcal{P} = \{R_1, \ldots, R_m\}$, then ϕ is the function whose value on R_k^o is the infimum of f on R_k^o. Similarly, there is a smallest step function ψ for which $\mathcal{P}$ is ψ-fine and $f \le \psi$; on R_k^o ψ equals the supremum of f on R_k^o. We have

$$\underline{\mathcal{S}}(f, \mathcal{P}) \le \int_R \phi \, dV \le \underline{\int_R} f \, dV \le \overline{\int_R} f \, dV \le \int_R \psi \, dV \le \overline{\mathcal{S}}(f, \mathcal{P}) \ .$$

Write $\phi(f, \mathcal{P})$ and $\psi(f, \mathcal{P})$ for ϕ and ψ. If $\mathcal{Q}$ is a refinement of $\mathcal{P}$, then $\phi(f, \mathcal{P}) \le \phi(f, \mathcal{Q})$ and $\psi(f, \mathcal{Q}) \le \psi(f, \mathcal{P})$.

IX.3.8.11. PROPOSITION. Let R be a rectangle in $\mathbb{R}^p$, and $f : \mathbb{R} \to \mathbb{R}$ a bounded function. The following are equivalent:

(i) f is Riemann integrable on R.

(ii) There are sequences (ϕ_n) and (ψ_n) of step functions supported on R with

$$\phi_n \le \phi_{n+1} \le f \le \psi_{n+1} \le \psi_n$$

on R for all n and

$$\lim_{n\to\infty} \int_R \phi_n \, dV = \sup_n \int_R \phi_n \, dV = \inf_n \int_R \psi_n \, dV = \lim_{n\to\infty} \int_R \psi_n \, dV \ .$$

PROOF: (ii) $\Rightarrow$ (i) since

$$\int_R \phi_n \, dV \le \underline{\int_R} f \, dV \le \overline{\int_R} f \, dV \le \int_R \psi_n \, dV$$

for each n by monotonicity.
For (i) $\Rightarrow$ (ii), choose a sequence $\mathcal{P}_n$ of full partitions of R such that $\mathcal{P}_{n+1}$ refines $\mathcal{P}_n$ for all n and $\lim_{n\to\infty}[\overline{\mathcal{S}}(f, \mathcal{P}_n) - \underline{\mathcal{S}}(f, \mathcal{P}_n)] = 0$. Set $\phi_n = \phi(f, \mathcal{P}_n)$, $\psi_n = \psi(f, \mathcal{P}_n)$ and apply IX.3.8.10.. 👍

Two more equivalent conditions will be added in IX.3.10.6..

Sets of Measure Zero

There is a more general and flexible notion of "negligibly small" sets than content zero:

IX.3.8.12. DEFINITION. A subset A of $\mathbb{R}^p$ has *measure zero* if for any $\epsilon > 0$ there is a sequence (R_k) of rectangles in $\mathbb{R}^p$ (degenerate rectangles are allowed) such that $A \subseteq \cup_{k=1}^{\infty} R_k$ and

$$\sum_{k=1}^{\infty} V(R_k) < \epsilon \ .$$

IX.3.8.13. Since degenerate rectangles are allowed in the definition, a set of zero content is a set of measure zero. A set of measure zero is not necessarily a Jordan region; in fact, it is not even necessarily bounded. Any subset of a set of measure zero is a set of measure zero. A finite or even countable union of sets of measure zero is a set of measure zero (). In particular, since a singleton set has measure zero, any countable subset of $\mathbb{R}^p$ has measure zero. Thus the set of all points with rational coordinates has measure zero.

We do have ():

IX.3.8.14. PROPOSITION. Let $A \subseteq \mathbb{R}^p$. Then

(i) If A is bounded, then A has zero content if and only if $\bar{A}$ has measure zero.

(ii) If A is a Jordan region, then A has measure zero if and only if it has zero content.

(iii) If A is compact (closed and bounded) and has measure zero, then A has zero content (and is thus a Jordan region).

IX.3.8.15. The notion of a set of measure zero is part of the theory of Lebesgue measure, which is discussed in detail in (). A notion of size or "measure" is defined for all reasonable ("measurable") subsets of $\mathbb{R}^p$, which extends Jordan content, i.e. the measure of a Jordan region is exactly its Jordan content. But many more sets are measurable; in fact, every subset of $\mathbb{R}^p$ which can be explicitly described turns out to be measurable. There is an extended notion of integral called the Lebesgue integral: "measurable" functions can be integrated over measurable sets (with some technical size restrictions related to boundedness).

We also make the following definition:

IX.3.8.16. DEFINITION. Let $A \subseteq \mathbb{R}^p$, and $f, g : A \to \mathbb{R}$. Then $f = g$ *almost everywhere on* A if

$$\{\boldsymbol{x} \in A : f(\boldsymbol{x}) \neq g(\boldsymbol{x})\}$$

is a set of measure zero. "Almost everywhere" is abbreviated *a.e.*

Characterization of Sets of Measure Zero

Sets of measure 0 can be characterized using integrals of step functions:

IX.3.8.17. THEOREM. Let R be a rectangle in $\mathbb{R}^p$, and $A \subseteq R$. Then A has measure 0 if and only if, for every $\epsilon > 0$, there is a nondecreasing sequence (ϕ_n) of step functions such that $\sup_n \phi_n(\boldsymbol{x}) \geq 1$ for all $\boldsymbol{x} \in A$ and $\int_R \phi_n \, dV < \epsilon$ for all n.

The proof consists of the following two lemmas.

IX.3.8.18. LEMMA. Let R be a rectangle in $\mathbb{R}^p$, and $A \subseteq R$. If, for every $\epsilon > 0$, there is a nonnegative step function ϕ such that $\phi(\boldsymbol{x}) \geq 1$ for all $\boldsymbol{x} \in A$ and $\int_R \phi \, dV < \epsilon$, then A has measure 0.

PROOF: Let $\epsilon > 0$ and ϕ the corresponding step function. Let $\phi = \sum \alpha_k \chi_{R_k^o}$ be the canonical representation of ϕ (IX.3.8.6.), and let $k_1, \ldots, k_n$ be the k for which $\alpha_k \geq 1$. Then $A \subseteq \cup_{j=1}^n R_{k_j}$ and

$$\sum_{j=1}^n V(R_{k_j}) \leq \sum_{j=1}^n \alpha_{k_j} V(R_{k_j}) \leq \int_R \phi \, dV < \epsilon \; .$$

Since $\epsilon > 0$ is arbitrary, A has measure 0. 👍

IX.3.8.19. LEMMA. Let R be a rectangle in $\mathbb{R}^p$, and $A \subseteq R$. If A has measure zero, then for any $\epsilon > 0$ there is a nondecreasing sequence (ϕ_n) of step functions such that $\sup_n \phi_n(\boldsymbol{x}) = +\infty$ for all $\boldsymbol{x} \in A$ and $\int_R \phi_n \, dV < \epsilon$ for all n.

PROOF: For each n there is a finite collection $\{R_{n,1}, \ldots, R_{n,m_n}\}$ of rectangles with $A \subseteq E_n := \cup_{k=1}^{m_n} R_{n,k}$ and $\sum_{k=1}^{m_n} V(R_{n,k}) < 2^{-n}\epsilon$. Then E_n is a Jordan region, $V(E_n) \leq 2^{-n}\epsilon$, and χ_{E_n} is a step function. Set

$$\phi_n = \sum_{k=1}^n \chi_{E_n} \; .$$

Then (ϕ_n) is a nondecreasing sequence of step functions, $\phi_n(\boldsymbol{x}) = n$ for $\boldsymbol{x} \in A$, and

$$\int_R \phi_n \, dV = \sum_{k=1}^n V(E_k) < \sum_{k=1}^n 2^{-k}\epsilon < \epsilon$$

for all n. 👍

IX.3.8.20. PROPOSITION. Let (ψ_n) be a nonincreasing sequence of nonnegative step functions supported on a fixed rectangle R. If

$$\lim_{n\to\infty} \int_R \psi_n \, dV = \inf_n \int_R \psi_n \, dV = 0$$

Then $\psi_n \to 0$ pointwise a.e.

PROOF: It suffices to show that $A_m = \{\boldsymbol{x} \in \mathbb{R} : \inf_n \psi_n(\boldsymbol{x}) \geq \frac{1}{m}\}$ has measure 0 for all $m \in \mathbb{N}$, since $\psi_n \to 0$ pointwise on the complement of $\cup_m A_m$. Fix m. For any $\epsilon > 0$, there is an n such that $\int_R \psi_n \, dV < \frac{\epsilon}{m}$. Then $\phi = m\psi_n$ is a step function with $\phi(\boldsymbol{x}) \geq 1$ for all $\boldsymbol{x} \in A_m$ and $\int_R \phi \, dV < \epsilon$. Thus A_m has measure 0 by IX.3.8.18.. 👍

We will prove an important converse in IX.3.9.6..

IX.3.9. The Monotone Convergence Theorem

The Monotone Convergence Theorem is an important continuity property of integrals:

IX.3.9.1. THEOREM. [MONOTONE CONVERGENCE THEOREM FOR RIEMANN INTEGRATION] Let R be a rectangle in $\mathbb{R}^p$, and (f_n) a nondecreasing sequence of nonnegative Riemann integrable functions on R converging pointwise a.e. on R to a Riemann integrable function f. Then

$$\int_R f\,dV = \lim_{n\to\infty} \int_R f_n\,dV = \sup_n \int_R f_n\,dV\ .$$

Note that it is the sequence which is nondecreasing, i.e. $f_n(\boldsymbol{x}) \le f_{n+1}(\boldsymbol{x})$ for all n and all $\boldsymbol{x} \in R$.

IX.3.9.2. There are two difficulties with this statement. First, the hypothesis that the limit function f is Riemann integrable on R is not automatic, even if the f_n are uniformly bounded. For example, let (r_n) be an enumeration of the rational numbers in $[0,1]$, and define $f_n(x)$ to be 1 if $x = r_k$ for some k, $1 \le k \le n$, and $f_n(x) = 0$ otherwise. Then f_n is 0 except at finitely many points, hence is Riemann integrable on $[0,1]$, but $f_n \to \chi_{\mathbb{Q}}$ pointwise on $[0,1]$, and $\chi_{\mathbb{Q}}$ is not Riemann integrable on $[0,1]$. Thus the Riemann integrability of f must be included as a hypothesis. To avoid this difficulty, replacing f_n by $f - f_n$ and using linearity of the Riemann integral the theorem may be more cleanly stated:

IX.3.9.3. THEOREM. Let R be a rectangle in $\mathbb{R}^p$, and (f_n) a nonincreasing sequence of nonnegative Riemann integrable functions on R converging pointwise a.e. on R to 0. Then

$$\lim_{n\to\infty} \int_R f_n\,dV = \inf_n \int_R f_n\,dV = 0\ .$$

IX.3.9.4. There is a second difficulty: proving this theorem. Although a direct proof can be given, it is quite complicated to do so. The simplest approach is to work via the Lebesgue integral and prove a more general theorem applicable to arbitrary Lebesgue measurable functions (XVIII.1.4.1.). In fact, overcoming both the difficulties with the Riemann integration statement of the theorem was one of the principal motivations for developing Lebesgue integration theory.

IX.3.9.5. We will content ourselves here with proving only the following special case, which will be crucial in developing the Daniell integral ().

IX.3.9.6. THEOREM. [MONOTONE CONVERGENCE THEOREM FOR STEP FUNCTIONS] Let R be a rectangle in $\mathbb{R}^p$, and (ψ_n) a nonincreasing sequence of nonnegative step functions on R converging pointwise a.e. on R to 0. Then

$$\lim_{n\to\infty} \int_R \psi_n\,dV = \inf_n \int_R \psi_n\,dV = 0\ .$$

PROOF: Let A be the set of $\boldsymbol{x} \in R$ for which $(\psi_n(\boldsymbol{x}))$ fails to converge to 0, and B the set of $\boldsymbol{x}$ at which at least one ψ_n is discontinuous, and $C = A \cup B$. Then C has measure 0. Let M be a bound for ϕ_1 on R.

Let $\epsilon > 0$. There is a sequence (R_k^o) of open rectangles covering C for which

$$\sum_{k=1}^{\infty} V(R_k^o) < \frac{\epsilon}{2M} \ .$$

If $E = R \setminus \cup_k R_k^o$, then for each $\boldsymbol{x} \in E$ there is an $n_{\boldsymbol{x}}$ and an open rectangle $R_{\boldsymbol{x}}^o$ containing $\boldsymbol{x}$ such that $\psi_{n_{\boldsymbol{x}}}$ takes a constant value less than $\frac{\epsilon}{2V(R)}$ on $R_{\boldsymbol{x}}^o$. The R_k^o and $R_{\boldsymbol{x}}^o$ cover R, so by compactness finitely many R_k^o and $R_{\boldsymbol{x}}^o$ cover E, say $R_{k_1}^o, \ldots, R_{k_r}^o, R_{\boldsymbol{x}_1}^o, \ldots, R_{\boldsymbol{x}_m}^o$. Let $R_{\boldsymbol{x}_k}$ be the closure of $R_{\boldsymbol{x}_k}^o$. By subdividing and eliminating overlaps we may assume the $R_{\boldsymbol{x}_k}$ are nonoverlapping, so $\sum_{k=1}^m V(R_{\boldsymbol{x}_k}) \leq V(R)$.

Set $N = \max\{n_{\boldsymbol{x}_1}, \ldots, n_{\boldsymbol{x}_m}\}$. Then if $n \geq N$, we have

$$\int_R \psi_n \, dV \leq \sum_{j=1}^{r} \int_{R_{k_j}^o} \psi_n \, dV + \sum_{k=1}^{m} \int_{R_{\boldsymbol{x}_k}} \psi_n \, dV < \frac{\epsilon}{2M} \sum_{k=1}^{\infty} MV(R_k^o) + \sum_{k=1}^{m} \frac{\epsilon}{2V(R)} V(R_{\boldsymbol{x}_k}) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$



IX.3.9.7. What this argument really shows is a special case of Egorov's Theorem (XVII.1.8.1.): under the hypotheses, ψ_n for n sufficiently large is actually uniformly close to 0 on the complement of a set of small measure. This is also reminiscent of Dini's Theorem ().

IX.3.10. Characterization of Riemann Integrable Functions

Using the Monotone Convergence Theorem (IX.3.9.6.), we can add a third equivalent condition to the characterization of Riemann integrability:

IX.3.10.1. THEOREM. Let R be a rectangle in $\mathbb{R}^p$, and $f : \mathbb{R} \to \mathbb{R}$ a bounded function. The following are equivalent:

(i) f is Riemann integrable on R.

(ii) There are sequences (ϕ_n) and (ψ_n) of step functions supported on R with

$$\phi_n \le \phi_{n+1} \le f \le \psi_{n+1} \le \psi_n$$

on R for all n and

$$\lim_{n\to\infty} \int_R \phi_n \, dV = \sup_n \int_R \phi_n \, dV = \inf_n \int_R \psi_n \, dV = \lim_{n\to\infty} \int_R \psi_n \, dV \ .$$

(iii) There are sequences (ϕ_n) and (ψ_n) of step functions supported on R with

$$\phi_n \le \phi_{n+1} \le f \le \psi_{n+1} \le \psi_n$$

on R for all n and $\phi_n \to f$ a.e. and $\psi_n \to f$ a.e. on R.

PROOF: (i) $\iff$ (ii) by IX.3.8.11.. For (ii) $\Rightarrow$ (iii), let ϕ_n, ψ_n be as in (ii). Then $(\psi_n - \phi_n)$ is a nonincreasing sequence of nonnegative step functions with

$$\lim_{n\to\infty} \int_R (\psi_n - \phi_n) \, dV = 0$$

and hence $\psi_n - \phi_n \to 0$ pointwise a.e., i.e. $\inf_n \psi_n(\boldsymbol{x}) = \sup_n \phi_n(\boldsymbol{x})$ for almost all $\boldsymbol{x} \in R$. Since $\sup_n \phi_n(\boldsymbol{x}) \le f(\boldsymbol{x}) \le \inf_n \psi_n(\boldsymbol{x})$ for all $\boldsymbol{x} \in R$, it follows that $\phi_n \to f$ a.e. and $\psi_n \to f$ a.e. (IX.3.8.20.).
Conversely, suppose ϕ_n and ψ_n are as in (iii). Then $(\psi_n - \phi_n)$ is a nonincreasing sequence of nonnegative step functions on R and $\psi_n - \phi_n \to 0$ pointwise a.e. Then by the Monotone Convergence Theorem we have

$$\lim_{n\to\infty} \int_R (\psi_n - \phi_n) \, dV = 0$$

so (ii) is satisfied. 👍

Upper and Lower Envelopes

IX.3.10.2. There is an alternate way of describing condition (iii) of IX.3.10.1. which leads to LEBESGUE'S characterization of Riemann integrability. Let f be a bounded function on a rectangle R. For each $\boldsymbol{a} \in R$, define

$$\underline{f}(\boldsymbol{a}) = \min\left(f(\boldsymbol{a}), \liminf_{\boldsymbol{x}\to\boldsymbol{a}} f(\boldsymbol{x})\right)$$

$$\overline{f}(\boldsymbol{a}) = \max\left(f(\boldsymbol{a}), \limsup_{\boldsymbol{x}\to\boldsymbol{a}} f(\boldsymbol{x})\right)$$

(these numbers are finite for all $\boldsymbol{a}$ since f is bounded). Then $\underline{f}$ and $\overline{f}$ are bounded functions on R, and $\underline{f} \le f \le \overline{f}$ on R; $\underline{f}(\boldsymbol{a}) = f(\boldsymbol{a}) = \overline{f}(\boldsymbol{a})$ if and only if f is continuous at $\boldsymbol{a}$. The functions $\overline{f}$ and $\underline{f}$ are called the *upper and lower envelopes* of f.

IX.3.10.3. Let $\mathcal{P}$ be a partition of R, and let $\phi(f,\mathcal{P})$ and $\psi(f,\mathcal{P})$ be the step functions defined in IX.3.8.10.. Then $\phi_n(f,\mathcal{P}) \le \underline{f}$ on the interior of any nondegenerate rectangle in $\mathcal{P}$, hence $\phi_n(f,\mathcal{P}) \le \underline{f}$ a.e. on R. Thus, if ϕ is any step function with $\phi \le f$ a.e. on R, then $\phi \le \underline{f}$ a.e. on R. Similarly, if ψ is any step function with $f \le \psi$ a.e. on R, then $\overline{f} \le \psi$ a.e. on R.

IX.3.10.4. PROPOSITION. Let R be a rectangle, and f a bounded function on R. Let $(\mathcal{P}_n)$ be a sequence of full partitions of R such that $\|\mathcal{P}_n\| \to 0$ as $n \to \infty$. Then $\phi(f,\mathcal{P}_n) \to \underline{f}$ a.e. and $\psi(f,\mathcal{P}_n) \to \overline{f}$ a.e. on R.

PROOF: The union of the degenerate rectangles in $\mathcal{P}_n$ has measure 0 (zero content) for each n. Hence if A is the set of points of R which are interior points of nondegenerate rectangles of $\mathcal{P}_n$ for every n, then $R \setminus A$ has measure 0. Fix $\boldsymbol{a} \in A$, and let $\epsilon > 0$. There is a $\delta > 0$ such that $f(\boldsymbol{x}) > \underline{f}(\boldsymbol{a}) - \epsilon$ for all $\boldsymbol{x} \in R \cap B_\delta(\boldsymbol{a})$. Then there is a N with $\|\mathcal{P}_n\| < \delta$ for $n \ge N$; for such an n the nondegenerate rectangle in $\mathcal{P}_n$ containing $\boldsymbol{a}$ is contained in $B_\delta(\boldsymbol{a})$ and thus $\underline{f}(\boldsymbol{a}) - \epsilon \le \phi(f,\mathcal{P}_n)(\boldsymbol{a}) \le \underline{f}(\boldsymbol{a})$. Therefore $\phi(f,\mathcal{P}_n)(\boldsymbol{a}) \to \underline{f}(\boldsymbol{a})$. So $\phi(f,\mathcal{P}_n) \to \underline{f}$ pointwise on A. Similarly, $\psi(f,\mathcal{P}_n) \to \overline{f}$ pointwise on A.

IX.3.10.5. In fact, it is easily verified () that $\underline{f}$ is precisely the supremum of all step functions ϕ satisfying $\phi \le f$, and $\overline{f}$ the infimum of all step functions ψ satisfying $f \le \psi$.

We can then add another equivalent condition for Riemann integrability:

IX.3.10.6. THEOREM. Let R be a rectangle in $\mathbb{R}^p$, and $f : \mathbb{R} \to \mathbb{R}$ a bounded function. The following are equivalent:

(i) f is Riemann integrable on R.

(ii) There are sequences (ϕ_n) and (ψ_n) of step functions supported on R with

$$\phi_n \le \phi_{n+1} \le f \le \psi_{n+1} \le \psi_n$$

on R for all n and

$$\lim_{n\to\infty} \int_R \phi_n \, dV = \sup_n \int_R \phi_n \, dV = \inf_n \int_R \psi_n \, dV = \lim_{n\to\infty} \int_R \psi_n \, dV \ .$$

(iii) There are sequences (ϕ_n) and (ψ_n) of step functions supported on R with

$$\phi_n \le \phi_{n+1} \le f \le \psi_{n+1} \le \psi_n$$

on R for all n and $\phi_n \to f$ a.e. and $\psi_n \to f$ a.e. on R.

(iv) We have $\underline{f} = \overline{f}$ a.e. on R (and in particular $\underline{f} = f$ a.e. and $\overline{f} = f$ a.e. on R).

PROOF: (i)–(iii) are equivalent by IX.3.10.1.. If ϕ_n, ψ_n are as in (iii), we have

$$\phi_n \le \underline{f} \le f \le \overline{f} \le \psi_n \text{ a.e.}$$

for each n; since $\phi_n \to f$ a.e. and $\psi_n \to f$ a.e. we have $\underline{f} = f = \overline{f}$ a.e.
Conversely, if (iv) is satisfied, let $(\mathcal{P}_n)$ be a sequence of full partitions of R such that $\|\mathcal{P}_n\| \to 0$ as $n \to \infty$, such that $\mathcal{P}_{n+1}$ refines $\mathcal{P}_n$ for all n. Then by IX.3.10.4. $\phi(f, \mathcal{P}_n) \nearrow f$ a.e. and $\psi(f, \mathcal{P}_n) \searrow f$ a.e. on R, so (iii) is satisfied. 👍

Noting that $\underline{f} = f = \overline{f}$ precisely at the points of continuity of f, we obtain LEBESGUE'S characterization:

IX.3.10.7. COROLLARY. Let R be a rectangle in $\mathbb{R}^p$, and $f : R \to \mathbb{R}$. Then f is Riemann integrable on R if and only if f is continuous almost everywhere and bounded on R.

This can be quickly extended to Jordan regions:

IX.3.10.8. COROLLARY. Let E be a Jordan region in $\mathbb{R}^p$, and $f : E \to \mathbb{R}$. Then f is Riemann integrable on E if and only if f is continuous almost everywhere and bounded on E.

PROOF: If f is continuous on E except on a set $A \subseteq E$ of measure zero, then $f\chi_E$ is continuous on any rectangle R containing E except on $A \cup \partial E$, which is a set of measure zero. 👍

Here is another useful observation:

IX.3.10.9. PROPOSITION. Let E be a Jordan region in $\mathbb{R}^n$, and $f : \bar{E} \to \mathbb{R}$ a bounded function. Then the following are equivalent:

(i) f is Riemann integrable over E.

(ii) f is Riemann integrable over $\bar{E}$.

(iii) f is Riemann integrable over E^o.

If these are satisfied, then

$$\int_E f\,dV = \int_{\bar{E}} f\,dV = \int_{E^o} f\,dV\ .$$

PROOF: (ii) $\Rightarrow$ (i) $\Rightarrow$ (iii) follow from IX.3.3.4..
(iii) $\Rightarrow$ (ii): If f is Riemann integrable over E^o, then it is continuous on E^o except on a set $A \subseteq E^o$ of measure zero. Thus f is continuous on $\bar{E}$ except on $A \cup \partial E$, which is a set of measure zero.
The last formula follows from IX.3.3.11. and IX.3.3.9.. 👍

IX.3.11. The Extended Riemann Integral

IX.3.11.1. In this section, we will extend the Riemann integral from Jordan regions to arbitrary open sets in $\mathbb{R}^p$. This extended integral will serve several purposes:

(i) It covers most (but not all!) instances of what can be called "improper Riemann integrals."

(ii) It makes the Change of Variables Theorem simpler and more natural.

(iii) It provides a good conceptual transition from the Riemann integral to the Lebesgue integral without measure-theoretic technicalities.

IX.3.11.2. We will only define the extended Riemann integral over *open* subsets of $\mathbb{R}^p$; thus it is not technically a generalization of the ordinary Riemann integral over Jordan regions since a Jordan region is not necessarily open. However, the distinction is unimportant since integration over open Jordan regions includes all essential features of the theory due to IX.3.2.23. and IX.3.10.9.. Following [**?**] (on which our exposition is partly based), we will call this theory the "extended Riemann integral" rather than the "improper Riemann integral" since it does not include all integrals which can be called "improper Riemann integrals."

IX.3.11.3. DEFINITION. (i) Let U be an open set in $\mathbb{R}^p$, and let $f : U \to \mathbb{R}$ be a nonnegative function. Then f is *Riemann integrable over U in the extended sense*, or *extended Riemann integrable over U*, if

(a) f is Riemann integrable over every compact Jordan region contained in U and

(b) $\sup\left[\int_E f\,dV\right] < \infty$, where the supremum is over all compact Jordan regions E contained in U.

If f is extended Riemann integrable over U, write

$$\int_U f\,dV = \sup\left[\int_E f\,dV\right]$$

where the supremum is over all compact Jordan regions E contained in U.

(ii) Let U be an open set in $\mathbb{R}^p$, and let $f : U \to \mathbb{R}$ be a function. Then f is *Riemann integrable over U in the extended sense* if both $f_+ = \max(f, 0)$ and $f_- = \max(-f, 0)$ are extended Riemann integrable over U. If this is the case, set

$$\int_U f\,dV = \int_U f_+\,dV - \int_U f_-\,dV\ .$$

Generating Sequences

Here is a crucial technical tool for the study of the extended integral:

IX.3.11.4. DEFINITION. Let U be an open set in $\mathbb{R}^p$. A *generating sequence* for U is an (increasing) sequence (E_n) of compact Jordan regions in U such that $U = \cup_{n=1}^{\infty} E_n$ and $E_n \subseteq E_{n+1}^o$ for all n.

IX.3.11.5. PROPOSITION. Let U be an open set in $\mathbb{R}^p$, and (E_n) a generating sequence for U. Then

(i) If K is any compact subset of U, then there is an n such that $K \subseteq E_n^o$.

(ii) If f is a nonnegative function on U, then f is extended Riemann integrable on U if and only if it is Riemann integrable on each E_n and $\sup_n \left[\int_{E_n} f\,dV\right] < \infty$. In this case,

$$\int_U f\,dV = \sup_n \left[\int_{E_n} f\,dV\right] .$$

If f is a general extended Riemann integrable function on U, then

$$\int_U f\,dV = \lim_{n\to\infty} \left[\int_{E_n} f\,dV\right] .$$

PROOF: (i) is an immediate consequence of compactness: $\{E_n^o\}$ is an open cover of K, which has a finite subcover.

(ii) If f is nonnegative and Riemann integrable on each E_n, then for any compact Jordan region $K \subseteq U$ we have, for $K \subseteq E_n^o$, f is Riemann integrable over K by IX.3.3.4. and

$$\int_K f\,dV \le \int_{E_n} f\,dV \le \int_U f\,dV$$

so taking the supremum over all K, we obtain the first formula. We then have that

$$\left(\int_{E_n} f\,dV\right)$$

is a bounded nondecreasing sequence converging to its supremum. The second formula then follows for general f by linearity. 👍

IX.3.11.6. PROPOSITION. Let U be an open set in $\mathbb{R}^p$. Then there exists a generating sequence for U.

PROOF: Let ρ be the standard metric on $\mathbb{R}^p$ (). Then

$$U = \{\boldsymbol{x} \in \mathbb{R}^p : \rho(\boldsymbol{x}, U^c) > 0\} .$$

Set

$$A_n = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\| \le n, \rho(\boldsymbol{x}, U^c) \ge 1/n\} .$$

Then A_n is compact, $A_n \subseteq A_{n+1}^o$, and $\cup_n A_n = U$. However, A_n is not necessarily a Jordan region since its boundary does not necessarily have zero content. But for each $\boldsymbol{x} \in A_n$, there is an open ball $B_{\boldsymbol{x}}$ centered at $\boldsymbol{x}$ such that $\bar{B}_{\boldsymbol{x}} \subseteq A_{n+1}^o$. The $B_{\boldsymbol{x}}$ form an open cover of A_n, which has a finite subcover; let E_n be the union of finitely many $\bar{B}_{\boldsymbol{x}}$ whose interiors cover A_n. Then E_n is a compact Jordan region, and we have $A_n \subseteq E_n \subseteq A_{n+1}^o \subseteq E_{n+1}^o$ for each n. 👍

Compare with ().

IX.3.11.7. Thus in computing the extended Riemann integral or proving its properties, we may work with a generating sequence. There are in general many generating sequences, and they are all equally good in principle; in fact, if (E_n) and (E'_n) are two generating sequences, by IX.3.11.5.(i) there is an increasing sequence (k_n) such that $E_n \subseteq E'_{k_n}$ and $E'_n \subseteq E_{k_n}$ for each n, so any two generating sequences are "almost the same." (Note too that $(E_n \cup E'_n)$ is also a generating sequence.) But in practice, in any situation there may be certain generating sequences which are easier to work with than others, and we are free to choose any convenient one.

IX.3.11.8. EXAMPLES. (i) Let $p = 1$ and $U = \mathbb{R}$. Set $f(x) = \frac{1}{x^2+1}$. We may take $E_n = [-n, n]$. Since f is continuous, it is Riemann integrable on each E_n. We have

$$\int_{E_n} f\,dV = \int_{-n}^{n} \frac{1}{x^2+1}\,dx = 2\tan^{-1} n$$

for each n. So f is extended Riemann integrable over $\mathbb{R}$, and $\int_{\mathbb{R}} f\,dV = \pi$. This is the usual improper Riemann integral

$$\int_{-\infty}^{+\infty} \frac{1}{x^2+1}\,dx\ .$$

In fact, the extended Riemann integral of any *nonnegative* function over $\mathbb{R}$ (or $(0, +\infty)$) is exactly the improper Riemann integral as defined in ().

(ii) Let $p = 1$, $U = (0, 1)$, and $f(x) = x^{-1/2}$. We may take $E_n = \left[\frac{1}{n}, 1 - \frac{1}{n}\right]$ for each n. Since f is continuous on U, it is Riemann integrable on each E_n, and

$$\int_{E_n} f\,dV = \int_{1/n}^{1-1/n} x^{-1/2}\,dx = 2\sqrt{1 - \frac{1}{n}} - 2\sqrt{\frac{1}{n}}\ .$$

Thus f is extended Riemann integrable over $(0, 1)$ and $\int_{(0,1)} f\,dV = 2$. This is the ordinary improper Riemann integral

$$\int_0^1 x^{-1/2}\,dx\ .$$

Thus an extended Riemann integrable function need not be bounded.

(iii) Let $p = 1$, $U = (0, +\infty)$, and $f(x) = \frac{\sin x}{x}$. We have that

$$\int_0^{+\infty} \frac{\sin x}{x}\,dx$$

exists as an improper Riemann integral in the sense of (). However,

$$\int_0^{2\pi n} \left(\frac{\sin x}{x}\right)_+ dx \geq c + \sum_{k=2}^{n} \frac{2}{\pi k}$$

where c is a constant, which is unbounded in n, so f is not extended Riemann integrable on U.

(iv) Here is an example where making a good choice of generating sequence is very helpful. Let $p = 2$, $U = \mathbb{R}^2$, and $f(x, y) = e^{-(x^2+y^2)}$. If we let $E_n = [-n, n] \times [-n, n]$, then (E_n) is a generating sequence for $\mathbb{R}^2$; but we have to calculate

$$\int_{E_n} f\,dV = \int_{-n}^{n}\int_{-n}^{n} e^{-(x^2+y^2)}\,dy\,dx$$

or the iterated integral in the other order, which cannot be calculated exactly. However, suppose instead we let E'_n be the closed disk of radius n centered at $\mathbf{0}$. We can easily calculate the integral over E'_n by changing to polar coordinates ():

$$\int_{E'_n} f\,dV = \int_0^{2\pi}\int_0^n e^{-r^2}\,r\,dr\,d\theta = \int_0^{2\pi}\left[\frac{1}{2} - \frac{e^{-n^2}}{2}\right]d\theta = \left[\frac{1}{2} - \frac{e^{-n^2}}{2}\right](2\pi) = \pi(1-e^{-n^2})$$

so f is extended Riemann integrable over $\mathbb{R}^2$ and $\int_{\mathbb{R}^2} f\,dV = \pi$.

IX.3.11.9. PROPOSITION. Let U be an open set in $\mathbb{R}^p$. Then

(i) If $f : U \to \mathbb{R}$ is Riemann integrable on each Jordan region in U, then f is extended Riemann integrable on U if and only if $|f|$ is.

(ii) If f and g are extended Riemann integrable functions on U and $\alpha \in \mathbb{R}$, then $f+g$ and αf are extended Riemann integrable on U, and

$$\int_U (f+g)\,dV = \int_U f\,dV \quad \text{and} \quad \int_U \alpha f\,dV = \alpha\int_U f\,dV\ .$$

PROOF: Fix a generating sequence (E_n) in U. We first prove (ii) if f, g, and α are nonnegative. For each n,

$$\int_{E_n}(f+g)\,dV = \int_{E_n} f\,dV + \int_{E_n} g\,dV \le \int_U f\,dV + \int_U g\,dV$$

so $f+g$ is extended Riemann integrable on U by IX.3.3.6. and IX.3.11.5., and we have

$$\int_U (f+g)\,dV = \lim_{n\to\infty}\int_{E_n}(f+g)\,dV = \lim_{n\to\infty}\left[\int_{E_n} f\,dV + \int_{E_n} g\,dV\right]$$

$$= \lim_{n\to\infty}\left[\int_{E_n} f\,dV\right] + \lim_{n\to\infty}\left[\int_{E_n} g\,dV\right] = \int_U f\,dV + \int_U g\,dV\ .$$

If $\alpha \ge 0$, then αf is nonnegative, and for each n

$$\int_{E_n}\alpha f\,dV = \alpha\int_{E_n} f\,dV \le \alpha\int_U f\,dV$$

so αf is extended Riemann integrable on U, and

$$\int_U \alpha f\,dV = \lim_{n\to\infty}\left[\int_{E_n}\alpha f\,dV\right] = \lim_{n\to\infty}\left[\alpha\int_{E_n} f\,dV\right] = \alpha\lim_{n\to\infty}\left[\int_{E_n} f\,dV\right] = \alpha\int_U f\,dV\ .$$

For (i), if f is extended Riemann integrable on U, then by definition f_+ and f_- are extended Riemann integrable on U, hence so is $|f| = f_+ + f_-$. Conversely, if $|f|$ is extended Riemann integrable on U, then so are f_+ and f_- since $0 \le f_+, f_- \le |f|$ ($f_+ = \frac{|f|+f}{2}$ and $f_- = \frac{|f|-f}{2}$ are Riemann integrable over each E_n).

To finish (ii), if f and g are extended Riemann integrable on U, so is $|f|+|g|$, and hence so is $|f+g|$ since $|f+g| \le |f|+|g|$ (obviously $|f+g|$ is Riemann integrable on each E_n). We then have

$$\int_U (f+g)\,dV = \lim_{n\to\infty} \int_{E_n} (f+g)\,dV = \lim_{n\to\infty} \left[\int_{E_n} f\,dV + \int_{E_n} g\,dV \right]$$

$$= \lim_{n\to\infty} \left[\int_{E_n} f\,dV \right] + \lim_{n\to\infty} \left[\int_{E_n} g\,dV \right] = \int_U f\,dV + \int_U g\,dV$$

as before. For scalar multiples, there is a slight subtlety. If $\alpha \ge 0$, we have $(\alpha f)_+ = \alpha f_+$ and $(\alpha f)_- = \alpha f_-$ and

$$\int_U \alpha f\,dV = \int_U (\alpha f)_+\,dV - \int_U (\alpha f)_-\,dV = \int_U \alpha f_+\,dV - \int_U \alpha f_-\,dV = \alpha \left[\int_U f_+\,dV - \int_U f_-\,dV \right] = \alpha \int_U f\,dV\,.$$

But if $\alpha < 0$, we have $(\alpha f)_+ = |\alpha| f_-$ and $(\alpha f)_- = |\alpha| f_+$, so

$$\int_U \alpha f\,dV = \int_U (\alpha f)_+\,dV - \int_U (\alpha f)_-\,dV = \int_U |\alpha| f_-\,dV - \int_U |\alpha| f_+\,dV$$

$$= |\alpha| \left[\int_U f_-\,dV - \int_U f_+\,dV \right] = -|\alpha| \int_U f\,dV = \alpha \int_U f\,dV\,.$$

Some other properties of the extended Riemann integral:

IX.3.11.10. PROPOSITION. Let U and W be open sets in $\mathbb{R}^p$, with $U \subseteq W$. If $f : W \to \mathbb{R}$ is extended Riemann integrable on W, then $f|_U$ is extended Riemann integrable on U. If f is additionally nonnegative on W, then

$$0 \le \int_U f\,dV \le \int_W f\,dV\,.$$

PROOF: Suppose f is nonnegative and extended Riemann integrable on W. If K is a compact Jordan region contained in U, then $K \subseteq W$, so f is Riemann integrable over K and $0 \le \int_K f\,dV \le \int_W f\,dV$. Thus the supremum of $\int_K f\,dV$ over all such K is finite and $\le \int_W f\,dV$, so f is extended Riemann integrable over U and $0 \le \int_U f\,dV \le \int_W f\,dV$. The case of general f now follows by applying this to f_+ and f_-.

IX.3.11.11. PROPOSITION. Let U and W be open sets in $\mathbb{R}^p$, and $f : U \cup W \to \mathbb{R}$. Then f is extended Riemann integrable on $U \cup W$ if and only if f is extended Riemann integrable on both U and W. If this is the case, f is automatically extended Riemann integrable on $U \cap W$ (where we make the convention that every function is integrable over $\emptyset$, with zero integral), and

$$\int_{U\cup W} f\,dV = \int_U f\,dV + \int_W f\,dV - \int_{U\cap W} f\,dV\,.$$

PROOF: If f is extended Riemann integrable on $U \cup W$, then it is extended Riemann integrable on U and W by IX.3.11.10.. So suppose f is extended Riemann integrable on U and W. Then f is extended Riemann integrable on $U \cap W$ by IX.3.11.10..

Let (E_n) and (E'_n) be generating sequences for U and W respectively. We claim that $(E_n \cup E'_n)$ and $(E_n \cap E'_n)$ are generating sequences for $U \cup V$ and $U \cap V$ respectively. $E_n \cup E'_n$ and $E_n \cap E'_n$ are compact Jordan regions (conventionally calling $\emptyset$ a compact Jordan region) by IX.3.2.25.. We have

$$E_n \cup E'_n \subseteq E^o_{n+1} \cup {E'}^o_{n+1} \subseteq (E_{n+1} \cup E'_{n+1})^o$$

$$E_n \cap E'_n \subseteq E^o_{n+1} \cap {E'}^o_{n+1} \subseteq (E_{n+1} \cap E'_{n+1})^o \ .$$

We clearly have $\cup_{n=1}^{\infty}(E_n \cup E'_n) = U \cup W$. If $\boldsymbol{x} \in U \cap W$, then $\boldsymbol{x} \in E_{n_1}$ and $\boldsymbol{x} \in E_{n_2}$ for some n_1, n_2. If $n = \max(n_1, n_2)$, then $\boldsymbol{x} \in E_n \cap E'_n$. Thus $\cup_{n=1}^{\infty}(E_n \cap E'_n) = U \cap W$.

For each n, we have that f is Riemann integrable on E_n and on E'_n, hence also on $E_n \cup E'_n$ and $E_n \cap E'_n$ by (), and

$$\int_{E_n \cup E'_n} f\,dV = \int_{E_n} f\,dV + \int_{E'_n} f\,dV - \int_{E_n \cap E'_n} f\,dV$$

for each n. The formula now follows by taking the limit as $n \to \infty$. 👍

The next result extends IX.3.11.5.(ii) and is a version of the Monotone Convergence Theorem.

IX.3.11.12. PROPOSITION. Let (U_n) be an increasing sequence of open sets in $\mathbb{R}^p$, and $U = \cup_{n=1}^{\infty} U_n$. If $f : U \to \mathbb{R}$ is nonnegative, then f is extended Riemann integrable on U if and only if f is extended Riemann integrable on each U_n and $\sup_n \left[\int_{U_n} f\,dV\right] < \infty$. In this case,

$$\int_U f\,dV = \sup_n \int_{U_n} f\,dV \ .$$

PROOF: If f is extended Riemann integrable on U, it is extended Riemann integrable on each U_n by IX.3.11.10., and $\int_{U_n} f\,dV \leq \int_U f\,dV$ for all n.

Conversely, suppose f is extended Riemann integrable on each U_n. Let K be a compact Jordan region in U. Since the U_n give an open cover for K which has a finite subcover, and the U_n are increasing, we have $K \subseteq U_n$ for some n, and thus f is Riemann integrable over K, and

$$\int_K f\,dV \leq \int_{U_n} f\,dV \ .$$

If $M := \sup_n \left[\int_{U_n} f\,dV\right] < \infty$, we have that $\int_K f\,dV \leq M$ for every compact Jordan region $K \subseteq U$; thus by definition f is extended Riemann integrable on U and $\int_U f\,dV \leq M$. Since the other inequality was observed at the beginning of the proof, the proof is complete. 👍

IX.3.11.13. COROLLARY. Let (U_n) be an increasing sequence of open sets in $\mathbb{R}^p$, and $U = \cup_{n=1}^{\infty} U_n$. If $f : U \to \mathbb{R}$ is extended Riemann integrable on U, then f is extended Riemann integrable on each U_n, and

$$\int_U f\, dV = \lim_{n\to\infty} \int_{U_n} f\, dV \ .$$

PROOF: Apply the proposition to f_+ and f_- and replace the supremum by limit. 👍

Properties of Extended Riemann Integrable Functions

IX.3.11.14. We have so far not used (or needed to use) Theorem () characterizing Riemann integrable functions. But this result gives a better picture of extended Riemann integrable functions, although there is not such a clean characterization.

IX.3.11.15. PROPOSITION. Let U be an open set in $\mathbb{R}^p$, and let f be an extended Riemann integrable function on U. Then f is continuous almost everywhere on U.

PROOF: Let (E_n) be a generating sequence for U. Since f is Riemann integrable on E_n, by () f is continuous on E_n except on a subset A_n of measure zero. Since $E_n \subseteq E_{n+1}^o$ for all n, it follows that f is continuous on U except on the set $\cup_{n=1}^{\infty} A_n$, which is a set of measure zero (). 👍

IX.3.11.16. Simply being continuous a.e. is not enough for a function to be extended Riemann integrable on U: although it does not need to be bounded on U, it at least has to be bounded on each compact Jordan region in U (this is automatic if the function is continuous on U, but not if it is only continuous a.e.) There is a cleaner way of expressing this:

IX.3.11.17. DEFINITION. Let U be an open set in $\mathbb{R}^p$, and $f : U \to \mathbb{R}$. Then f is *locally bounded* on U if for every $\boldsymbol{x} \in U$ there is an open ball $B_{\boldsymbol{x}}$ centered at $\boldsymbol{x}$, contained in U, on which f is bounded.

IX.3.11.18. Note that if f is either bounded or continuous on U, then it is locally bounded on U.

IX.3.11.19. PROPOSITION. Let U be an open set in $\mathbb{R}^p$, and $f : U \to \mathbb{R}$. Then f is bounded on every compact subset of U if and only if it is locally bounded on U.

PROOF: This is a compactness/local compactness argument. If f is bounded on every compact subset of U, and $\boldsymbol{x} \in U$, then there is an open ball B centered at $\boldsymbol{x}$ with $\bar{B} \subseteq U$. Since $\bar{B}$ is compact, f is bounded on B. Thus f is locally bounded on U. Conversely, if f is locally bounded on U and $K \subseteq U$ is compact, for each $\boldsymbol{x} \in K$ there is an open ball $B_{\boldsymbol{x}}$ centered at $\boldsymbol{x}$, contained in U, and a constant $M_{\boldsymbol{x}}$ such that $|f| \leq M_{\boldsymbol{x}}$ on $B_{\boldsymbol{x}}$. The $B_{\boldsymbol{x}}$ form an open cover of K which has a finite subcover $\{B_{\boldsymbol{x}_1}, \ldots, B_{\boldsymbol{x}_n}\}$. If $M = \max(M_{\boldsymbol{x}_1}, \ldots, M_{\boldsymbol{x}_n})$, then $|f| \leq M$ on K. 👍

IX.3.11.20. Corollary. Let U be an open set in $\mathbb{R}^p$, and $f : U \to \mathbb{R}$. Then f is Riemann integrable over every compact Jordan region in U if and only if f is continuous a.e. and locally bounded on U.

IX.3.11.21. Thus if f is extended Riemann integrable on U, then it is continuous a.e. and locally bounded on U. The converse is not quite true: the Riemann integrals of $|f|$ over compact Jordan regions in U also have to be uniformly bounded.

The Extended Integral vs. the Ordinary Integral

IX.3.11.22. If U is an open set which is also a Jordan region, we have defined two integrals over U: the (ordinary) Riemann integral and the extended Riemann integral. They are not the same: only bounded functions are Riemann integrable, but extended Riemann integrable functions can be unbounded (IX.3.11.8.(ii); in fact, for any nonempty U there are unbounded extended Riemann integrable functions on U). And even if f is both Riemann integrable and extended Riemann integrable on U, it is not obvious the two integrals have the same value. We will (temporarily) write $(R)\int_U f\,dV$ and $(E)\int_U f\,dV$ for the two integrals when defined.

IX.3.11.23. Theorem. Let U be an open Jordan region in $\mathbb{R}^p$, and $f : U \to \mathbb{R}$. Then f is Riemann integrable on U if and only if it is bounded on U and extended Riemann integrable on U. If this is the case, the two integrals of f over U are equal.

Proof: First suppose f is nonnegative. If f is Riemann integrable on U, then it is bounded, and Riemann integrable on any compact Jordan region $K \subseteq U$, with

$$(R)\int_K f\,dV \le (R)\int_U f\,dV$$

by IX.3.3.4.; thus f is extended Riemann integrable on U, and

$$(E)\int_U f\,dV \le (R)\int_U f\,dV\ .$$

Conversely, if f is bounded and extended Riemann integrable on U, then it is continuous a.e. on U (IX.3.11.15.), and hence Riemann integrable on U by IX.3.10.8..

It only remains to show (for nonnegative f) that if f is Riemann integrable on U, then $(R)\int_U f\,dV \le (E)\int_U f\,dV$. Let R be a rectangle containing U, and $\mathcal{P}$ a partition of R. Let $R_1, \dots, R_n$ be the rectangles in $\mathcal{P}$ contained in U, and $K = R_1 \cup \cdots \cup R_n$. Then K is a compact Jordan region contained in U. If m_j is the infimum of f on R_j, then we have

$$\underline{\mathcal{S}}(f\chi_U, \mathcal{P}) = \sum_{j=1}^{n} m_j V(R_j) \le (R)\int_K f\,dV \le (E)\int_U f\,dV$$

and the inequality follows by taking the supremum over all $\mathcal{P}$.

For general f, apply the above to f_+ and f_-. 👍

IX.3.12. Exercises

IX.3.12.1. Use the continuous version of Fubini's Theorem (IX.3.4.7.; IX.3.5.2. suffices) to prove the following version of the equality of mixed partials VIII.2.2.4. (compare with VIII.2.2.8.):

THEOREM. Let U be an open set in $\mathbb{R}^2$, and $f : U \to \mathbb{R}$ a function such that $\dfrac{\partial^2 f}{\partial x\,\partial y}$ and $\dfrac{\partial^2 f}{\partial y\,\partial x}$ exist and are continuous on U. Then $\dfrac{\partial^2 f}{\partial x\,\partial y} = \dfrac{\partial^2 f}{\partial y\,\partial x}$ on U.

(a) Show using the Fundamental Theorem of Calculus that if $R = [a,b] \times [c,d] \subseteq U$, then

$$\int_a^b \int_c^d \frac{\partial^2 f}{\partial y\,\partial x}(x,y)\,dy\,dx = \int_c^d \int_a^b \frac{\partial^2 f}{\partial x\,\partial y}(x,y)\,dx\,dy \ .$$

(b) If there is an $(x_0,y_0) \in U$ with $\frac{\partial^2 f}{\partial y\,\partial x}(x_0,y_0) > \frac{\partial^2 f}{\partial x\,\partial y}(x_0,y_0)$, then there is a rectangle R around (x_0,y_0) contained in U such that $\frac{\partial^2 f}{\partial y\,\partial x} - \frac{\partial^2 f}{\partial x\,\partial y} > 0$ on R, and hence

$$\int_R \left[\frac{\partial^2 f}{\partial y\,\partial x} - \frac{\partial^2 f}{\partial x\,\partial y}\right] dV > 0 \ .$$

Draw a similar conclusion if there is an $(x_0,y_0) \in U$ with $\frac{\partial^2 f}{\partial y\,\partial x}(x_0,y_0) < \frac{\partial^2 f}{\partial x\,\partial y}(x_0,y_0)$.

(c) Use Fubini's Theorem to obtain a contradiction from (a) and (b).

(d) Extend to functions of several variables and higher-order mixed partials.

IX.3.12.2. Explain why Theorem IX.3.1.26. is not an immediate corollary of Proposition IX.3.1.16.. Identify what else in addition to IX.3.1.16. is needed, and prove it.

IX.3.12.3. Let A be a Jordan region in $\mathbb{R}^n$, and $f : A \to \mathbb{R}$ a nonnegative continuous function. Let

$$B = \{(\boldsymbol{x}, y) : \boldsymbol{x} \in A, 0 \le y \le f(\boldsymbol{x})\} \subseteq \mathbb{R}^{n+1} \ .$$

Show that B is a Jordan region in $\mathbb{R}^{n+1}$, and that

$$V_{n+1}(B) = \int_A f\,dV_n \ .$$

IX.3.12.4. Let E be the subset of $R = [0,1]^2 \subseteq \mathbb{R}^2$ consisting of the points on the x-axis and, for each rational $x = \frac{p}{q}$ in lowest terms, the points in a Cantor set of positive measure () on the vertical segment

$$\left\{(x,y) : 0 \le y \le \frac{1}{q}\right\} \ .$$

(i) Show that E is a closed set of measure 0, hence of content 0.

(ii) Show that $f = \chi_E$ is Riemann integrable over R, but that $f_{[x]}$ is not Riemann integrable on $[0,1]$ if x is rational.

(iii) Modify this example to one for which $f^{[y]}$ is also not Riemann integrable for y rational.

IX.3.12.5. Evaluate

$$\int_0^1 \int_x^1 e^{-y^2}\, dy\, dx$$

by applying Fubini's Theorem to interchange the order of integration.

IX.3.12.6. Give the details of the proof of Theorem IX.3.4.14..

IX.3.12.7. (a) Prove IX.3.5.10..
(b) Generalize to the case where S is replaced by an arbitrary Jordan region in $\mathbb{R}^q$ (cf. IX.3.4.14.).

IX.4. The Jacobian and Change of Variable

The literature is somewhat inconsistent about the definition of the Jacobian of a differentiable function, and the notation for it (cf. IX.4.2.7.). We will take a fairly general path and use a version of the definition and notation which seems to encapsulate the most information.

IX.4.1. Inner Products on Exterior Powers

In this section, we will mostly consider only finite-dimensional vector spaces over $\mathbb{R}$, although some of the theory works in greater generality.

There are at least two natural ways to define an induced inner product on exterior powers of an inner product space. They are not too different: they are just scalar multiples of each other. We begin with perhaps the most useful one, which has the best geometric interpretation.

IX.4.1.1. Let $(V, \langle\cdot,\cdot\rangle)$ be a (real) inner product space. The inner product defines an isometric linear map $\phi : V \to V^*$ by $[\phi(v)](w) = \langle v, w\rangle$ (). Then ϕ extends to an injective graded algebra-homomorphism $\bigwedge \phi$ from $\bigwedge(V)$ to $\bigwedge(V^*)$: $\bigwedge \phi = \bigoplus_k \bigwedge^k \phi$, where $\bigwedge^k \phi$ is an injective linear map from $\bigwedge^k(V)$ to $\bigwedge^k(V^*)$. Identifying $\bigwedge^k(V^*)$ with a subspace of $(\bigwedge^k(V))^*$ (), we get an induced map $\psi_k : \bigwedge^k(V) \to (\bigwedge^k(V))^*$. (If V is finite-dimensional, then all these maps are isomorphisms.)

IX.4.1.2. Now define a map $\bigwedge^k(V) \times \bigwedge^k(V) \to \mathbb{R}$ by

$$\langle x, y\rangle = [\psi_k(x)](y)$$

for $x, y \in \bigwedge^k(V)$.

IX.4.1.3. Proposition. The map $\langle\cdot,\cdot\rangle_k$ is well defined and an inner product on $\bigwedge^k(V)$.

This inner product has a much more concrete description if V is finite-dimensional. Let $\{v_1, \ldots, v_n\}$ be an ordered orthonormal basis for V. For each strictly increasing k-multiindex $I = (i_1, \ldots, i_k)$, set $v_I = v_{i_1} \wedge \cdots \wedge v_{i_k} \in \bigwedge^k(V)$.

IX.4.1.4. Proposition. The set $B = \{v_I\}$, where I runs over all strictly increasing k-multiindices, is an orthonormal basis for $\bigwedge^k(V)$ (with respect to the inner product $\langle\cdot,\cdot\rangle_k$).

Proof: B is a basis for $\bigwedge^k(V)$ (). We show that B is an orthonormal set (showing again that it is a basis by dimension counting). 👍

IX.4.1.5. Example. Suppose $k = n$ in the above situation. Then $\bigwedge^n(V)$ is one-dimensional, spanned by $v_1 \wedge \cdots \wedge v_n$, which is a unit vector. The inner product on $\bigwedge^n(V) \cong \mathbb{R}$ is the usual one from the identification of $v_1 \wedge \cdots \wedge v_n$ with 1.

Another Inner Product

Although the above inner product is a nice one, and the one we will use in the sequel, there is another inner product which comes from the subspace construction of the exterior power.

IX.4.1.6. Let V be a real inner product space. Then $\mathcal{T}^k(V)$ has a natural induced inner product (), and $\bigwedge^k(V) \cong \mathcal{A}^k(V)$ is a subspace, so the inner product on $\mathcal{T}^k(V)$ restricts to an inner product on $\mathcal{A}^k(V)$ we denote by $(\cdot,\cdot)_k$. (This definition works even if V is not finite-dimensional, but we will restrict to the finite-dimensional case.)

IX.4.1.7. PROPOSITION. Let $\{v_1,\dots,v_n\}$ be an orthonormal basis in V. Let $k \leq n$, and if $I = (i_1,\dots,i_k)$ is a strictly increasing k-multiindex, let $v_I = v_{i_1} \wedge \cdots \wedge v_{i_k}$. Then, if I and J are strictly increasing k-multiindices, we have

$$(v_I, v_J)_k = \begin{cases} \frac{1}{k!} & \text{if} \quad I = J \\ \\ 0 & \text{if} \quad I \neq J \end{cases} .$$

PROOF: We have

$$v_I = \frac{1}{k!} \sum_{\sigma \in S_k} v_{\sigma(i_1)} \otimes \cdots \otimes V_{\sigma(i_k)}$$

and similarly for v_j, so

$$(v_I, v_J)_k = \frac{1}{(k!)^2} \sum_{\sigma \in S_k} \sum_{\tau \in S_k} sign(\sigma) sign(\tau) \langle v_{i_{\sigma(1)}}, v_{i_{\tau(1)}} \rangle \cdots \langle v_{i_{\sigma(k)}}, v_{i_{\tau(k)}} \rangle \ .$$

If $I \neq J$, there is a $j \in J \setminus I$, so every term will have a factor which is the inner product of one of the v_i with v_j, so $(v_I, v_J) = 0$.
If $I = J$, any term with $\tau \neq \sigma$ will have a factor of $\langle v_{i_{\sigma(m)}}, v_{i_{\tau(m)}} \rangle$ for an m with $\sigma(m) \neq \tau(m)$, so only the terms with $\sigma = \tau$ are nonzero. Each of these is 1, and there are $k!$ of these terms. 👍

Thus, since the v_I are an orthonormal basis for $\bigwedge^k(V)$ for $\langle \cdot,\cdot \rangle_k$, we obtain:

IX.4.1.8. PROPOSITION. If V is a finite-dimensional (real) inner product space, then for each k we have $(\cdot,\cdot)_k = \frac{1}{k!}\langle \cdot,\cdot \rangle_k$ on $\bigwedge^k(V)$.

IX.4.1.9. Thus if $\| \cdot \|_\otimes$ denotes the norm on $\bigwedge^k(V)$ from $(\cdot,\cdot)_k$, and $\| \cdot \|$ the ordinary norm from $\langle \cdot,\cdot \rangle_k$, we have

$$\| \cdot \|_\otimes = \frac{1}{\sqrt{k!}} \| \cdot \| \ .$$

To see where the factor of $\sqrt{k!}$ comes from, note that if $\{v_1,\dots,v_m\}$ are orthogonal unit vectors in an inner product space, then

$$\|v_1 + \cdots + v_m\| = \sqrt{m} \ .$$

IX.4.1.10. This suggests that in the identification of $\bigwedge^k(V)$ with $\mathcal{A}^k(V)$ in (), if $\mathbb{F} = \mathbb{R}$ we should use the map $\sqrt{k!}A_k$ instead of A_k, i.e.

$$v_1 \wedge \cdot \wedge v_k \mapsto \frac{1}{\sqrt{k!}} \sum_{\sigma \in S_k} v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(k)}$$

and thus a different normalization of the wedge product than the two standard ones (one "halfway between"), cf. III.11.19.13.. This of course is special to the case $\mathbb{F} = \mathbb{R}$ and does not even make sense for a general field $\mathbb{F}$.

Adjoints

IX.4.1.11. Let V and W be finite-dimensional (real) inner product spaces. If $T \in \mathcal{L}(V, W)$, then T has an adjoint $T^* \in \mathcal{L}(W, V)$. For each k, $\bigwedge^k T$ has an adjoint $(\bigwedge^k T)^* \in \mathcal{L}(\bigwedge^k(W), \bigwedge^k(V))$. T^* also induces linear maps $\bigwedge^k(T^*) \in \mathcal{L}(\bigwedge^k(W), \bigwedge^k(V))$ for each k.

IX.4.1.12. Proposition. If $T \in \mathcal{L}(V, W)$, then $(\bigwedge^k T)^* = \bigwedge^k(T^*)$ for each k.

IX.4.1.13. As a special case (using III.11.17.44.), if T is orthogonal (an isometric bijection), so is $\bigwedge^k T$ for every k.

IX.4.2. The Jacobian(s) of a Linear Transformation

IX.4.2.1. Let V and W be finite-dimensional real inner product spaces, and $T : V \to W$ a linear transformation. Then for each k we have an induced linear transformation $\bigwedge^k T : \bigwedge^k(V) \to \bigwedge^k(W)$. If $dim(V) = n$ and $dim(W) = m$, then $\bigwedge^k T = 0$ for $k > \min(n, m)$, so this is only interesting for $k \leq \min(n, m)$. The most important case is when $k = \min(n, m)$; the case $k < \min(n, m)$ is rarely used even in geometric measure theory.

IX.4.2.2. Definition. The *k-Jacobian* of T is $J^k_T = \| \bigwedge^k T \|$ (operator norm with respect to the induced inner products).

IX.4.2.3. Actually, the inner products $(\cdot, \cdot)_k$ can be used on $\bigwedge^k(V)$ and $\bigwedge^k(W)$ to evaluate J^k_T, since they each differ from $\langle \cdot, \cdot \rangle_k$ by the same factor $k!$. But the geometric interpretation using $(\cdot, \cdot)_k$ is not as nice.

Here is a key example:

IX.4.2.4. Example. Suppose $k = m = n$. Then $\bigwedge^n(V)$ and $\bigwedge^n(W)$ are one-dimensional inner product spaces and can be identified with $\mathbb{R}$ by choosing a unit vector in each. Then $\bigwedge^n T$ is multiplication by a scalar λ. The sign of λ depends on the choices of the unit vectors, but $|\lambda| = \| \bigwedge^n T \| = J^n_T$ is well determined. To evaluate $|\lambda|$, fix ordered orthonormal bases $\{v_1, \ldots, v_n\}$ of V and $\{w_1, \ldots, w_n\}$ of W; then T has an $n \times n$ matrix $A = [T]$ with respect to these ordered bases. Then, if the unit vectors are chosen to be $v_1 \wedge \cdots \wedge v_n$

and $w_1 \wedge \cdots \wedge w_n$ respectively, $\lambda = det(A)$ (), and thus $J_T^n = |det(A)|$. (To see directly that this does not depend on the choice of the ordered orthonormal bases, choosing different ordered orthonormal bases amounts to pre- and/or postmultiplying A by orthogonal matrices, which does not change the determinant except possibly for sign since the determinant of an orthogonal matrix is ± 1.)

The Signed Jacobian

IX.4.2.5. We now specialize to the case $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ (this is not a real restriction, since it essentially just amounts to fixing ordered orthonormal bases for V and W.) Then T has a matrix $A = [T]$ (with respect to the standard bases), and if $m = n$ $J_T^n = |det(A)|$. But $det(A)$ carries more information than $|det(A)|$: the sign of $det(A)$ (if nonzero) indicates whether T is orientation-preserving or orientation-reversing.

IX.4.2.6. DEFINITION. Let $T \in \mathcal{L}(\mathbb{R}^n)$. The *(signed) Jacobian* of T is $J_T = det(T) = det(A)$.

IX.4.2.7. The notation may be slightly confusing: $J_T^n = |J_T|$, not J_T. Some authors define the Jacobian of T to be J_T and others $|J_T|$. Others call A the "Jacobian matrix" of T and J_T (or $|J_T|$) the "Jacobian determinant" of T. Various notations for the Jacobian (however defined) are also used. We have chosen definitions and notation which seem to retain the most information.

IX.4.2.8. The difference between J_T and $J_T^n = |J_T|$ (in the case $n = m$) is closely related to the difference between oriented and unoriented integration (the factor J_T or $|J_T|$ appears in change-of-variables formulas). In geometric measure theory, there are usually no natural orientations, and unoriented integration is frequently used; but integration of differential forms uses oriented integration. In the case where k, m, n are not all equal (mostly just used in geometric measure theory), there are no natural orientations, so there is no natural signed version of J_T^k.

Geometric Interpretation

IX.4.2.9. First consider the case $k = n = m$. By IX.4.2.4., J_T^n is the factor by which volumes are multiplied by under T (). Specifically, if $\{v_1, \ldots, v_n\}$ is an orthonormal basis for V, then $[\bigwedge^n T](v_1 \wedge \cdots \wedge v_n)$ is the volume of the parallelepiped in W with edges $Tv_1, \ldots, Tv_n$. [To see that all volumes are multiplied by a common factor, note that if T is invertible, i.e. $J_T^n > 0$, then T sends Lebesgue measurable sets in $V \cong \mathbb{R}^n$ to Lebesgue measurable sets in W (also identified with $\mathbb{R}^n$ via an orthonormal basis), and we may define a measure μ on $\mathbb{R}^n$ by $\mu(E) = \lambda(T(E))$. Then μ is a translation-invariant measure on $\mathbb{R}^n$ and hence a scalar multiple of Lebesgue measure.]

IX.4.2.10. Now consider the general case where $k \leq \min(n, m)$. Let X be a k-dimensional subspace of V, and Y a k-dimensional subspace of Y containing $T(X)$ ($Y = T(X)$ if T is one-to-one on X). Let S be the restriction of T to X, regarded as an element of $\mathcal{L}(X, Y)$. The embedding of X into V extends to an injective algebra-homomorphism of $\bigwedge X$ into $\bigwedge(V)$ which is isometric on each summand (), and similarly for the embedding of Y into W. Thus, if $\{v_1, \ldots, v_k\}$ is an orthonormal basis for X, then $\|[\bigwedge^k T](v_1 \wedge \cdots \wedge v_k)\|$ is the k-volume of the parallelepiped in W with edges $Tv_1, \ldots, Tv_k$, and is the factor by which k-volumes of general (measurable) sets in X are multiplied by under T. This factor may vary for different X's; we will show that J_T^k is the maximum (supremum) of these factors over all X.

IX.4.2.11. Suppose $V = W$ and T is positive (). Then all eigenvalues of T are nonnegative, and T is diagonalizable (), i.e. there is an ordered orthonormal basis $\{v_1, \ldots, v_n\}$ for V consisting of eigenvectors of T with eigenvalues $\lambda_i \geq 0$ which we may assume are in decreasing (nonincreasing) order. Then the vectors v_I for strictly increasing I form an orthonormal basis for $\bigwedge^k(V)$, and each is an eigenvector for $\bigwedge T$: if $I = (i_1, \ldots, i_k)$, then v_I is an eigenvector with eigenvalue $\lambda_{i_1} \cdots \lambda_{i_k}$. Thus $\bigwedge^k T$ is positive and diagonalized by the v_I, so $\| \bigwedge^k T\|$ is the largest eigenvalue, $\lambda_1 \cdots \lambda_k$, which is $\|[\bigwedge^k T](v_1 \wedge \cdots \wedge v_k)\|$.

IX.4.2.12. For general T, we can apply this argument to T^*T, since by IX.4.1.12. and III.11.17.44.

$$\| \bigwedge^k T\|^2 = \|(\bigwedge^k T)^*(\bigwedge^k T)\| = \|(\bigwedge^k T^*)(\bigwedge^k T)\| = \| \bigwedge^k (T^*T)\|$$

and so

$$\| \bigwedge^k T\| = \sqrt{\lambda_1 \cdots \lambda_k} = \|[\bigwedge^k T](v_1 \wedge \cdots \wedge v_k)\|$$

where $(\lambda_1, \ldots, \lambda_n)$ are the eigenvalues of T^*T in nonincreasing order and $\{v_1, \ldots, v_n\}$ a corresponding orthonormal basis of eigenvectors as above, i.e. J_T^k is the square root of the product of the k largest eigenvalues of T^*T. Geometrically, it is the largest k-volume of a parallelepiped with edges $Tw_1, \ldots, Tw_k$, for $w_1, \ldots, w_k$ an orthonormal set in V of size k.

IX.4.2.13. This argument does not work for a general linear transformation from $\bigwedge^k(V)$ to $\bigwedge^k(W)$ since there are many unit vectors in $\bigwedge^k(V)$ which are not simple. But operators of the form $\bigwedge^k T$ for $T \in \mathcal{L}(V, W)$ are special. Note that the linear map $T \mapsto \bigwedge^k T$ from $\mathcal{L}(V, W)$ to $\mathcal{L}(\bigwedge^k(V), \bigwedge^k(W))$ has small range in general since $dim(\mathcal{L}(\bigwedge^k(V), \bigwedge^k(W)))$ is typically much larger than $dim(\mathcal{L}(V, W))$, e.g. if k is near $n/2$.

Evaluating Jacobians

Now let $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ for convenience, so we have the standard ordered orthonormal bases. Then T is multiplication by an $m \times n$ matrix A.

IX.4.2.14. If $k \leq \min(n, m)$, then by the above we have that J_T^k is the square root of the product of the k largest eigenvalues of A^tA. In the case $k < \min(n, m)$, there may be no better systematic way to compute J_T^k: even if $k = 1$, where $J_T^1 = \|T\|$, there is no formula in terms of the entries of A (XX.12.6.6.).

IX.4.2.15. If $k = n \leq m$, though, J_T^n is just the square root of $det(A^tA)$, which is nothing but $|det(A)|$, giving a simpler formula.

IX.4.2.16. There is another formula in the case $k = n \leq m$. To motivate it, consider the case $k = n = 1 < m$. The matrix A is a column vector $\boldsymbol{w}$, and $J_T^1 = \|T\| = \|\boldsymbol{w}\|$, so $(J_T^1)^2$ is the sum of the squares of the entries in A.

Now suppose $k = n < m$. $\bigwedge^n(\mathbb{R}^n)$ is one-dimensional, spanned by the unit vector $\boldsymbol{e}_1^{(n)} \wedge \cdots \wedge \boldsymbol{e}_n^{(n)}$; the image is a one-dimensional subspace of $\bigwedge^n(\mathbb{R}^m)$, spanned by the vector $\boldsymbol{y} = (T\boldsymbol{e}_1^{(n)}) \wedge \cdots \wedge (T\boldsymbol{e}_n^{(n)})$, and $J_T^n = \|\boldsymbol{y}\|$. The vectors $\boldsymbol{e}_I^{(m)}$, where I ranges over strictly increasing n-multiindices for $\mathbb{R}^m$, form an orthonormal basis for $\bigwedge^n(\mathbb{R}^m)$, so $\|\boldsymbol{y}\|^2$ is the sum of the squares of the coefficients.

If $I = (i_1, \ldots, i_m)$, then the coefficient α_I of $\boldsymbol{y}$ with respect to $\boldsymbol{e}_I^{(m)}$ is the projection $Q_I\boldsymbol{y}$, where Q_I is the one-dimensional projection of $\bigwedge^n(\mathbb{R}^m)$ onto the span of $\boldsymbol{e}_I^{(m)}$, identified with $\mathbb{R}$. Then $Q_I = \bigwedge^n P_I$,

where P_I is the projection of $\mathbb{R}^m$ onto the span of $\{\boldsymbol{e}_{i_1}^{(m)}, \ldots, \boldsymbol{e}_{i_n}^{(m)}\} \cong \mathbb{R}^n$. Thus this coefficient is $J_{T_I}^n$, where $T_I = P_I T$ regarded as an element of $\mathcal{L}(\mathbb{R}^n)$. The matrix A_I of T_I consists of the columns of A corresponding to I, so the coefficient is the determinant of A_I. We thus have:

IX.4.2.17. PROPOSITION. Let $n \leq m$, and $T \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$ with matrix A. Then $(J_T^n)^2$ is the sum of the squares of the determinants of the $n \times n$ minors of A.

IX.4.2.18. In the case $k = m < n$, we may apply the above arguments to T^*, since $J_T^m = J_{T^*}^m$, to obtain that J_T^m is $det(AA^t)$ and is also the square root of the sum of the squares of the determinants of the $m \times m$ minors of A.

IX.4.3. The Jacobian(s) of a Differentiable Function

In this section, we work exclusively in the case $V = \mathbb{R}^n$, $W = \mathbb{R}^m$, mainly for convenience. We will extend the various notions of Jacobian from linear functions to general differentiable functions.

IX.4.3.1. DEFINITION. Let U be an open set in $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^m$ a function, and $\boldsymbol{a} \in U$. If $\boldsymbol{f}$ is differentiable at $\boldsymbol{a}$, then for each k we define

$$J_{\boldsymbol{f}}^k(\boldsymbol{a}) = J_{d\boldsymbol{f}(\boldsymbol{a})}^k$$

and, if $n = m$,

$$J_{\boldsymbol{f}}(\boldsymbol{a}) = J_{d\boldsymbol{f}(\boldsymbol{a})} = det(D\boldsymbol{f}(\boldsymbol{a})) \ .$$

IX.4.3.2. If $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear, then T is differentiable at each $\boldsymbol{a}$ and dT is the constant function T; thus, for each $\boldsymbol{a}$, $J_T(\boldsymbol{a}) = J_T$ and $J_T^k(\boldsymbol{a}) = J_T^k$ for each k, so the notation is consistent.

We will call $J_{\boldsymbol{f}}^k(\boldsymbol{a})$ the *k-Jacobian of $\boldsymbol{f}$ at $\boldsymbol{a}$* and, if $n = m$, $J_{\boldsymbol{f}}(\boldsymbol{a})$ the *Jacobian of $\boldsymbol{f}$ at $\boldsymbol{a}$*. Some authors call $D\boldsymbol{f}(\boldsymbol{a})$ the "Jacobian matrix of $\boldsymbol{f}$ at $\boldsymbol{a}$" and (if $n = m$) $J_{\boldsymbol{f}}(\boldsymbol{a})$ (or $|J_{\boldsymbol{f}}(\boldsymbol{a})|$) the "Jacobian determinant of $\boldsymbol{f}$ at $\boldsymbol{a}$."

Geometric Interpretation

IX.4.3.3. If U is an open set in $\mathbb{R}^n$ and $\boldsymbol{f} : U \to \mathbb{R}^n$ is a $\mathcal{C}^1$ function, and $\boldsymbol{a} \in U$, then $|J_{\boldsymbol{f}}(\boldsymbol{a})|$ is the "local factor" by which volumes of "infinitesimally small" regions around $\boldsymbol{a}$ are multiplied by under $\boldsymbol{f}$ (the sign of $J_{\boldsymbol{f}}(\boldsymbol{a})$, if nonzero, indicates whether $\boldsymbol{f}$ is locally orientation-preserving or orientation-reversing at $\boldsymbol{a}$). More precisely, note that if E is a Jordan region whose closure is contained in U, then $f(E)$ is a Jordan region (), and we have:

IX.4.3.4. PROPOSITION.

$$|J_{\boldsymbol{f}}(\boldsymbol{a})| = \lim_{\epsilon \to 0} \frac{Vol(\boldsymbol{f}(B_\epsilon(\boldsymbol{a})))}{Vol(B_\epsilon(\boldsymbol{a}))}$$

where $B_\epsilon(\boldsymbol{a})$ is the open ball of radius ϵ around $\boldsymbol{a}$ (cubes could also be used). More generally, for any $\epsilon > 0$ there is a $\delta > 0$ such that, whenever E is a Jordan region containing $\boldsymbol{a}$ whose closure is contained in U and whose diameter is less than δ,

$$\Big| Vol(\boldsymbol{f}(E)) - |J_{\boldsymbol{f}}(\boldsymbol{a})| Vol(E) \Big| \leq \epsilon \, Vol(E) \ .$$

IX.4.3.5. Replacing "Volume" by n-dimensional Lebesgue measure and "Jordan region" by "Lebesgue measurable set," the statement remains true.

IX.4.4. Change of Variable

Chapter X

Differential Forms and Vector Analysis

"The real difficulty with forms is imagining what they are."

J. & B. Hubbard[1]

The subject of differential forms can at first be regarded as a way of explaining, justifying, and using Leibniz notation. But it becomes much more than that: by combining Leibniz notation with the formalism of exterior algebra and some elementary multivariable calculus, all of vector analysis, notably Stokes' Theorem (the Fundamental Theorem of Vector Calculus) in its various forms, can be efficiently and elegantly obtained, as well as some significant algebraic topology. And the theory can be nicely extended to smooth manifolds, greatly expanding its scope and usefulness; indeed, it is one of the bases for the subject of differential geometry. The thesis of [Spi65] is that, once the results of vector calculus are obtained (some of which are admittedly not easy), almost the entire difficulty of the theory of differential forms lies in making the proper definitions, developing good notation, and properly phrasing the results; the actual proofs are mostly quite easy.

One big advantage of the differential form approach to vector analysis is that it works essentially identically in any dimension. Vector analysis in dimensions two and three can be done without forms (and is in many references), and gives important theorems like Green's Theorem, the Divergence Theorem, and the classical Stokes' Theorem which are special to those dimensions. But these theorems must each be done separately in that approach. Using forms they are naturally all applications of the general Stokes' Theorem. And they do not have analogs in higher dimensions. Such subjects as electromagnetism and relativity in physics need vector analysis in four dimensions (and string theory needs it in multiple dimensions), which can only reasonably be done with the forms approach.

We will work only in the $\mathcal{C}^\infty$ setting, and for us *smooth* will mean $\mathcal{C}^\infty$. Much of the theory can be done with relaxed differentiability hypotheses, but the definitions and results then tend to be clumsy to state carefully, and the increased generality is of dubious practical benefit. Some authors rather imprecisely take "smooth" to mean "continuously differentiable as many times as necessary" (2 is often enough, sometimes even 1).

[1] [HH99, p. 516].

X.1. Vector Fields

X.1.1. The Tangent and Cotangent Bundle on $\mathbb{R}^n$

X.1.1.1. DEFINITION. Let U be an open set in $\mathbb{R}^n$. The *tangent bundle* of U is

$$TU = U \times \mathbb{R}^n \ .$$

The *cotangent bundle* of U is

$$T^*U = U \times (\mathbb{R}^n)^*$$

where $(\mathbb{R}^n)^*$ is the dual space to $\mathbb{R}^n$ (). Give TU and T^*U the product topology and the product smooth structure; then they are smooth $2n$-manifolds.

Since $(\mathbb{R}^n)^*$ is naturally isomorphic to $\mathbb{R}^n$ (), the tangent and cotangent bundles are also naturally isomorphic and are only notationally different. But it will be very convenient, even crucial, to distinguish between them.

The case $U = \mathbb{R}^n$ is especially important; $T\mathbb{R}^n \cong \mathbb{R}^{2n}$, but we must carefully distinguish the two copies of $\mathbb{R}^n$ and think of $T\mathbb{R}^n$ as $\mathbb{R}^n \times \mathbb{R}^n$, not $\mathbb{R}^{2n}$.

X.1.1.2. It may not be obvious why either of the words in the term "tangent bundle" is used. The "bundle" is used since the tangent (and cotangent) bundles are real vector bundles over U (trivial bundles in this case) in the sense of (). And the "tangent" and "cotangent" make sense if the more general context of () is considered. Note also that these terms have nothing to do with trigonometry.

X.1.1.3. There are continuous (smooth) projections $\pi : TU \to U$ and $\rho : TU \to \mathbb{R}^n$ with $\pi(\boldsymbol{p}, \boldsymbol{v}) = \boldsymbol{p}$, $\rho(\boldsymbol{p}, \boldsymbol{v}) = \boldsymbol{v}$. For $\boldsymbol{p} \in U$, set

$$T_{\boldsymbol{p}}U = \pi^{-1}(\{\boldsymbol{p}\}) = \{(\boldsymbol{p}, \boldsymbol{v}) : \boldsymbol{v} \in \mathbb{R}^n\} \cong \mathbb{R}^n \ .$$

$T_{\boldsymbol{p}}U$ is called the *tangent space* to U at $\boldsymbol{p}$. Similarly, there are projections, also denoted π and ρ, from T^*U to U and $(\mathbb{R}^n)^*$, and $T^*_{\boldsymbol{p}}U = \pi^{-1}(\{\boldsymbol{p}\})$ is the *cotangent space* to U at $\boldsymbol{p}$. (We should write π^* and ρ^* for these projections, but this notation will be needed elsewhere.) For fixed $\boldsymbol{p} \in U$, we will regard $T_{\boldsymbol{p}}U$ and $T^*_{\boldsymbol{p}}U$ as vector spaces over $\mathbb{R}$, and we can identify $T^*_{\boldsymbol{p}}U$ with $(T_{\boldsymbol{p}}U)^*$.

X.1.1.4. We can think of the tangent bundle construction as attaching a separate copy of $\mathbb{R}^n$ at each point of U, in such a way that the copies "vary smoothly" over U. It is a way of carefully codifying the familiar informal procedure of drawing vectors "from $\boldsymbol{p}$ to $\boldsymbol{q}$" with tail at $\boldsymbol{p}$ and head at $\boldsymbol{q}$, and identifying such a vector with the vector from $\boldsymbol{0}$ to $\boldsymbol{q} - \boldsymbol{p}$. For fixed $\boldsymbol{p}$, the tangent space $T_{\boldsymbol{p}}\mathbb{R}^n$ can be thought of as consisting of the vectors with tail at $\boldsymbol{p}$, thinking of $(\boldsymbol{p}, \boldsymbol{q} - \boldsymbol{p})$ as the "vector from $\boldsymbol{p}$ to $\boldsymbol{q}$," i.e. identifying $(\boldsymbol{p}, \boldsymbol{v})$ with the vector $\boldsymbol{v}$ drawn with tail at $\boldsymbol{p}$ (the vector from $\boldsymbol{p}$ to $\boldsymbol{p} + \boldsymbol{v}$).

X.1.2. Sections and Vector Fields

X.1.2.1. Informally, a (smooth) vector field on $\mathbb{R}^n$, or an open set U in $\mathbb{R}^n$, consists of attaching a vector in $\mathbb{R}^n$ at each point of U so that the vectors "vary smoothly." A vector field on U is often defined to be a smooth function from U to $\mathbb{R}^n$, and this is not a bad way to do it, but using the tangent bundle is better:

X.1.2.2. DEFINITION. Let U be an open set in $\mathbb{R}^n$. A *(smooth) section* of TU is a smooth function $\sigma : U \to TU$ such that, for each $\boldsymbol{p} \in U$, $\sigma(\boldsymbol{p}) \in T_{\boldsymbol{p}}U$ (i.e. $\pi \circ \sigma$ is the identity on U). A smooth section of TU is called a *vector field* on U.

The definition of a vector field from X.1.2.1. is then just obtained by composing a section with ρ (the correspondence $\sigma \leftrightarrow \rho \circ \sigma$ is a bijection). It is customary to use letters like X and Y to denote vector fields. If X is a vector field on U, for $\boldsymbol{p} \in U$ we denote by $X_{\boldsymbol{p}}$ the vector $\rho \circ X(\boldsymbol{p})$ in $\mathbb{R}^n$, i.e. $X(\boldsymbol{p}) = (\boldsymbol{p}, X_{\boldsymbol{p}})$.

X.1.2.3. Let $\Sigma(TU)$ be the set of vector fields on U (smooth sections of TU). There are some natural algebraic operations on $\Sigma(TU)$, using the vector space structure on each $T_{\boldsymbol{p}}U$:

If $X, Y \in \Sigma(TU)$, then $X + Y \in \Sigma(TU)$, where $(X+Y)_{\boldsymbol{p}} = X_{\boldsymbol{p}} + Y_{\boldsymbol{p}}$ for $\boldsymbol{p} \in U$.

If $X \in \Sigma(TU)$ and $f \in \mathcal{C}^\infty(U)$, then $fX \in \Sigma(TU)$, where $(fX)_{\boldsymbol{p}} = f(\boldsymbol{p})X_{\boldsymbol{p}}$ for $\boldsymbol{p} \in U$. In particular, if $\alpha \in \mathbb{R}$, then $\alpha X \in \Sigma(TU)$, where $(\alpha X)_{\boldsymbol{p}} = \alpha(X_{\boldsymbol{p}})$.

It is easily checked that these operations make $\Sigma(TU)$ into a $\mathcal{C}^\infty(U)$-module, and in particular a real vector space (infinite-dimensional!).

X.1.2.4. Since TU is a product bundle, there are constant sections: if $\boldsymbol{v} \in \mathbb{R}^n$, define $X^{\boldsymbol{v}}$ by $X^{\boldsymbol{v}}(\boldsymbol{p}) = (\boldsymbol{p}, \boldsymbol{v})$ (i.e. $X^{\boldsymbol{v}}_{\boldsymbol{p}} = \boldsymbol{v}$ for all $\boldsymbol{p} \in U$). In particular, for each standard basis vector $\boldsymbol{e}_j$ in $\mathbb{R}^n$, there is a constant vector field $X^{\boldsymbol{e}_j}$. We write

$$\frac{\partial}{\partial x_j} = X^{\boldsymbol{e}_j}$$

for $1 \le j \le n$, i.e. $\frac{\partial}{\partial x_j}$ is a vector field with $\left(\frac{\partial}{\partial x_j}\right)_{\boldsymbol{p}} = \boldsymbol{e}_j$ for each $\boldsymbol{p} \in U$. The reason for this seemingly off-the-wall notation will become apparent in X.1.4.2..

X.1.2.5. If $X \in \Sigma(TU)$, then for each $\boldsymbol{p} \in U$ the vectors $\{(\boldsymbol{p}, \boldsymbol{e}_1), \ldots, (\boldsymbol{p}, \boldsymbol{e}_n)\}$ form a basis for $T_{\boldsymbol{p}}U$, so there are unique scalars $g_1(\boldsymbol{p}), \ldots, g_n(\boldsymbol{p})$ such that

$$X_{\boldsymbol{p}} = \sum_{j=1}^{n} g_j(\boldsymbol{p})\boldsymbol{e}_j$$

for each $\boldsymbol{p} \in U$, and the g_j give functions from U to $\mathcal{R}$, which are obviously smooth ($g_j = \phi_j \circ \rho \circ X$, where ϕ_j is the projection of $\mathbb{R}^n$ onto the span of $\boldsymbol{e}_j$). Thus

$$X = \sum_{j=1}^{n} g_j \frac{\partial}{\partial x_j}$$

and so the $\frac{\partial}{\partial x_j}$ form a basis for $\Sigma(TU)$ as a $\mathcal{C}^\infty(U)$-module (i.e. $\Sigma(TU)$ is a free $\mathcal{C}^\infty(U)$-module of dimension n).

X.1.3. Change of Basis

X.1.4. Lie Differentiation

X.1.4.1. If X is a vector field on $U \subseteq \mathbb{R}^n$, one can differentiate smooth functions "in the direction of X." If $f \in \mathcal{C}^\infty(U)$, define $\frac{\partial f}{\partial X} : U \to \mathbb{R}$ by

$$\frac{\partial f}{\partial X}(\boldsymbol{p}) = \frac{\partial f}{\partial X_{\boldsymbol{p}}}(\boldsymbol{p})$$

where $\frac{\partial f}{\partial X_{\boldsymbol{p}}}(\boldsymbol{p})$ is the directional derivative of f at $\boldsymbol{p}$ with respect to $X_{\boldsymbol{p}}$ (VIII.2.3.2.; note that $X_{\boldsymbol{p}}$ is not a unit vector in general). The function $\frac{\partial f}{\partial X}$ is called the *Lie derivative* of f in the direction of X. The function $\frac{\partial f}{\partial X}$ is often written $L_X f$ or, even more commonly, Xf (do not confuse Xf with fX which is just the product (X.1.2.3.)).

X.1.4.2. If $X^{\boldsymbol{e}_j} = \frac{\partial}{\partial x_j}$, then

$$X^{\boldsymbol{e}_j} f(\boldsymbol{p}) = \frac{\partial f}{\partial X^{\boldsymbol{e}_j}}(\boldsymbol{p}) = \frac{\partial f}{\partial x_j}(\boldsymbol{p})$$

so $X^{\boldsymbol{e}_j} f = \frac{\partial}{\partial x_j} f = \frac{\partial f}{\partial X^{\boldsymbol{e}_j}} = \frac{\partial f}{\partial x_j}$, explaining the notation for the basic vector fields. More generally, if

$$X = \sum_{j=1}^{n} g_j \frac{\partial}{\partial x_j}$$

with the $g_j \in \mathcal{C}^\infty(U)$, i.e. $X_{\boldsymbol{p}} = (g_1(\boldsymbol{p}), \ldots, g_n(\boldsymbol{p}))$ for $\boldsymbol{p} \in U$, we have, by VIII.3.5.4.,

$$(Xf)(\boldsymbol{p}) = \left[\left(\sum_{j=1}^{n} g_j \frac{\partial}{\partial x_j} \right) f \right] (\boldsymbol{p}) = \nabla f(\boldsymbol{p}) \cdot X_{\boldsymbol{p}} = \sum_{j=1}^{n} g_j(\boldsymbol{p}) \frac{\partial f}{\partial x_j}(\boldsymbol{p})$$

so the notation is consistent; $X \mapsto \frac{\partial}{\partial X}$ is $\mathcal{C}^\infty$-linear, and $Xf = \frac{\partial f}{\partial X} \in \mathcal{C}^\infty(U)$.

X.1.4.3. If $X \in \Sigma(TU)$, then the operator $\frac{\partial}{\partial X} : \mathcal{C}^\infty(U) \to \mathcal{C}^\infty(U)$ is a derivation (), i.e. is $\mathbb{R}$-linear and satisfies the product rule

$$\frac{\partial}{\partial X}(fg) = f \frac{\partial g}{\partial X} + g \frac{\partial f}{\partial X}$$

or

$$X(fg) = f(Xg) + g(Xf)$$

for $f, g \in \mathcal{C}^\infty(U)$.

X.1.4.4. If $1 \leq j, k \leq n$, then $\frac{\partial}{\partial x_j}$ and $\frac{\partial}{\partial x_k}$ commute as operators on $\mathcal{C}^\infty(U)$ by Clairaut's Theorem. But if $X, Y \in \Sigma(TU)$, then $\frac{\partial}{\partial X}$ and $\frac{\partial}{\partial Y}$ do not commute in general. If

$$X = \sum_{j=1}^{n} g_j \frac{\partial}{\partial x_j}, \quad Y = \sum_{j=1}^{n} h_j \frac{\partial}{\partial x_j}$$

we have, for $f \in \mathcal{C}^\infty(U)$,

$$(XY)f = X(Yf) = X\left(\sum_{k=1}^{\infty} h_k \frac{\partial f}{\partial x_k}\right) = \sum_{j=1}^{n}\sum_{k=1}^{n} g_j \frac{\partial}{\partial x_j}\left(h_k \frac{\partial f}{\partial x_k}\right)$$

$$= \sum_{j=1}^{n}\sum_{k=1}^{n} g_j \left(\frac{\partial h_k}{\partial x_j}\frac{\partial f}{\partial x_k} + h_k \frac{\partial^2 f}{\partial x_j \partial x_k}\right) = \sum_{j=1}^{n}\sum_{k=1}^{n} g_j \frac{\partial h_k}{\partial x_j}\frac{\partial f}{\partial x_k} + \sum_{j=1}^{n}\sum_{k=1}^{n} g_j h_k \frac{\partial^2 f}{\partial x_j \partial x_k}$$

$$(YX)f = Y(Xf) = Y\left(\sum_{j=1}^{\infty} g_j \frac{\partial f}{\partial x_j}\right) = \sum_{k=1}^{n}\sum_{j=1}^{n} h_k \frac{\partial}{\partial x_k}\left(g_j \frac{\partial f}{\partial x_j}\right)$$

$$= \sum_{k=1}^{n}\sum_{j=1}^{n} h_k \left(\frac{\partial g_j}{\partial x_k}\frac{\partial f}{\partial x_j} + g_j \frac{\partial^2 f}{\partial x_k \partial x_j}\right) = \sum_{k=1}^{n}\sum_{j=1}^{n} h_k \frac{\partial g_j}{\partial x_k}\frac{\partial f}{\partial x_j} + \sum_{k=1}^{n}\sum_{j=1}^{n} g_j h_k \frac{\partial^2 f}{\partial x_k \partial x_j}$$

$$(XY - YX)f = \sum_{j=1}^{n}\sum_{k=1}^{n} g_j \frac{\partial h_k}{\partial x_j}\frac{\partial f}{\partial x_k} - \sum_{k=1}^{n}\sum_{j=1}^{n} h_k \frac{\partial g_j}{\partial x_k}\frac{\partial f}{\partial x_j}$$

since the second-order terms cancel by Clairaut's Theorem. Interchanging indices j and k in the first double sum, we have

$$(XY - YX)f = \sum_{k=1}^{n}\sum_{j=1}^{n} g_k \frac{\partial h_j}{\partial x_k}\frac{\partial f}{\partial x_j} - \sum_{k=1}^{n}\sum_{j=1}^{n} h_k \frac{\partial g_j}{\partial x_k}\frac{\partial f}{\partial x_j}$$

$$= \sum_{j=1}^{n}\left[\sum_{k=1}^{n}\left(g_k \frac{\partial h_j}{\partial x_k} - h_k \frac{\partial g_j}{\partial x_k}\right)\right]\frac{\partial f}{\partial x_j}$$

and thus

$$[X,Y] := (XY - YX) = \sum_{j=1}^{n}\left[\sum_{k=1}^{n}\left(g_k \frac{\partial h_j}{\partial x_k} - h_k \frac{\partial g_j}{\partial x_k}\right)\right]\frac{\partial}{\partial x_j} \in \Sigma(TU)\ .$$

$[X,Y]$ is called the *Lie bracket* of X and Y.

X.1.4.5. Note that XY and YX are second-order (not vector fields or derivations), but $[X,Y]$ is a vector field (derivation), since the second-order derivative terms cancel. There is a nice geometric interpretation of the Lie bracket (). See also Exercise X.1.7.1..

X.1.4.6. If X and Y are constant vector fields, then $[X,Y] = 0$, i.e. X and Y commute. Thus if we choose a different basis $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ for $\mathbb{R}^n$ and take $\frac{\partial}{\partial y_1}, \ldots, \frac{\partial}{\partial y_n}$ as the corresponding vector fields, then these vector fields commute. Thus Clairaut's Theorem is "coordinate-free" (and in particular does not rely on the fact that $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ is orthonormal). See X.1.7.1.(d) for a generalization.

X.1.5. Integral Curves

X.1.5.1. If X is a vector field on $U \subseteq \mathbb{R}^n$, and $\boldsymbol{p} \in U$, we would like to find a parametrized curve $\gamma(t)$ through $\boldsymbol{p}$ such that $\gamma'(t) = X_{\gamma(t)}$ for all t, i.e. the vector $X_{\boldsymbol{q}}$ is the tangent vector to the curve at the point $\boldsymbol{q}$, for every $\boldsymbol{q}$ on the curve. This is possible, at least locally:

X.1.5.2. THEOREM. Let X be a vector field on $U \subseteq \mathbb{R}^n$, $\boldsymbol{p} \in U$, and $t_0 \in \mathbb{R}$. Then there is an open neighborhood V of $\boldsymbol{p}$, contained in U, and an $\epsilon > 0$ such that, for every $\boldsymbol{q} \in V$ there is a unique parametrized curve $\gamma_{\boldsymbol{q}} : (t_0 - \epsilon, t_0 + \epsilon) \to U$ such that $\gamma_{\boldsymbol{q}}(t_0) = \boldsymbol{q}$ and $\gamma'_{\boldsymbol{q}}(t) = X_{\gamma_{\boldsymbol{q}}(t)}$ for all $t \in (t_0 - \epsilon, t_0 + \epsilon)$. Furthermore, each such $\gamma_{\boldsymbol{q}}$ is $\mathcal{C}^\infty$.

PROOF: This is a straightforward application of XI.1.3.4.. The differential equation

$$\boldsymbol{y}' = X_{\boldsymbol{y}}$$

is an autonomous first-order vector differential equation. To put it in standard form, set

$$X_{\boldsymbol{q}} = \boldsymbol{g}(\boldsymbol{q}) = (g_1(\boldsymbol{q}), \ldots, g_n(\boldsymbol{q}))$$

for $\boldsymbol{q} \in U$; then the differential equation becomes

$$\boldsymbol{y}' = \boldsymbol{g}(\boldsymbol{y})$$

or, in coordinates, the system

$$\begin{aligned} y'_1 &= g_1(y_1, \ldots, y_n) \\ &\vdots \\ y'_n &= g_n(y_1, \ldots, y_n) \end{aligned}$$

and $\boldsymbol{g}$ is $\mathcal{C}^\infty$, hence locally Lipschitz, so the IVP $\boldsymbol{y}' = \boldsymbol{g}(\boldsymbol{y})$, $\boldsymbol{y}(t_0) = \boldsymbol{q}$ has a unique solution, which is $\mathcal{C}^\infty$, in a neighborhood of t_0, for $\boldsymbol{q}$ in a neighborhood of $\boldsymbol{p}$. 👍

X.1.5.3. Since the differential equation is autonomous, the solution does not depend on t_0, in the sense that if $\gamma^0_{\boldsymbol{q}}$ is the solution to the IVP for $t_0 = 0$, then for any t_0 the function $\gamma^{t_0}_{\boldsymbol{q}}$ given by $\gamma^{t_0}_{\boldsymbol{q}}(t) = \gamma^0_{\boldsymbol{q}}(t - t_0)$ is the solution for that t_0. This makes sense, since the image curves (for fixed $\boldsymbol{q}$) are all the same, traversed at the same "speed."

Note that the integral "curve" is degenerate if $X_{\boldsymbol{p}} = \boldsymbol{0}$: $\gamma^{t_0}_{\boldsymbol{p}}$ is the constant function with value $\boldsymbol{p}$.

X.1.5.4. Much less than $\mathcal{C}^\infty$ is needed for the function X (or $\boldsymbol{g}$) to get unique integral curves: $\boldsymbol{g}$ needs only to be locally Lipschitz, not necessarily even $\mathcal{C}^1$. In fact, the subject of integral curves is essentially the same thing as the theory of autonomous systems of first-order differential equations.

X.1.5.5. The integral curve through a point $\boldsymbol{p}$ is only defined locally. Local solutions can be cobbled together into a maximal interval of definition, which is an open interval. The maximal interval of definition of the curve will have a (finite) endpoint in one direction if the curve approaches the boundary of U in that direction, and only then (X.1.6.4., X.1.7.2.). But if the maximal interval goes to $\pm\infty$, the curve does not necessarily approach the boundary of U (X.1.6.4.).

X.1.6. The Local Flow

X.1.6.1. Thus each point $\boldsymbol{p}$ of U has an open neighborhood V which partitions into pairwise disjoint integral curves. These curves "vary continuously," i.e. if we define

$$\phi_t(\boldsymbol{x}) = \gamma^0_{\boldsymbol{x}}(t)$$

for $|t| < \epsilon$ and $\boldsymbol{x} \in V$, then $\phi : (-\epsilon, \epsilon) \times V \to U$ is continuous by XI.1.5.5.. In fact, much more is true:

X.1.6.2. THEOREM. Let X be a vector field on $U \subseteq \mathbb{R}^n$, and $\boldsymbol{p} \in U$. Then there is an open neighborhood V of $\boldsymbol{p}$, an $\epsilon > 0$, and a $\mathcal{C}^\infty$ function $\psi : (-\epsilon, \epsilon) \times V \to U$ such that

(i) For each $t \in (-\epsilon, \epsilon)$, ϕ_t defined by $\phi_t(\boldsymbol{x}) = \psi(t, \boldsymbol{x})$ is a diffeomorphism from V to an open set $V_t \subseteq U$.

(ii) $\phi_0(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in V$.

(iii) If $s, t, s + t \in (-\epsilon, \epsilon)$, then $\phi_{s+t} = \phi_s \circ \phi_t$.

(iv) If $\boldsymbol{q} \in V$, then $\gamma(t) = \phi(t, \boldsymbol{q})$ is the unique integral curve $\gamma^0_{\boldsymbol{q}}(t)$ of X.1.5.2..

X.1.6.3. In other words, "flowing" each point along its integral curve gives a *local flow* on small open subsets of U. This flow is not globally defined in general, even if $U = \mathbb{R}^n$ (X.1.6.4.). Piecing the local flows together into a "maximal" flow is much more problematic than with integral curves (since the ϵ can go to 0 toward the edge of U), and cannot be done satisfactorily in general (but see () for the case of compact manifolds).

X.1.6.4. EXAMPLE. Define a vector field X on $\mathbb{R}^2$ by $X_{(x,y)} = (0, y^2)$. If $\boldsymbol{p} = (x_0, y_0)$ with $y_0 \neq 0$, the integral curve $\gamma^0_{\boldsymbol{p}}$ is given by

$$\gamma^0_{\boldsymbol{p}}(t) = \left(x_0, \frac{1}{c - t}\right)$$

where $c = \frac{1}{y_0}$. If $y_0 > 0$, the trajectory is the (open) vertical ray in the upper half plane with x-coordinate x_0. As $t \to -\infty$, $\gamma^0_{\boldsymbol{p}}(t) \to (x_0, 0)$. But $\gamma^0_{\boldsymbol{p}}(t)$ goes to infinity as $t \to c-$, so the maximal interval on which the integral curve is defined is $(-\infty, c)$. If $y_0 < 0$, so $c = \frac{1}{y_0} < 0$, the trajectory is the ray in the lower half plane with x-coordinate x_0; $\gamma^0_{\boldsymbol{p}}(t) \to (x_0, 0)$ as $t \to +\infty$, and $\gamma^0_{\boldsymbol{p}}(t)$ goes to infinity as $t \to c+$, so the maximal interval on which the integral curve is defined is $(c, +\infty)$. There is no global flow defined on any open interval around 0 since the larger $|y_0|$ is, the faster the integral curve goes to infinity (i.e. $c \to 0$ as $|y_0| \to \infty$).
The integral curve $\gamma^0_{\boldsymbol{p}}$ is the constant function $\boldsymbol{p}$ if $\boldsymbol{p} = (x_0, 0)$.

X.1.7. Exercises

X.1.7.1. Prove that the Lie Bracket of vector fields () has the following properties for vector fields X, Y, Z:
(a) $[X + Y, Z] = [X, Z] + [Y, Z]$ and $[\lambda X, Y] = \lambda[X, Y]$ for $\lambda \in \mathbb{R}$.
(b) $[Y, X] = -[X, Y]$. In particular, $[X, X] = 0$. So from (a), $[X, Y + Z] = [X, Y] + [X, Z]$ and $[X, \lambda Y] = \lambda[X, Y]$ for $\lambda \in \mathbb{R}$.
(c) $[X, [Y, Z]] + [Y, [Z, X]] + [Z, [X, Y]] = 0$ (Jacobi Identity).

Thus the set of smooth vector fields on U has the structure of a Lie algebra from $[\cdot, \cdot]$.

(d) If j is fixed, then

$$\left[\frac{\partial}{\partial x_j}, \sum_{k=1}^{\infty} g_k \frac{\partial}{\partial x_k}\right] = \sum_{k=1}^{\infty} \frac{\partial g_k}{\partial x_j} \frac{\partial}{\partial x_k} \ .$$

X.1.7.2. Let X be a vector field on $U \subseteq \mathbb{R}^n$, and $\boldsymbol{p} \in U$. Suppose the integral curve $\gamma^0_{\boldsymbol{p}}$ is defined on $[0, t_0)$ for some $t_0 > 0$, and suppose the curve $\gamma^0_{\boldsymbol{p}}([0, t_0))$ has a cluster point $\boldsymbol{q} \in U$ as $t \to t_0-$. Show that $\gamma^0_{\boldsymbol{p}}$ can be extended to $[0, t_0 + \epsilon)$ for some $\epsilon > 0$ with $\gamma^0_{\boldsymbol{p}}(t_0) = \boldsymbol{q}$. [Reindex the integral curve so that $t_0 = 0$, and consider the local flow at $\boldsymbol{q}$. Note that $\|X\|$ is bounded near $\boldsymbol{q}$.] Obtain a similar result if $t_0 < 0$.

X.2. Differential Forms

In the original conception of Leibniz notation, an expression like dx was regarded as an "infinitesimal" or "differential." From an equation $y = f(x)$ was obtained the relation $dy = f'(x)\,dx$ (although LEIBNIZ himself did not use functional or prime notation, which was not yet invented), interpreted to mean that at the point x, the infinitesimal (differential) dy was related to the infinitesimal dx by a factor of $f'(x)$, i.e. $\frac{dy}{dx} = f'(x)$ (LEIBNIZ would just say $dy = \frac{dy}{dx}\,dx$), and that

$$y = \int dy = \int f'(x)\,dx\ .$$

These ideas, while a useful heuristic, were too vague to be careful mathematics.

We want to make careful mathematical sense out of expressions like dx, especially in higher dimensions, and what it means to integrate them. Our interpretation may seem unrecognizably different from that of LEIBNIZ, but will give the same formulas and results (rigorously!)

> "The notation df/dx, always a little too tempting, has inspired many (usually meaningless) definitions of dx and df separately, the sole purpose of which is to make the equation
>
> $$df = \frac{df}{dx} \cdot dx$$
>
> work out. If $f : \mathbb{R}^2 \to \mathbb{R}$ then df is *defined*, classically, as
>
> $$df = \frac{\partial f}{\partial x}\,dx + \frac{\partial f}{\partial y}\,dy$$
>
> (whatever dx and dy mean).
>
> [We will make] rigorous definitions which enable us to prove the above equations as theorems. It is a touchy question whether or not these modern definitions represent a real improvement over classical formalism; this the reader must decide for himself [sic]."
>
> *M. Spivak*[2]

(In my opinion, they do.)

X.2.1. Differential Forms on $\mathbb{R}^1$

X.2.1.1. So what should dx, or $f'(x)\,dx$, mean? It should be an *(oriented!) integrand* (whatever that means), i.e. if it is applied to an interval I from a to b (oriented, i.e. we could have $a < b$ or $a > b$),

$$dx(I) = \int_a^b dx = b - a$$

or, more generally,

$$[f'(x)\,dx](I) = \int_a^b f'(x)\,dx = f(b) - f(a)$$

[2] [Spi65, p. 45].

so it should be some kind of object which can be "integrated" over oriented intervals. (Actually, whether the "integrand" is itself oriented is a little obscure; the orientation can be regarded as confined to the interval it is applied to. In our approach, both will be oriented. Orientation leads to what I find is one of the biggest technical difficulties of working with forms: keeping careful track of signs. But it is ultimately very pretty to see how everything works out so neatly.)

X.2.1.2. To follow and motivate the approach we take, we must distinguish between the scalars $\mathbb{R}$ and the one-dimensional vector space $\mathbb{R}^1$, which consists of the 1-tuples (α) or the column 1-vectors $[\alpha]$. The usual identification $\alpha \leftrightarrow (\alpha) \leftrightarrow [\alpha]$ of $\mathbb{R}$ with $\mathbb{R}^1$ uses the standard basis $\{(1)\} = \{[1]\}$ for $\mathbb{R}^1$. This distinction may seem artificial, but turns out to be conceptually crucial, especially when we extend to higher dimensions. Under this distinction, we will regard a closed bounded interval in $\mathbb{R}$ as a compact oriented submanifold with boundary in $\mathbb{R}^1$, i.e. we will integrate forms over such submanifolds.

X.2.1.3. We must also distinguish between $\mathbb{R}^1$ and $(\mathbb{R}^1)^*$: $(\mathbb{R}^1)^*$ is the set of linear functionals on $\mathbb{R}^1$, which assign to each vector in $\mathbb{R}^1$ a scalar. $(\mathbb{R}^1)^*$ is usually also identified with $\mathbb{R}$, identifying the scalar α with the linear functional given by multiplication by α (or more precisely the linear functional $[x] \mapsto \alpha x$). In this connection, if f is a (smooth) function from $U \subseteq \mathbb{R}$ to $\mathbb{R}$, we have two standard points of view for the derivative at a point $p \in U$:

(i) If f is regarded as a function from U to $\mathbb{R}$, then $f'(p)$ is a real number (point of view of elementary one-variable calculus).

(ii) If f is regarded as a function from $U \subseteq \mathbb{R}^1$ to $\mathbb{R}^1$, then the (total) differential $df(p)$ is a linear transformation from $\mathbb{R}^1$ to $\mathbb{R}^1$ (point of view of vector calculus).

We will take a third, somewhat intermediate point of view:

(iii) If f is regarded as a function from $U \subseteq \mathbb{R}^1$ to $\mathbb{R}$, then $df(p) = f'(p)$ is a linear functional on $\mathbb{R}^1$ (element of $(\mathbb{R}^1)^*$), namely the linear functional of multiplication by the scalar $f'(p)$.

Thus we may think of df as a (smooth) function from U to $(\mathbb{R}^1)^*$, whose value at a point $p \in U$ is the linear functional corresponding to multiplication by the scalar $f'(p)$; thus df is a (smooth) section of the cotangent bundle of U. So we are led to the following definition:

X.2.1.4. Definition. Let U be an open subset of $\mathbb{R}^1$. A smooth section of the cotangent bundle T^*U is called a *differential 1-form* on U. Write $\Omega^1(U)$ for the set of differential 1-forms on U.

X.2.1.5. We write df for the differential 1-form on U described above. In particular, we may regard dx as the differential 1-form corresponding to the function $f(x) = x$, i.e. as the constant section of T^*U corresponding to the number 1.

X.2.1.6. Just as for vector bundles, we may add differential 1-forms pointwise and multiply them pointwise by smooth functions; thus $\Omega^1(U)$ becomes a $\mathcal{C}^\infty(U)$-module. It is a free $\mathcal{C}^\infty(U)$-module of rank 1 with basis $\{dx\}$, i.e. every differential 1-form ω on U can be written uniquely as $\omega = g\,dx$ for a $\mathcal{C}^\infty$ function $g : U \to \mathbb{R}$, where $\omega(p)$ is the linear functional corresponding to multiplication by $g(p)$, for each $p \in U$. If G is an antiderivative for g on U, then $\omega = dG$, so every differential 1-form on U is of the form df for some (not quite unique) $f \in \mathcal{C}^\infty(U)$.

X.2.1.7. The word "differential" in the term *differential 1-form* really has no meaning in this context, and is used for historical reasons because of the way the concept and notation arose, and because we use the term *differential* to mean an important construction in the theory (partially described above). We often omit the "differential," i.e. we will just write *1-form* to mean *differential 1-form* (one could more generally define 1-forms $g\,dx$ for functions g with less, even no, differentiability, but we will not consider such forms). Differential 1-forms are also often called *exterior 1-forms* for reasons described below.

Integrating 1-Forms

X.2.1.8. Let U be an open set in $\mathbb{R}^1$, ω a 1-form on U, and I an oriented closed bounded interval in U from a to b (we could have $a < b$ or $a > b$). We want to define the integral of ω over I. We have $\omega = g\,dx$ for some smooth function $g : U \to \mathbb{R}$. Let f be an antiderivative for g on U.

There are (at least) three approaches to defining $\int_I \omega$:

(i) Declare by fiat that

$$\int_I \omega = f(b) - f(a)$$

(this is independent of the choice of f). This approach requires no knowledge about Riemann integration.

(ii) Pretend we haven't yet developed Riemann integration. Develop (or redevelop) the entire theory allowing for oriented intervals: if $a > b$, a partition of the interval from a to b runs from right to left, with the increments Δx_k negative. The entire theory, including the FTC (properly phrased), works in this slight generality. We then define

$$\int_I \omega = \int_a^b g(x)\,dx = f(b) - f(a)$$

where the integral is the Riemann integral extended to this oriented situation.

(iii) First develop the usual theory of Riemann integration as in (). Now parametrize the interval from a to b by taking a smooth function $\phi : [c, d] \to U$ $(c < d)$ with $\phi(c) = a$ and $\phi(d) = b$, and define

$$\int_I \omega = \int_c^d g(\phi(t))\phi'(t)\,dt$$

where the integral is the usual Riemann integral. The change-of-variables formula gives that this number is well defined independent of the choice of the parametrization ϕ, and the FTC gives that $\int_I \omega = f(b) - f(a)$.

X.2.1.9. All three approaches give the same value, $f(b) - f(a)$, for $\int_I \omega$. The first approach is not very satisfactory, though, since it does not extend in any obvious way to the higher-dimensional case and leads to few if any applications. Approaches (ii) and (iii) are much better. They extend nicely to higher dimensions, are somewhat complementary, and lead to many important applications.

0-Forms on $\mathbb{R}^1$

X.2.1.10. If $U \subseteq \mathbb{R}^1$, we also define a *(differential) 0-form* on U to be just an element of $\mathcal{C}^\infty(U)$, i.e. $\Omega^0(U) = \mathcal{C}^\infty(U)$ (this is technically not quite the correct definition, cf. (), but will do for now). Thus if $f \in \Omega^0(U)$, then f has a differential $df \in \Omega^1(U)$, and every 1-form is the differential of a 0-form.

Integrating 0-Forms

X.2.1.11. "Integrating" a 0-form over an oriented 0-submanifold of $\mathbb{R}^1$ is straightforward and simple-minded. A compact 0-submanifold of $U \subseteq \mathbb{R}^1$ is just a finite set $X = \{p_1, \ldots, p_n\}$ in U. An orientation of a point p is a choice σ_p of $+$ or $-$ (or, if you will, ± 1); thus an orientation of X is a choice of sign for each p_k. If f is a smooth function on U, define

$$\int_X f = \sum_{k=1}^{n} \sigma_{p_k} f(p_k) \ .$$

Stokes' Theorem in $\mathbb{R}^1$

X.2.1.12. Now let I be an oriented interval in $U \subseteq \mathbb{R}^1$ from a to b. Set $\partial I = \{a, b\}$ (the boundary of I). Give ∂I the orientation with $\sigma_b = +$ and $\sigma_a = -$ (the *induced orientation* from the orientation of I). If $\omega = f$ is a 0-form on U, so that $d\omega = f'\,dx$, the FTC says that

$$\int_I d\omega = f(b) - f(a)$$

but this is exactly the integral of ω over ∂I; so the FTC can be rephrased

$$\int_I d\omega = \int_{\partial I} \omega$$

for any 0-form ω on U. The general Stokes' Theorem () asserts that this formula holds if U is an open set in $\mathbb{R}^n$, ω is a $(k-1)$-form on U, and I a smooth k-submanifold of U.

X.2.2. Differential 1-Forms and Line Integrals on $\mathbb{R}^n$

X.2.3. Higher Differential Forms on $\mathbb{R}^n$

X.2.3.1. If U is an open set in $\mathbb{R}^n$, for $k \geq 0$ we denote by $\Omega^k(U)$ the set of k-forms on U. $\Omega^k(U)$ is naturally a real vector space, and $\Omega^k(U) = 0$ if $k > n$.

Multi-index Notation

X.2.3.2. If $1 \leq k \leq n$, a *k-multiindex* for $\mathbb{R}^n$ is a k-tuple of natural numbers between 1 and n, e.g. $I = (i_1, \ldots, i_k)$. We normally only consider k-multiindices where the terms are strictly increasing (and in particular distinct), but we do not make this a blanket assumption.

X.2.3.3. If $I = (i_1, \ldots, i_k)$ is a k-multiindex, write dx_I for the k-form $dx_{i_1} \cdots dx_{i_k}$. Then $dx_I = 0$ if the i_m are not all distinct, and $dx_I = (-1)^\sigma dx_J$, where J is the set $\{i_1, \ldots, i_k\}$ in increasing order and σ is the sign of the reordering permutation.

With this notation, a general k-form ω is uniquely of the form

$$\omega = \sum_I f_I \, dx_I$$

where the sum is over k-multiindices in strictly increasing order.

Wedge Product

Although we are generally omitting the wedges in writing differential forms, we should think of them as being there. This allows the operations of exterior algebra to be applied to forms, although the application is not quite immediate. We can take the wedge product of a k-form and an ℓ-form on $\mathbb{R}^n$ to yield a $(k+\ell)$-form.

X.2.3.4. The simplest definition uses the standard basis. Every k-form on $\mathbb{R}^n$ is a sum of terms of the form

$$f\,dx_{i_1}\,dx_{i_2}\cdots dx_{i_k}$$

with the i_m strictly increasing and similarly for ℓ-forms. So we take

$$(f\,dx_{i_1}\,dx_{i_2}\cdots dx_{i_k})\wedge(g\,dx_{j_1}\,dx_{j_2}\cdots dx_{j_\ell}) = fg\,dx_{i_1}\,dx_{i_2}\cdots dx_{i_k}\,dx_{j_1}\,dx_{j_2}\cdots dx_{j_\ell}$$

reordered to strictly increasing order using

$$dx_i\,dx_i = 0 \text{ for all } i \text{ and } dx_i\,dx_j = -dx_j\,dx_i \text{ for all } i \neq j$$

(so the wedge is zero unless the i's and j's are all distinct) and extend to sums by linearity (applying this definition term-by-term). This is consistent with regarding the k-form $f\,dx_{i_1}\cdots dx_{i_k}$ as the wedge of the 0-form f and the k-form $dx_{i_1}\cdots dx_{i_k}$.

X.2.4. Integration of Differential Forms on $\mathbb{R}^n$

X.2.4.1. It is a delicate issue what subsets of $\mathbb{R}^n$ we should (or even reasonably can) define the integral of a k-form over. We should at least be able to integrate over

smooth (oriented) compact k-submanifolds with boundary

k-rectangles

and, more generally, "smooth oriented k-submanifolds with corners." We will at least need all such sets to be compact Jordan regions, and would like the collection of admissible subsets to be closed under diffeomorphisms (in a general sense).

X.2.5. The Differential of a Form

The main feature of differential forms which sets them apart from mere exterior algebra is the differential, the fundamental analysis component.

The Differential of a Function

X.2.5.1. A smooth real-valued function f on a region U in $\mathbb{R}^n$ has a "(total) differential" df defined as in (): if $\boldsymbol{x}_0 \in U$, $df(\boldsymbol{x}_0)$ is the linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^1$ with matrix

$$Df(\boldsymbol{x}_0) = \left[\begin{array}{cccc} \frac{\partial f}{\partial x_1}(\boldsymbol{x}_0) & \frac{\partial f}{\partial x_2}(\boldsymbol{x}_0) & \cdots & \frac{\partial f}{\partial x_n}(\boldsymbol{x}_0) \end{array} \right]$$

which is a linear functional on $\mathbb{R}^n$ (an element of $(\mathbb{R}^n)^*$). Thus df can be regarded as a smooth function from U to $(\mathbb{R}^n)^*$, i.e. a smooth section of the cotangent bundle of U, which is a 1-form. Thus we have three standard pictures of the derivative of a smooth function $f : U \to \mathbb{R}$:

The (total) derivative Df, a function from U to the $1 \times n$ matrices.

The gradient ∇f, a vector field on U.

The differential df, a 1-form on U.

Each of these pictures is useful in various contexts.

X.2.5.2. The 1-form df can of course be written in terms of the standard basis; it is easily checked that

$$df = \frac{\partial f}{\partial x_1}\, dx_1 + \frac{\partial f}{\partial x_2}\, dx_2 + \cdots + \frac{\partial f}{\partial x_n}\, dx_n$$

(Exercise X.2.12.1.) which is certainly the expected formula.

Using this notation, we can state a general version of the Fundamental Theorem of Calculus:

X.2.5.3. THEOREM. Let f be a smooth function on a region $U \subseteq \mathbb{R}^n$, and C be an oriented smooth curve in U from $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$, with $\partial C = \{\boldsymbol{x}_0, \boldsymbol{x}_1\}$ oriented with $\sigma_{\boldsymbol{x}_1} = 1$ and $\sigma_{\boldsymbol{x}_0} = -1$ (the induced orientation). Then

$$\int_C df = \int_{\partial C} f \ .$$

This is a special case of Stokes' Theorem (). It can be immediately proved by taking a smooth parametrization of C and applying the usual FTC and the Chain Rule of vector calculus.

X.2.5.4. We also have the product rule

$$d(fg) = f\, dg + g\, df$$

($f, g \in \mathcal{C}^\infty(U)$) which is easily verified by a simple calculation left to the reader.

Differential of a General Form

X.2.5.5. We will now define the differential (often called the *exterior differential*) of a general form. It is enough to define the differential of the k-form

$$f\,dx_{j_1}\,dx_{j_2}\cdots dx_{j_k} \in \Omega^k(U)$$

to be

$$(df)\wedge(dx_{j_1}\,dx_{j_2}\cdots dx_{j_k})$$

which is a $(k+1)$-form. The differential of a general k-form is then defined by linearity. This defines a map $d:\Omega^k(U)\to\Omega^{k+1}(U)$. (We should technically call this d_k since it is a different map for each k, but the k is customarily omitted.) In multiindex notation, we have

$$d\left(\sum_I f_I\,dx_I\right)=\sum_I (df_I)\wedge dx_I$$

where the sum is over all k-multiindices with strictly increasing terms. In fact, using the observations of X.2.3.3., it is easily seen that this formula holds for general sums over k-multiindices which are not necessarily strictly increasing.

X.2.5.6. We could just make this definition by fiat, and later see that it is the useful one; but it is better to give at least a little motivation. Consider the case $U=\mathbb{R}^2$ and $k=1$. The general 1-form is $\omega=f\,dx+g\,dy$, and the general 2-form is $h\,dx\,dy$. If we want $d\omega$ to be a 2-form with the property that if R is a (solid) rectangle in $\mathbb{R}^2$ with boundary ∂R parametrized once around counterclockwise, then

$$\int_{\partial R}\omega=\int_R d\omega$$

we calculate that h must be $\frac{\partial g}{\partial x}-\frac{\partial f}{\partial y}$ (Exercise X.2.12.3.). Then note that by the above definition

$$d\omega=(df)\wedge dx+(dg)\wedge dy=\left(\frac{\partial f}{\partial x}\,dx+\frac{\partial f}{\partial y}\,dy\right)dx+\left(\frac{\partial g}{\partial x}\,dx+\frac{\partial g}{\partial y}\,dy\right)dy$$

$$=\frac{\partial f}{\partial x}\,dx\,dx+\frac{\partial f}{\partial y}\,dy\,dx+\frac{\partial g}{\partial x}\,dx\,dy+\frac{\partial g}{\partial y}\,dy\,dy=-\frac{\partial f}{\partial y}\,dx\,dy+\frac{\partial g}{\partial x}\,dx\,dy=\left(\frac{\partial g}{\partial x}-\frac{\partial f}{\partial y}\right)dx\,dy\ .$$

See Theorem X.2.5.9. for further motivation.

X.2.5.7. The general coordinate formula for the differential is very complicated and is of doubtful usefulness to even write down. A few more special cases are simple and illuminating:

1-form:

$$d(f\,dx_j)=\left(\frac{\partial f}{\partial x_1}\,dx_1+\cdots+\frac{\partial f}{\partial x_n}\right)\wedge dx_j$$

$$=\frac{\partial f}{\partial x_1}\,dx_1\,dx_j+\cdots+\frac{\partial f}{\partial x_{j-1}}\,dx_{j-1}\,dx_j-\frac{\partial f}{\partial x_{j+1}}\,dx_j\,dx_{j+1}-\cdots-\frac{\partial f}{\partial x_n}\,dx_j\,dx_n\ .$$

A general 1-form is canonically a sum of these.

$(n-1)$-form:

$$d(f\,dx_1\,dx_2\cdots dx_{j-1}\,dx_{j+1}\cdots dx_n) = \left(\frac{\partial f}{\partial x_1}\,dx_1 + \cdots + \frac{\partial f}{\partial x_n}\,dx_n\right) \wedge (dx_1\,dx_2\cdots dx_{j-1}\,dx_{j+1}\cdots dx_n)$$

$$= \frac{\partial f}{\partial x_j}\,dx_j\,dx_1\cdots dx_{j-1}\,dx_{j+1}\cdots dx_n = (-1)^{j-1}\frac{\partial f}{\partial x_j}\,dx_1\cdots dx_n\ .$$

A general $(n-1)$-form is canonically a sum of these.

"constant" form:

$$d(dx_I) = d(1\,dx_{i_1}\cdots dx_{i_k}) = (d1)\wedge(dx_{i_1}\cdots dx_{i_k}) = 0$$

where $I = (i_1,\dots,i_k)$ is an arbitrary k-multiindex and 1 is the constant function 1 (note that $d1 = 0$).

Of course, if ω is an n-form, then $d\omega = 0$.

Properties

X.2.5.8. PROPOSITION. Let U be an open set in $\mathbb{R}^n$.
(i) Let $\omega_1,\omega_2\in\Omega^k(U)$ $(0\le k\le n)$. Then

$$d(\omega_1+\omega_2) = d\omega_1 + d\omega_2\ .$$

(ii) Let $\omega_1\in\Omega^k(U)$ and $\omega_2\in\Omega^\ell(U)$. Then

$$d(\omega_1\wedge\omega_2) = (d\omega_1)\wedge\omega_2 + (-1)^k\omega_1\wedge(d\omega_2)\ .$$

(iii) Let $\omega\in\Omega^k(U)$. Then $d(d\omega) = 0$.

PROOF: (i) is obvious; in fact it is obvious that d is linear (respects both addition and scalar multiplication).
(ii): By linearity, it suffices to prove this for $\omega_1 = f\,dx_I$ and $\omega_2 = g\,dx_J$, where I is a k-multiindex and J an ℓ-multiindex. Then

$$\omega_1\wedge\omega_2 = fg\,(dx_I\wedge dx_J) = fg\,dx_K$$

where for $I = (i_1,\dots,i_k)$ and $J = (j_1,\dots,j_\ell)$ we take K to be the $(k+\ell)$-multiindex $(i_1,\dots,i_k,j_1,\dots,j_\ell)$; K is not necessarily strictly increasing, but by X.2.5.4. and the last observation of X.2.5.5.

$$d(\omega_1\wedge\omega_2) = d(fg)\wedge dx_K = (f\,dg + g\,df)\wedge dx_K = (f\,dg + g\,df)\wedge dx_I\wedge dx_J$$

$$= g\,[df\wedge dx_I]\wedge dx_J + f\,[dg\wedge dx_I\wedge dx_J]$$

$$= g\,[df\wedge dx_I]\wedge dx_J + (-1)^k f\,[dx_I\wedge dg\wedge dx_J]$$

$$= [df\wedge dx_I]\wedge[g\,dx_J] + (-1)^k\,[f\,dx_I]\wedge[dg\wedge dx_J] = (d\omega_1)\wedge\omega_2 + (-1)^k\omega_1\wedge(d\omega_2)\ .$$

(iii) By (ii), we have, for any k-multiindex I,

$$d(d(f\,dx_I)) = d((df)\wedge(dx_I)) = d(df)\wedge dx_I - (df)\wedge d(dx_I)\ .$$

The second term is 0 (cf. X.2.5.7.), so it suffices to show that $d(df) = 0$ for any f.

$$d(df) = d\left(\sum_{j=1}^{n} \frac{\partial f}{\partial x_j}\, dx_j\right) = \sum_{j=1}^{n} \left[\sum_{i<j} \frac{\partial^2 f}{\partial x_i\, \partial x_j}\, dx_i\, dx_j - \sum_{i>j} \frac{\partial^2 f}{\partial x_i\, \partial x_j}\, dx_j\, dx_i\right] .$$

Using Clairaut's Theorem, the terms cancel in pairs, so the sum is 0. 👍

A Characterization

The properties in X.2.5.8., along with the basic formula for 0-forms, uniquely characterize the differential:

X.2.5.9. THEOREM. Let U be an open set in $\mathbb{R}^n$. Then the differential as defined in X.2.5.5. is the unique set of functions from $\Omega^k(U)$ to $\Omega^{k+1}(U)$ $(0 \le k \le n)$ satisfying the properties of X.2.5.8. and for which df $(f \in \mathcal{C}^\infty(U) = \Omega^0(U))$ is as defined in X.2.5.1..

X.2.6. Stokes' Theorem for Rectangles in $\mathbb{R}^n$

X.2.6.1. Our general setup in this section is as follows. Let U be an open set in $\mathbb{R}^n$, and

$$R = \prod_{j=1}^{n} [a_j, b_j]$$

an n-rectangle in U. The boundary ∂R is a union of $2n$ $(n-1)$-faces which we denote 0_j and 1_j $(1 \le j \le n)$, where

$$0_j = \{(x_1, \ldots, x_{j-1}, a_j, x_{j+1}, \ldots, x_n) \in R\}$$
$$1_j = \{(x_1, \ldots, x_{j-1}, b_j, x_{j+1}, \ldots, x_n) \in R\}$$

(the faces for different j overlap in faces of dimension $n-2$). Denote by R_j the product $\prod_{i \ne j} [a_i, b_i] \subseteq \mathbb{R}^{n-1}$.

Induced Orientation

X.2.6.2. The rectangle R inherits a natural orientation from $\mathbb{R}^n$ using the standard basis. The induced orientation on ∂R () is easy to describe: on each $(n-1)$-face it is the one with "outward-pointing normal." On the faces 1_j it is $(-1)^{j-1}$ times the one from the map

$$\pi_{j1} : (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \mapsto (x_1, \ldots, x_{j-1}, b_j, x_{j+1}, \ldots, x_n) : R_j \to 1_j$$

and for 0_j it is the negative of $(-1)^{j-1}$ times the one from the corresponding map

$$\pi_{j0} : (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \mapsto (x_1, \ldots, x_{j-1}, a_j, x_{j+1}, \ldots, x_n) : R_j \to 0_j$$

so if $f \in \mathcal{C}^\infty(U)$ we have

$$\int_{1_j} f\, dx_1 \cdots dx_{j-1}\, dx_{j+1} \cdots dx_n = (-1)^{j-1} \int_{R_j} f \circ \pi_{j1}\, dV$$

$$\int_{0_j} f\,dx_1\cdots dx_{j-1}\,dx_{j+1}\cdots dx_n = -(-1)^{j-1}\int_{R_j} f\circ\pi_{j0}\,dV$$

where $\int_{R_j}\cdot dV$ is the usual Riemann integral.

Note also that

$$\int_{1_i} f\,dx_1\cdots dx_{j-1}\,dx_{j+1}\cdots dx_n = \int_{0_i} f\,dx_1\cdots dx_{j-1}\,dx_{j+1}\cdots dx_n = 0$$

if $i \neq j$.

Statement and Proof of the Theorem

X.2.6.3. THEOREM. [STOKES' THEOREM FOR RECTANGLES] Let U be an open set in $\mathbb{R}^n$, R an n-rectangle in U, and $\omega \in \Omega^{n-1}(U)$. Give R the natural orientation and ∂R the induced orientation. Then

$$\int_R d\omega = \int_{\partial R}\omega\ .$$

PROOF: By linearity it suffices to prove the theorem for $\omega = f\,dx_1\cdots dx_{j-1}\,dx_{j+1}\cdots dx_n$ for some j, $1 \leq j \leq n$. We then have

$$d\omega = (-1)^{j-1}\frac{\partial f}{\partial x_j}\,dx_1\cdots dx_n$$

(X.2.5.7.). So we have

$$\int_R d\omega = (-1)^{j-1}\int_R \frac{\partial f}{\partial x_j}\,dV\ .$$

Using the above notation, by Fubini's Theorem (IX.3.4.13.) this equals

$$(-1)^{j-1}\int_{R_j}\left[\int_{a_j}^{b_j}\frac{\partial f}{\partial x_j}(x_1,\ldots,x_{j-1},t,x_{j+1},\ldots,x_n)\,dt\right]dV$$

$$= (-1)^{j-1}\int_{R_j}\left[f(x_1,\ldots,x_{j-1},b,x_{j+1},\ldots,x_n) - f(x_1,\ldots,x_{j-1},a,x_{j+1},\ldots,x_n)\right]dV$$

by the Fundamental Theorem of Calculus (IX.2.9.5.). As above, this equals

$$(-1)^{j-1}\int_{R_j} f\circ\pi_{j1}\,dV - (-1)^{j-1}\int_{R_j} f\circ\pi_{j0}\,dV = \int_{1_j}\omega + \int_{0_j}\omega = \int_{\partial R}\omega$$

since $\int_{1_i}\omega = \int_{0_i}\omega = 0$ for $i \neq j$. 👍

X.2.6.4. It is claimed in [Spi65] that Stokes' Theorem is "trivial." While there is some justifiable truth to this statement, it is overblown because (among other reasons) the proof depends on

The Fundamental Theorem of Calculus (IX.2.9.5.)

Fubini's Theorem (IX.3.4.13.)

neither of which is trivial (granted, the proof of Stokes' Theorem is pretty easy modulo these results); and the general Stokes' Theorem, even in $\mathbb{R}^n$ (), also depends on

The Change of Variables Theorem () (unless the integral on chains is defined by fiat)

The Partition of Unity Theorem (XIII.11.10.15.)

which are definitely nontrivial.

X.2.7. Pullbacks of Forms

If U and V are open sets in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $\boldsymbol{f}$ is a smooth map from U to V, there is a systematic way to "pull back" a form ω on V to a form $\boldsymbol{f}^*\omega$ on U respecting wedge products, differentials and, to the greatest extent possible, integrals. Pullbacks are closely related to the Change of Variable Theorem (). There is an elegant coordinate-free definition of the pullback (), but we will take a lower-tech approach using coordinates which will result in more useful explicit formulas.

Pullback of a Function

X.2.7.1. There is an obvious way to pull back an element g of $\Omega^0(V) = \mathcal{C}^\infty(V)$ to an element $\boldsymbol{f}^*(g)$ of $\Omega^0(U) = \mathcal{C}^\infty(U)$: set $\boldsymbol{f}^*(g) = g \circ \boldsymbol{f}$. In fact, this is forced by the requirement that $\boldsymbol{f}^*$ preserve "integration": if $\boldsymbol{x} \in U$ and $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, and σ is $+1$ or -1, the integral of $g \in \mathcal{C}^\infty(V)$ over the oriented 0-submanifold $Y = \{\boldsymbol{y}; \sigma\}$ is $\sigma g(\boldsymbol{y})$. So to preserve integration we need to have the integral $\sigma[\boldsymbol{f}^*(g)](\boldsymbol{x})$ of $\boldsymbol{f}^*(g)$ over the oriented 0-submanifold $\{\boldsymbol{x}; \sigma\}$ of U to be $\sigma g(\boldsymbol{y}) = \sigma g(\boldsymbol{f}(\boldsymbol{x}))$. This must be true for all $\boldsymbol{x} \in U$.

Pullback of a 1-Form

X.2.7.2. Now suppose $\omega \in \Omega^1(V)$. We want to define $\boldsymbol{f}^*\omega \in \Omega^1(U)$ so that $\boldsymbol{f}^*$ is linear and preserves wedge products and differentials. By linearity, it suffices to assume $\omega = g\,dy_j$ for $g \in \Omega^0(V)$ and $1 \le j \le m$. We need

$$\boldsymbol{f}^*(g\,dy_j) = \boldsymbol{f}^*(g \wedge dy_j) = \boldsymbol{f}^*(g) \wedge \boldsymbol{f}^*(dy_j) = \boldsymbol{f}^*(g) \wedge d(\boldsymbol{f}^*(y_j)) \ .$$

But y_j is just the j-th coordinate map of V, $y_j(\alpha_1, \ldots, \alpha_m) = \alpha_j$, so $\boldsymbol{f}^*(y_j) = y_j \circ \boldsymbol{f} = f_j$, where f_j is the j'th coordinate function of $\boldsymbol{f}$. So $\boldsymbol{f}^*\omega$ must be $(g \circ \boldsymbol{f})\,df_j$. In other words, we must define

$$\boldsymbol{f}^*\left(\sum_{j=1}^m g_j\,dy_j\right) = \sum_{j=1}^m (g_j \circ \boldsymbol{f})\,df_j \ .$$

X.2.7.3. Let us check that this definition respects integration in the following sense. Suppose I is an oriented 1-manifold in U, i.e. a simple smooth curve from $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$, and that $J = \boldsymbol{f}(I)$ is an oriented 1-manifold in V, i.e. a simple smooth curve from $\boldsymbol{y}_0 = \boldsymbol{f}(\boldsymbol{x}_0)$ to $\boldsymbol{y}_1 = \boldsymbol{f}(\boldsymbol{x}_1)$ ($\boldsymbol{f}$ does not have to be one-to-one on I). Then, if ω is a 1-form on V,

X.2.7.4. In fact, the definition of the integral of a 1-form is based on a pullback. Suppose U is an open set in $\mathbb{R}^n$ and C a simple oriented smooth curve from $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$. A smooth parametrization of C is a smooth function γ from $[a,b] \subseteq \mathbb{R}^1$ to C with $\gamma(a) = \boldsymbol{x}_0$ and $\gamma(b) = \boldsymbol{x}_1$ (technically we should extend γ to a slightly larger open interval in $\mathbb{R}^1$). If ω is a 1-form on U, then $\int_C \omega$ is *defined* to be $\int_{[a,b]} \gamma^*\omega$ (the Change of Variables Theorem in $\mathbb{R}^1$ then shows that this is well defined). If $\omega = \sum_{j=1}^n g_j\, dx_j$, we have

$$\int_C \omega = \int_{[a,b]} \gamma^*\omega = \int_{[a,b]} \left(\sum_{j=1}^n (g_j \circ \gamma)\, d\gamma_j \right) = \int_a^b \left(\sum_{j=1}^n (g_j(\gamma(t))\, \gamma_j'(t) \right) dt \ .$$

This is exactly the line integral along C of the vector field $\boldsymbol{g} = (g_1, \ldots, g_n)$ associated with ω ().

Pullback of a k-Form

X.2.7.5. The way to extend to a general k-form is then obvious (and forced):

$$\boldsymbol{f}^*(g\, dy_{j_1} \cdots dy_{j_k}) = (g \circ \boldsymbol{f})\, df_{j_1} \wedge \cdots \wedge df_{j_k}$$

(the wedges are often omitted from this formula) and the pullback of a general k-form defined by linearity.

Pullback of an n-Form

X.2.7.6. The coordinate formula for a general pullback has a nice form if $k = n$. First consider the case $k = m = n$. The general n-form on V is $\omega = g\, dy_1 \cdots dy_n$, and

$$f^*\omega = (g \circ \boldsymbol{f})\, df_1 \wedge \cdots \wedge df_n$$

$$= (g \circ \boldsymbol{f}) \left[\left(\frac{\partial f_1}{\partial x_1}\, dx_1 + \cdots \frac{\partial f_1}{\partial x_n}\, dx_n \right) \wedge \cdots \wedge \left(\frac{\partial f_n}{\partial x_1}\, dx_1 + \cdots \frac{\partial f_n}{\partial x_n}\, dx_n \right) \right]$$

When the expression in brackets is expanded by linearity and reordering coordinates to

$$(g \circ \boldsymbol{f})\, h\, dx_1 \cdots dx_n$$

the function h is exactly $det(D\boldsymbol{f})$ (Exercise ()). [In fact, using the coordinate-free approach of (), this is an elegant way to define and characterize the determinant, cf. ().]

X.2.7.7. Now consider the general case where $k = n$ (it will only be interesting if $n \leq m$). If $I = (i_1, \ldots, i_n)$ is an n-multiindex in $\mathbb{R}^m$, let $\pi_I : \mathbb{R}^m \to \mathbb{R}^n$ be

$$\pi_I(x_1, \ldots, x_m) = (x_{i_1}, \ldots, x_{i_n})$$

and $\boldsymbol{f}_I = \pi_I \circ \boldsymbol{f}$. One can think of $\boldsymbol{f}_I$ as the "I'th coordinate function of $\boldsymbol{f}$." Then $\boldsymbol{f}_I : U \to \mathbb{R}^n$, and a calculation () shows that

$$\boldsymbol{f}^* \left(\sum_I g_I\, dy_I \right) = \left[\sum_I (g_I \circ \boldsymbol{f})\, det(D\boldsymbol{f}_I) \right] dx_1 \cdots dx_n \ .$$

(This formula also works if the I's in the sum are not strictly increasing.)

This formula agrees with the previous one if $k = n = m$, and with the line integral if $k = n = 1$.

A Coordinate-free Approach

X.2.8. Chains and the General Stokes' Theorem in $\mathbb{R}^n$

X.2.9. Surfaces in $\mathbb{R}^3$

X.2.10. Divergence and Curl

X.2.11. Analysis on Manifolds

X.2.12. Exercises

X.2.12.1. Verify the formula in X.2.5.2..

X.2.12.2. Give the details of the proof of Theorem X.2.5.3.. (Note that $\int_{\partial C} f = f(\boldsymbol{x}_1) - f(\boldsymbol{x}_0)$.)

X.2.12.3. Carry out the calculation of $\int_{\partial R} \omega$ in X.2.5.6. and verify the formula for h, as follows. Let $R = [a,b] \times [c,d]$.

(a) Calculate

$$\int_{\partial R} (f\,dx + g\,dy) = \int_a^b f(x,c)\,dx + \int_c^d g(b,y)\,dy + \int_b^a f(x,d)\,dx + \int_d^c g(a,y)\,dy$$

$$= \int_c^d [g(b,y) - g(a,y)]\,dy - \int_a^b [f(x,d) - f(x,c)]\,dx = \int_c^d \left[\int_a^b \frac{\partial g}{\partial x}(x,y)\,dx\right] dy - \int_a^b \left[\int_c^d \frac{\partial f}{\partial y}(x,y)\,dy\right] dx\ .$$

(b) Note that dx and dy each stand for two different things in (a) and explain the notation.

(c) Apply Fubini's Theorem (IX.3.4.10.).

X.3. Differential Topology

X.3.1. De Rham Cohomology

X.3.2. Vector Bundles

X.3.3. Topological K-Theory

Chapter XI

Differential Equations

> "The most widely and successfully used concept in mathematical descriptions of scientific problems has been the differential equation, even though success has often been impeded by the fact that differential equations can be extremely difficult to analyze. These difficulties, rather than thwarting progress, have instead stimulated more research and better understanding. As a result, the subject of differential equations contains a wealth of deep theories and subtle methods of analysis."
>
> *E. A. Herman*[1]

The subject of Differential Equations consists of the study of differential equations, how to solve them (exactly or approximately), and the properties the solutions must have. A *differential equation* is an equation involving an unknown function and its derivatives. A *solution* to a differential equation is a sufficiently differentiable function satisfying the equation, usually also satisfying some specified initial or boundary conditions. The unknown function may be a function of one independent variable or more than one. A differential equation involving a function of one independent variable is an *ordinary differential equation* or *ODE*; a differential equation involving a function of more than one independent variable is a *partial differential equation* or *PDE*. One can also consider systems of ODEs or PDEs involving several unknown functions. In general, as might be easily predicted, the theory of PDEs is much more difficult than the theory of ODEs, and most active research in Differential Equations today involves PDEs.

Differential equations are of fundamental importance in both pure and applied mathematics. For example, the basic laws of physics, beginning with $F = ma$, are really differential equations [if $y(t)$ is position at time t, and $F(y)$ is the force at position y, then the equation is

$$my''(t) = F(y(t)) \ .$$

The force could also be a function of both y and t. There are variations possible such as taking both y and F to be vector-valued.]

The subject of Differential Equations is one of the largest areas of mathematics, bridging pure and applied mathematics. Entire books are written on the solutions to important specific differential equations. A considerable part of Numerical Analysis involves computing numerical approximations to solutions of

[1] [**?**, p. 161].

differential equations. We will only treat the most basic aspects of the theory, with an emphasis on the analysis tools needed to obtain them, and we will concentrate on establishing general theorems about existence, uniqueness, and properties of solutions rather than the elaborate techniques used to actually find them explicitly.

> "Newton's *Principia* was impressive, with its revelation of deep mathematical laws underlying natural phenomena. But what happened next was even more impressive. Mathematicians tackled the entire panoply of physics – sound, light, heat, fluid flow, gravitation, electricity, magnetism. In every case, they came up with differential equations that described the physics, often very accurately.
>
> The long-term implications have been remarkable. Many of the most important technological advances, such as radio, television, and commercial jet aircraft depend, in numerous ways, on the mathematics of differential equations. The topic is still the subject of intense research activity, with new applications emerging almost daily. It is fair to say that Newton's invention of differential equations, fleshed out by his successors in the 18th and 19th centuries, is in many ways responsible for the society in which we now live. This only goes to show what is lurking just behind the scenes, if you care to look."
>
> *Ian Stewart*[2]

The following warning from [**?**, p. vii] should be well heeded:



> "In elementary courses one frequently gets the impression that [differential] equations can usually be 'solved,' i.e., that explicit formulas for solutions (in terms of powers, exponentials, trigonometric functions, and the like) can usually be found. Nothing could be further from the truth. In fact, only very exceptional equations can be explicitly integrated – those that appear in the exercise sections of classical textbooks."

The situation is similar to the fact that it is rarely possible to find an explicit formula for a function implicitly defined by an equation $F(x, y) = 0$, even for nice F ((); in fact, such an equation can be thought of as a "differential equation of order zero").

[2] *The Story of Mathematics*, p. 136.

XI.1. Existence and Uniqueness of Solutions to Ordinary Differential Equations

XI.1.1. Elementary Considerations and Terminology

XI.1.1.1. It is customary (although by no means universal) in the theory of ODEs to denote the independent variable by t, since in many applications the independent variable represents time. The unknown function is typically denoted x or y; we will standardize on using y as being less confusing to those of us used to using x as an independent variable. While t is always assumed to be a real variable, usually ranging over an interval, y can be real-valued or vector-valued (we will use x's to denote elements of the range space and y's to denote functions from $\mathbb{R}$ to the range space).

XI.1.1.2. An ODE is thus an equation involving $t, y, y', \ldots, y^{(n)}$, which can be put in the form

$$F(t, y, y', \ldots, y^{(n)}) = 0$$

by putting everything on one side of the equation. Such an equation is called an *ordinary differential equation of order n* (most differential equations arising in physics and other applications have order 1 or 2). For the most part, we consider only differential equations of the *standard form* (or *Cauchy form*)

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)})$$

which is an abbreviation for

$$y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t))$$

where f is a specified real-valued (or, more generally, vector-valued) function with reasonable properties (usually at least continuous) and the equation is to hold for all t in a suitable subset of $\mathbb{R}$. Most reasonable ODEs can in principle be put into this form using the Implicit Function Theorem (Exercise XI.1.10.2.), and most ODE's arising in applications are already of this form or can easily be put into it explicitly (with the exception of certain linear ODEs; cf. ()).

If f is a function of only $y, y', \ldots, y^{(n-1)}$ (i.e. not explicitly of t), the differential equation is called *autonomous*; when t is interpreted as time, an autonomous differential equation is "time-independent." Many, but by no means all, differential equations in applications are autonomous.

A solution y to an n'th order ODE is necessarily a function which is n-times differentiable; for an equation in standard form, if f is continuous, y must be $\mathcal{C}^n$ since the right side will be a continuous function of t. If f is $\mathcal{C}^r$, then $y^{(n)}$ is $\mathcal{C}^r$ by the Chain Rule, so y is $\mathcal{C}^{n+r}$.

XI.1.1.3. We almost always want to consider only solutions to differential equations on intervals, generally open intervals; sometimes solutions given by formulas have domains consisting of a disjoint union of open intervals, in which case the restriction to each of the subintervals is considered a separate solution. Most commonly we want a solution $y(t)$ on an interval I taking a specified value x_0 at a specified number $t_0 \in I$ (our use of x_0 for the initial value instead of y_0 is nonstandard, but the reasons for using this notation will be apparent later, and it is consistent with our convention described in XI.1.1.1.). A differential equation with a specified value at an "initial" point t_0 is called an *initial-value problem* or *IVP* (for a higher-order differential equation, initial values for lower-order derivatives are also needed, or values at more than one point, called a *boundary-value problem* or *BVP*). While differential equations typically have many solutions, with various intervals of definition, in reasonable cases initial-value problems have unique solutions.

XI.1.1.4. In fact, the general principle of the theory of differential equations is that, subject to some mild technical restrictions, the general solution to an n'th order differential equation is a family of functions determined by n more or less independent constants (parameters), which can be expressed in various ways. (Conversely, a family of smooth functions determined by n independent parameters is the general solution set to an n'th order differential equation, again subject to mild technical restrictions, cf. Exercise XI.1.10.5..)

However, it is usually impossible to find an explicit or closed formula for the general solution to even simple differential equations (cf. XI.1.1.5.(i), XII.12.3.13., XII.12.3.18.). Implicit relations can be found in somewhat greater generality, but even these can be difficult or impossible to come by in many cases. Power-series solutions can often be found if the differential equation is sufficiently regular. But in many cases the best one can do is generate approximate numerical solutions, and there are now powerful and efficient techniques for doing this.

XI.1.1.5. Examples. Here are some simple examples of first- and second-order differential equations that illustrate the general nature of the subject.

(i) Let f be a real-valued continuous function on an interval I. A solution to the first-order differential equation

$$y' = f(t)$$

is simply an antiderivative for f on I. There is thus a solution by V.8.5.9., in fact infinitely many, all differing by constants. If $t_0 \in I$ is fixed and x_0 is a fixed real number, then there is a unique solution y on I satisfying the differential equation (i.e. an antiderivative for f on I) and also the initial condition $y(t_0) = x_0$.

(ii) Consider the autonomous first-order differential equation

$$y' = ky$$

on $\mathbb{R}$, where $k \in \mathbb{R}$ is a constant. A solution is $y(t) = Ae^{kt}$, where A is an arbitrary real constant. To see that these are all the solutions, suppose y is a solution which is not identically 0, and fix t_0 with $y(t_0) \neq 0$. Then y is nonzero in an interval around t_0, and we have, for t in this interval,

$$\frac{d}{dt}\left(\log|y(t)|\right) = \frac{y'(t)}{y(t)} = k$$

so $\log|y(t)| = kt + C$ for some constant C. Thus $|y(t)| = e^C e^{kt}$ for some C, at least for t near enough to t_0. But it follows easily that $y(t) = \pm e^C e^{kt}$ for t near t_0 (with constant sign depending on the sign of $y(t_0)$), and this solution extends uniquely to a solution on all of $\mathbb{R}$. If we specify the initial condition $y(t_0) = x_0$, we have a unique solution with $A = x_0 e^{-kt_0}$, valid on all of $\mathbb{R}$. The "time-independence" of the equation is that if $y = f(t)$ is a solution, then so is $y = f(t + b)$ for any constant b.

(iii) Let $k > 0$ be a constant, and consider the autonomous second-order differential equation

$$y'' = k^2 y\ .$$

The function

$$y(t) = Ae^{kt} + Be^{-kt}$$

is a solution on all of $\mathbb{R}$ for arbitrary real constants A and B. The differential equation

$$y'' = -k^2 y$$

has as solution

$$y(t) = A\cos kt + B\sin kt$$

for constants A, B. It can be shown that in both cases, every solution is of this form (). Simple calculations show that if we fix $t_0 \in \mathbb{R}$ and real constants x_0 and x_0', then in either case there is a unique solution (i.e. unique choice of A and B) with $y(t_0) = x_0$ and $y'(t_0) = x_0'$.

Here are some more examples which illustrate other characteristics of solutions to differential equations:

XI.1.1.6. EXAMPLES. (i) Consider the autonomous first-order differential equation

$$y' = |y| \ .$$

The constant function $f_0(t) = 0$ is a solution, as is $f_A(t) = Ae^t$ for $A > 0$ and $f_A(t) = Ae^{-t}$ for $A < 0$. It can be shown (Exercise ()) that every solution on an interval I is the restriction of one of these solutions to I. These solutions are $\mathcal{C}^\infty$ on all of $\mathbb{R}$ despite the fact that $|y|$ is not a differentiable function of y. It is easily checked that for every t_0 and x_0 there is a unique choice of $A \in \mathbb{R}$ for which f_A is a solution to the IVP

$$y' = |y|, \quad y(t_0) = x_0 \ .$$

The function $f(t, y) = |y|$ is continuous on $\mathbb{R}^2$; although it is not differentiable in y for $y = 0$, it satisfies a uniform Lipschitz condition in y: $|f(t, y_2) - f(t, y_1)| \leq |y_2 - y_1|$ for all $t, y_1, y_2 \in \mathbb{R}$.

(ii) Consider the autonomous first-order differential equation

$$y' = y^2 \ .$$

The constant function 0 is a solution, as is

$$f_c(t) = \frac{1}{c - t}$$

for $c \in \mathbb{R}$. It can be shown (Exercise ()) that every solution on an interval I is the restriction of one of these solutions to I. For any t_0 and x_0, the initial-value problem

$$y' = y^2, \quad y(t_0) = x_0$$

has a unique solution in an interval around t_0: if $x_0 = 0$, the solution is the constant function 0, and if $x_0 \neq 0$, the solution is the restriction of f_c, where $c = t_0 + \frac{1}{x_0}$. This solution is only defined on the interval $(-\infty, c)$ or $(c, +\infty)$ around t_0, not on all of $\mathbb{R}$. The (maximum) domain of the solution depends on t_0 and x_0.

(iii) Consider the autonomous first-order differential equation

$$y' = \frac{3}{2} y^{1/3}$$

(the constant $\frac{3}{2}$ just makes subsequent calculations easy). The constant function 0 is a solution, as are

$$\phi_c^+(t) = \begin{cases} 0 & \text{if} \quad t < c \\ (t - c)^{3/2} & \text{if} \quad t \geq c \end{cases}$$

$$\phi_c^-(t) = \begin{cases} 0 & \text{if} \quad t < c \\ -(t-c)^{3/2} & \text{if} \quad t \geq c \end{cases}$$

for any $c \in \mathbb{R}$. Note that these solutions are defined and $\mathcal{C}^1$ (but not $\mathcal{C}^2$) on all of $\mathbb{R}$. It can be shown (Exercise ()) that every solution on an interval I is the restriction of one of these solutions to I. Note that $f(t,y) = \frac{3}{2}y^{1/3}$ is continuous on $\mathbb{R}^2$, but does not satisfy a Lipschitz condition in y around $y = 0$.

If $t_0, x_0 \in \mathbb{R}$, the initial-value problem

$$y' = \frac{3}{2}y^{1/3}, \quad y(t_0) = x_0$$

always has a solution. If $x_0 \neq 0$, the unique solution is $\phi_c^{\pm}$, where $c = t_0 - x_0^{2/3}$ and the sign is the sign of x_0. If $x_0 = 0$, there are many solutions: the constant function 0 is a solution, as are ϕ_c^+ and ϕ_c^- for any $c \geq t_0$. The solutions 0, $\phi_{t_0}^+$, and $\phi_{t_0}^-$ do not agree on any interval around t_0, so the solutions are not locally unique.

(iv) Consider the autonomous first-order differential equation

$$y' = 3y^{2/3} \ .$$

The constant function 0 is a solution, as are

$$\phi_c(t) = (t-c)^3$$

$$\psi_c^+(t) = \begin{cases} 0 & \text{if} \quad t < c \\ (t-c)^3 & \text{if} \quad t \geq c \end{cases}$$

$$\psi_c^-(t) = \begin{cases} (t-c)^3 & \text{if} \quad t < c \\ 0 & \text{if} \quad t \geq c \end{cases}$$

for any $c \in \mathbb{R}$, and

$$\psi_{c,d}(t) = \begin{cases} (t-c)^3 & \text{if} \quad t < c \\ 0 & \text{if} \quad c \leq t \leq d \\ (t-d)^3 & \text{if} \quad t > d \end{cases}$$

for any $c, d \in \mathbb{R}$, $c < d$ (we could write $\psi_{c,c} = \phi_c$). Note that these solutions are defined on all of $\mathbb{R}$; ϕ_c is $\mathcal{C}^\infty$ (even analytic) on $\mathbb{R}$, and $\psi_c^{\pm}$ and $\psi_{c,d}$ ($c < d$) are $\mathcal{C}^2$ but not $\mathcal{C}^3$ on $\mathbb{R}$. It can be shown (Exercise ()) that every solution on an interval I is the restriction of one of these solutions to I. Note that $f(t,y) = 3y^{2/3}$ is continuous on $\mathbb{R}^2$, but does not satisfy a Lipschitz condition in y around $y = 0$.

If $t_0, x_0 \in \mathbb{R}$, the initial-value problem

$$y' = 3y^{2/3}, \quad y(t_0) = x_0$$

always has a solution. If $x_0 \neq 0$, there is locally a unique solution ϕ_c, where $c = t_0 - x_0^{1/3}$. However, this solution is not globally unique: $\psi_c^{\pm}$ (where the sign is the sign of x_0) is also a solution, as is $\psi_{c,d}$ for any $d > c$ if $x_0 < 0$ and $\psi_{d,c}$ for any $d < c$ if $x_0 > 0$. If $x_0 = 0$, there are many solutions: the constant function 0 is a solution, as are ϕ_{t_0}, ψ_c^+ for any $c \geq t_0$, ψ_c^- for any $c \leq t_0$, and $\psi_{c,d}$ for $c \leq t_0 \leq d$. The solutions 0, ϕ_{t_0}, $\psi_{t_0}^+$, and $\psi_{t_0}^-$ do not agree on any interval around t_0, so the solutions are not locally unique.

XI.1.2. Existence and Uniqueness Theorem for First-Order ODEs

In this subsection, we prove a general theorem about existence and uniqueness of local solutions to a first-order initial-value problem

$$y' = f(t, y), \quad y(t_0) = x_0$$

where y is an unknown function from an interval I in $\mathbb{R}$ to a (real) Banach space $\mathcal{X}$. In most elementary applications $\mathcal{X}$ will be $\mathbb{R}$, or $\mathbb{R}^n$ for some n; we use a general Banach space since the result is then cleaner to state as well as more general, and the proof is the same. This fundamental theorem is closely related to the Implicit and Inverse Function Theorems, at least in spirit (cf. XI.1.10.12.). In particular, it is a local result, but we have control over the size of the interval on which the solution is valid, which in some key special cases gives global solutions. This result was first obtained by E. PICARD in 1890; our exposition is adapted from [**?**]. The result is called the *Picard Existence Theorem*, *Picard-Lindelöf Theorem*, or *Cauchy-Lipschitz Theorem* in various references (the Cauchy-Lipschitz Theorem is properly the crucial special case XI.1.2.7., obtained before PICARD'S general result).

XI.1.2.1. The setup of this theorem is as follows. Let $\mathcal{X}$ be a Banach space and $x_0 \in \mathcal{X}$. Let $\bar{B}_r(x_0)$ be the closed ball in $\mathcal{X}$ centered at x_0 with radius $r > 0$. Let $t_0 \in \mathbb{R}$, and I an interval in $\mathbb{R}$ with $t_0 \in I$. Let f be a continuous function from $I \times \bar{B}_r(x_0)$ to $\mathcal{X}$. We suppose in addition there is a constant M such that $\|f(t,x)\| \leq M$ for all $(t,x) \in I \times \bar{B}_r(x_0)$, and that f satisfies a uniform Lipschitz condition in x (or, as we usually write, in y) on $I \times \bar{B}_r(x_0)$, i.e. there is a constant K such that for every $t \in I$ and $x_1, x_2 \in \bar{B}_r(x_0)$ we have

$$\|f(t,x_2) - f(t,x_1)\| \leq K\|x_2 - x_1\| \ .$$

We seek an interval $I_0 \subseteq I$, $t_0 \in I_0$, and a differentiable function $y : I_0 \to \mathcal{X}$ such that $y'(t) = f(t, y(t))$ for $t \in I_0$ and $y(t_0) = x_0$, and to show that any two such functions agree on I_0.

XI.1.2.2. THEOREM. [EXISTENCE AND UNIQUENESS THEOREM FOR FIRST-ORDER ODES] Using the above notation, set

$$I_0 = I \cap \left[t_0 - \frac{r}{M}, t_0 + \frac{r}{M}\right] \ .$$

Then there is a unique differentiable function $y : I_0 \to \mathcal{X}$ such that $y'(t) = f(t, y(t))$ for $t \in I_0$ and $y(t_0) = x_0$. Additionally, y is $\mathcal{C}^1$ on I_0. If $\tilde{I}$ is any subinterval of I_0 containing t_0, the restriction of y to $\tilde{I}$ is the unique solution to the IVP on $\tilde{I}$.

PROOF: It suffices to show there is a unique solution on any closed subinterval J of I_0 containing t_0. The strategy of the proof is to use the (Extended) Banach Fixed-Point Theorem (VI.2.10.10.) to obtain the solution on J as the unique fixed point of a suitable strict power contraction on a complete metric space. To this end, set $\epsilon = \frac{r}{M}$ and let Y be the set of continuous functions from J to the closed ball $\bar{B}_r(x_0)$ in $\mathcal{X}$ of radius r centered at x_0. Define a metric ρ on Y by

$$\rho(\phi, \psi) = \max_{t \in J} \|\phi(t) - \psi(t)\| \ .$$

Then Y is a complete metric space under ρ (it is a closed subset of $C(J, \mathcal{X})$ ()). Define a map $T : Y \to Y$ by

$$[T(\phi)](t) = x_0 + \int_{t_0}^{t} f(u, \phi(u))\, du \ .$$

It is obvious from () that $T(\phi)$ is a continuous function from J to $\mathcal{X}$. We must show that $T(\phi) \in Y$ if $\phi \in Y$, and that there is an n and a constant $\alpha < 1$ such that $\rho(T^n(\phi), T^n(\psi)) \le \alpha\rho(\phi, \psi)$ for all $\phi, \psi \in Y$.

We have, for $t \in J$ and $\phi \in Y$,

$$\|[T(\phi)](t) - x_0\| = \left\| \int_{t_0}^{t} f(u, \phi(u))\, du \right\| \le \int_{t_0}^{t} \|f(u, \phi(u)\|\, du \le M|t - t_0| \le M\epsilon = r$$

so $T(\phi) \in Y$. And we have, for $\phi, \psi \in Y$ and $t \in J$,

$$\|T(\phi)(t) - T(\psi)(t)\| = \left\| \int_{t_0}^{t} [f(u, \phi(u)) - f(u, \psi(u))]\, du \right\|$$

$$\le \int_{t_0}^{t} \|f(u, \phi(u)) - f(u, \psi(u))\|\, du \le K \int_{t_0}^{t} \|\phi(u) - \psi(u)\|\, du$$

$$\le K\epsilon \max_{u \in J} \|\phi(u) - \psi(u)\| = K\epsilon\rho(\phi, \psi)\ .$$

Similarly, for each n we have

$$\|T^n(\phi) - T^n(\psi)\| \le K \left| \int_{t_0}^{t} \|T^{n-1}(\phi)(u_1) - T^{n-1}(\psi)(u_1)\|\, du_1 \right|$$

$$\le K^2 \left| \int_{t_0}^{t} \int_{t_0}^{u_1} \|T^{n-2}(\phi)(u_2) - T^{n-2}(\psi)(u_2)\|\, du_2\, du_1 \right|$$

$$\le \cdots \le K^n \left| \int_{t_0}^{t} \int_{t_0}^{u_1} \cdots \int_{t_0}^{u_{n-1}} \|\phi(u_n) - \psi(u_n)\|\, du_n \cdots du_1 \right|$$

$$\le K^n \rho(\phi, \psi) \left| \int_{t_0}^{t} \int_{t_0}^{u_1} \cdots \int_{t_0}^{u_{n-1}} du_n \cdots du_1 \right| = \frac{K^n |t - t_0|^n}{n!} \rho(\phi, \psi) \le \frac{(K\epsilon)^n}{n!} \rho(\phi, \psi)\ .$$

Since t is arbitrary, we have

$$\rho(T^n(\phi), T^n(\psi)) \le \frac{(K\epsilon)^n}{n!} \rho(\phi, \psi)\ .$$

If n is sufficiently large, then

$$\frac{(K\epsilon)^n}{n!} < 1$$

and thus T^n is a strict contraction from Y to Y.

Thus, by VI.2.10.10., T has a unique fixed point y. We thus have

$$y(t) = t_0 + \int_{t_0}^{t} f(u, y(u))\, du$$

for $t \in J$, and since $g(u) = f(u, y(u))$ is continuous, by () we have that $y'(t) = f(t, y(t))$ for $t \in J$. Also obviously $y(t_0) = x_0$. Conversely, any solution on J must be a fixed point of T, and thus y is the unique solution on J.

If J' is a closed subinterval of I_0 containing t_0 different from J, say $J \subseteq J'$, the unique solution on J' restricted to J must be the unique solution on J.

Since $g(t) = f(t, y(t))$ is continuous, y' is continuous, i.e. y is $\mathcal{C}^1$. 👍

XI.1.2.3. The restriction of the domain of the solution to $[t_0 - \frac{r}{M}, t_0 + \frac{r}{M}]$ simply insures that the unique solution near t_0, when continued to the endpoints of the interval, does not leave the range of y-values for which f is defined (the restriction is a real one even if r can be chosen arbitrarily large; cf. Example XI.1.1.6.(ii)); the constant $\frac{r}{M}$ is conservative in general (cf. Exercise XI.1.10.6.). Note that the size of the Lipschitz constant K does not affect the length of the interval on which the existence of a unique solution is guaranteed by the theorem. Note also that if t_0 is in the interior of I (the usual situation in which the theorem is applied), it is also in the interior of I_0.

XI.1.2.4. The proof of Theorem XI.1.2.2. is constructive: it not only shows the existence of a unique solution, but actually gives an iterative algorithm for finding it. In fact, if ϕ is any continuous function from I_0 to $\mathcal{X}$ with $\phi(t_0) = x_0$ (e.g. the constant function x_0), then the sequence $(T^n(\phi))$ converges uniformly on I_0 to the solution. This "method of successive approximations," called *Picard iteration*, is a practical computational method for finding approximations to the solution in many cases, although there are usually better numerical techniques.

XI.1.2.5. In fact, we have the following error estimate for the successive approximation: if ϕ_0 is the constant function x_0 and $\phi_n = T^n(\phi_0)$ for $n \in \mathbb{N}$, then, for $t \in I_0$,

$$\|\phi_0(t) - \phi_1(t)\| = \left\| \int_{t_0}^{t} f(u, y_0)\, du \right\| \leq \left| \int_{t_0}^{t} \|f(t, y_0)\|\, du \right| \leq M|t - t_0|$$

and by induction

$$\|\phi_n(t) - \phi_{n+1}(t)\| \leq \frac{MK^n}{(n+1)!}|t - t_0|^{n+1} \leq \frac{M}{K}\frac{(Kr/M)^{n+1}}{(n+1)!}$$

$$\|\phi_n(t) - y(t)\| \leq \frac{M}{K}\sum_{k=n+1}^{\infty}\frac{(Kr/M)^k}{k!}$$

The sum is the error in the n'th Taylor approximation to $e^{Kr/M}$, hence by ()

$$\rho(\phi_n, y) \leq \frac{Me^{Kr/M}}{K(n+1)!}\ .$$

Thus the ϕ_n converge uniformly quite rapidly to y.

We then get a local version with a cleaner statement but less precise conclusion:

XI.1.2.6. Corollary. [Existence and Uniqueness Theorem for First-Order ODEs, Local Version] Let $\mathcal{X}$ be a Banach space, $x_0 \in \mathcal{X}$, $t_0 \in \mathbb{R}$, and $f(t, y)$ a continuous function from a neighborhood U of (t_0, x_0) in $\mathbb{R} \times \mathcal{X}$ to $\mathcal{X}$. Suppose f satisfies a uniform Lipschitz condition in y on U. Then there is a $\delta > 0$ such that the first-order initial value problem

$$y' = f(t, y), \quad y(t_0) = x_0$$

has a unique solution on $[t_0 - \delta, t_0 + \delta]$. Additionally, y is $\mathcal{C}^1$ on $[t_0 - \delta, t_0 + \delta]$.

PROOF: An r and a closed bounded interval I containing t_0 in its interior can be chosen such that $I \times \bar{B}_r(x_0)$ is contained in U. By compactness, $f(t, x_0)$ is bounded, say by N, on I. If K is the Lipschitz constant, then for $y \in \bar{B}_r(x_0)$ we have

$$\|f(t,y)\| \leq \|f(t,x_0)\| + \|f(t,y) - f(t,x_0)\| \leq N + Kr$$

so f is bounded on $I \times \bar{B}_r(y_0)$. Apply the theorem with this I and r. 👍

As a special case, we get the following version which covers most standard applications. This is the version commonly stated and (sometimes) proved in Differential Equations texts.

XI.1.2.7. COROLLARY. [EXISTENCE AND UNIQUENESS THEOREM FOR FIRST-ORDER ODES, STANDARD VERSION] Let R be a rectangle in $\mathbb{R}^2$, and (t_0, x_0) an interior point of R. Let $f(t, y)$ be a function from R to $\mathbb{R}$ with f and $\frac{\partial f}{\partial y}$ continuous on R. Then there is a $\delta > 0$ such that the initial-value problem

$$y' = f(t,y), \quad y(t_0) = x_0$$

has a unique solution on $(t_0 - \delta, t_0 + \delta)$. Additionally, y is $\mathcal{C}^1$ on $(t_0 - \delta, t_0 + \delta)$.

PROOF: Since $\frac{\partial f}{\partial y}$ is continuous, it is bounded in a neighborhood of (t_0, x_0), and thus by the MVT (V.8.2.11.) f satisfies a uniform Lipschitz condition in y on this neighborhood. 👍

Using the theorem, one can get an explicit estimate for the size δ can be. Actually R can be any neighborhood of (t_0, x_0).

Global Solutions

XI.1.2.8. The above results only give a solution in a (usually small) interval around t_0. Even if f is defined on all of $I \times \mathcal{X}$, i.e. the hypotheses of XI.1.2.2. are satisfied for all r (with the Lipschitz constant K depending on r), Example XI.1.1.6.(ii) shows that the solution may not be defined on all of I. The maximum interval on which the solution is defined can even depend on the choice of x_0 (this might be predicted since the constant M can depend on x_0 as well as on r).

But if f satisfies a uniform Lipschitz condition in y on all of $I \times \mathcal{X}$ (or even somewhat more generally), the control on the domain in XI.1.2.2. allows us to cobble together local solutions into a solution defined on all of I:

XI.1.2.9. THEOREM. [EXISTENCE AND UNIQUENESS THEOREM FOR FIRST-ORDER ODES, GLOBAL VERSION] Let $\mathcal{X}$ be a Banach space, I an interval in $\mathbb{R}$, and $t_0 \in I$. Let $f(t, y)$ be a continuous function from $I \times \mathcal{X}$ to $\mathcal{X}$ satisfying a uniform Lipschitz condition in y on every closed bounded subinterval of I. Then, for any $x_0 \in \mathcal{X}$, the initial-value problem

$$y' = f(t,y), \quad y(t_0) = x_0$$

has a unique solution defined on all of I. Additionally, the solution is $\mathcal{C}^1$ on I.

PROOF: It suffices to show that if $J = [a, b]$ is any closed bounded subinterval of I containing t_0 in its interior, then there is a unique solution to the IVP $y' = f(t, y)$, $y(t_0) = x_0$ on J. Fix such a J. If K is the Lipschitz constant on J, also fix L with $0 < L < \frac{1}{K}$.

First let $t_1 \in J$ and $x_1 \in \mathcal{X}$, and consider the IVP $y' = f(t, y)$, $y(t_1) = x_1$. The function $g(t) = f(t, x_1)$ is bounded on J by compactness, say by N. Then for any r, f is bounded on $J \times \bar{B}_r(x_1)$ by $N + Kr$ (cf. the proof of XI.1.2.6.). Thus one bound on the length of the interval on which a unique solution is guaranteed by Theorem XI.1.2.2. is $\frac{r}{N+Kr}$. If r is taken large enough, this is greater than L. Thus Theorem XI.1.2.2. gives a unique solution to this IVP on $J \cap [t_1 - L, t_1 + L]$.

Now return to the original IVP $y' = f(t, y)$, $y(t_0) = x_0$. Theorem XI.1.2.2. gives a unique solution $y = \phi_0(t)$ on $J_0 = J \cap [t_0 - L, t_0 + L]$. We now successively extend this unique solution to the right until it extends to b. If $t_0 + L \geq b$, we are done. Otherwise, fix t_1 with $t_0 + \frac{L}{2} < t_1 < t_0 + L$. Set $x_1 = \phi_0(t_1)$. Theorem XI.1.2.2. gives a unique solution $y = \phi_1(t)$ to the IVP $y' = f(t, y)$, $y(t_1) = x_1$ on $J_1 = J \cap [t_1 - L, t_1 + L]$. We must have $\phi_1 = \phi_0$ on $J_0 \cap J_1$ since both are solutions to the IVP $y' = f(t, y)$, $y(t_1) = x_1$ on $J_0 \cap J_1$, and thus we have extended ϕ_0 to a solution on $J_0 \cup J_1$.

If $t_1 + L \geq b$, we are done extending to the right. If not, we have $t_1 + L \geq (t_0 + L) + \frac{L}{2}$, i.e. we have extended to the right by at least $\frac{L}{2}$. Repeat the process, choosing t_2 with $t_1 + \frac{L}{2} < t_2 < t_1 + L$, x_2, J_2, ϕ_2, etc. Since at each stage we extend to the right by at least $\frac{L}{2}$, we reach b in a finite number of steps.

Extend ϕ_0 to the left of t_0 to a in the same manner. We obtain a solution ϕ on J in this manner. Any solution to the original IVP must agree with ϕ on a subset of J which is both relatively open and closed, hence must agree with ϕ everywhere on J. 👍

XI.1.3. Systems of First-Order Differential Equations

The result XI.1.2.2. is general enough to immediately give solutions not only to a single differential equation, but also to systems of differential equations: we want differentiable functions $y_1, \ldots, y_n$ on an interval I satisfying the n equations

$$y_k' = f_k(t, y_1, \ldots, y_n) \quad (1 \leq k \leq n) \ .$$

We also want the y_k to have specified values at some $t_0 \in I$.

XI.1.3.1. The setup is similar to the one for single equations, but we change notation slightly. We have a Banach space $\mathcal{X}$, an interval $I \subseteq \mathbb{R}$, a point $t_0 \in I$, n points $x_1^{(0)}, \ldots, x_n^{(0)}$ in $\mathcal{X}$, $r > 0$, $M > 0$, $K > 0$, and n continuous functions

$$f_1, \ldots, f_n : I \times \bar{B}_r(x_1^{(0)}) \times \cdots \times \bar{B}_r(x_n^{(0)}) \to \mathcal{X}$$

with the property that $\|f_k(t, x_1, \ldots, x_n)\| \leq M$ for $t \in I$ and $\|x_j - x_j^{(0)}\| \leq r$ $(1 \leq j, k \leq n)$, and

$$\|f_k(t, x_1^{(2)}, \ldots, x_n^{(2)}) - f_k(t, x_1^{(1)}, \ldots, x_n^{(1)})\| \leq K \max_{1 \leq j \leq n} (\|x_j^{(2)} - x_j^{(1)}\|)$$

for $1 \leq k \leq n$ whenever $t \in I$ and $x_j^{(1)}, x_j^{(2)} \in \bar{B}_r(x_j^{(0)})$ for all j (the f_k satisfy a uniform Lipschitz condition in $x_1, \ldots, x_n$ on I).

XI.1.3.2. Theorem. [Existence and Uniqueness Theorem for Systems of First-Order ODEs] Using the above notation, set

$$I_0 = I \cap \left[t_0 - \frac{r}{M}, t_0 + \frac{r}{M}\right] .$$

Then there are unique differentiable functions $y_1, \ldots, y_n : I_0 \to \mathcal{X}$ such that

$$y_k'(t) = f_k(t, y_1(t), \ldots, y_n(t))$$

$(1 \leq k \leq n)$ for $t \in I_0$ and $y_k(t_0) = x_k^{(0)}$. Additionally, the y_k are $\mathcal{C}^1$ on I_0. If $\tilde{I}$ is any subinterval of I_0 containing t_0, the restriction of the y_k to $\tilde{I}$ is the unique solution to the IVP on $\tilde{I}$.

Proof: This is an almost immediate corollary of XI.1.2.2.. Set $\tilde{\mathcal{X}} = \mathcal{X}^n$ with the ∞-norm (), and $x_0 = (x_1^{(0)}, \ldots, x_n^{(0)})$. Define $f : I \times \bar{B}_r(x_0) \to \tilde{\mathcal{X}}$ by

$$f(t, (x_1, \ldots, x_n)) = (f_1(t, x_1, \ldots, x_n), \ldots, f_n(t, x_1, \ldots, x_n)) .$$

Then the hypotheses of XI.1.2.2. are satisfied, and the coordinate functions of the solution y give the y_k. 👍

The interval on which the existence of a unique solution is guaranteed can sometimes be shown to be larger than I_0; see Exercise XI.1.10.7..

As in the single equation case, we get a local version with a cleaner statement but less precise conclusion:

XI.1.3.3. Corollary. [Existence and Uniqueness Theorem for Systems of First-Order ODEs, Local Version] Let $\mathcal{X}$ be a Banach space, $x_1^{(0)}, \ldots, x_n^{(0)} \in \mathcal{X}$, $t_0 \in \mathbb{R}$, and $f_1, \ldots, f_n$ continuous functions from a neighborhood U of $(t_0, x_1^{(0)}, \ldots, x_n^{(0)})$ in $\mathbb{R} \times \mathcal{X}^n$ to $\mathcal{X}$. Suppose each f_k satisfies a uniform Lipschitz condition in $x_1, \ldots, x_n$ on U. Then there is a $\delta > 0$ such that the first-order initial-value problem

$$y_k' = f_k(t, y_1, \ldots, y_n), \quad y_k(t_0) = x_k^{(0)} \quad (1 \leq k \leq n)$$

has a unique solution on $(t_0 - \delta, t_0 + \delta)$. Additionally, the y_k are $\mathcal{C}^1$ on $(t_0 - \delta, t_0 + \delta)$.

We also get a global version from XI.1.2.9.:

XI.1.3.4. Theorem. [Existence and Uniqueness Theorem for Systems of First-Order ODEs, Global Version] Let $\mathcal{X}$ be a Banach space, I an interval in $\mathbb{R}$, and $t_0 \in I$. Let $f_1, \ldots, f_n$ be continuous functions from $I \times \mathcal{X}^n$ to $\mathcal{X}$ satisfying a uniform Lipschitz condition in $x_1, \ldots, x_n$ on every closed bounded subinterval of I. Then, for any $x_1^{(0)}, \ldots, x_n^{(0)} \in \mathcal{X}$, the initial-value problem

$$y_k' = f_k(t, y_1, \ldots, y_n), \quad y_k(t_0) = x_k^{(0)} \quad (1 \leq k \leq n)$$

has a unique solution defined on all of I. Additionally, the solutions are $\mathcal{C}^1$ on I.

XI.1.4. Higher-Order Differential Equations and Systems

We can use the results for systems of first-order equations to obtain similar existence and uniqueness theorems for higher-order differential equations, by a simple but clever trick. We want to solve the initial-value problem

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y(t_0) = x_0, y'(t_0) = x_0', \ldots, y^{(n-1)}(t_0) = x_0^{(n-1)}$$

where y is an n-times differentiable function from an interval in $\mathbb{R}$ to a Banach space $\mathcal{X}$ and f is a suitably nice function from a neighborhood of $(t_0, x_0, x_0', \ldots, x_0^{(n-1)})$ in $\mathbb{R} \times \mathcal{X}^n$ to $\mathcal{X}$.

The trick is to define new unknown functions $y_1, \ldots, y_{n-1}$ which satisfy the system of equations

$$y' = y_1, y_1' = y_2, \ldots, y_{n-2}' = y_{n-1}, y_{n-1}' = f(t, y, y_1, \ldots, y_{n-1})$$

so $y_k = y^{(k)}$. Then we have a system of n first-order equations in n unknown functions $y, y_1, \ldots, y_{n-1}$, and we can apply the various existence and uniqueness theorems for such systems. The statements are:

XI.1.4.1. Theorem. [Existence and Uniqueness Theorem for Higher-Order ODE's] Let $\mathcal{X}$ be a Banach space, I an interval in $\mathbb{R}$, $t_0 \in I$, $x_0, x_0', \ldots, x_0^{(n-1)} \in \mathcal{X}$, $r > 0$, and

$$f : I \times \bar{B}_r(x_0) \times \bar{B}_r(x_0') \times \cdots \times \bar{B}_r(x_0^{(n-1)}) \to \mathcal{X} \ .$$

Suppose f is continuous and bounded by N on $I \times \bar{B}_r(x_0) \times \bar{B}_r(x_0') \times \cdots \times \bar{B}_r(x_0^{(n-1)})$ and satisfies a uniform Lipschitz condition in the y-variables. Set

$$M = \max\{N, \|y_0'\| + r, \ldots, \|y_0^{(n-1)}\| + r\}$$

$$I_0 = I \cap \left[t_0 - \frac{r}{M}, t_0 + \frac{r}{M}\right] \ .$$

Then there is a unique n-times differentiable function $y : I_0 \to \mathcal{X}$ with $y(t_0) = x_0$, $y'(t_0) = x_0'$, $\ldots$, $y^{(n-1)}(t_0) = x_0^{(n-1)}$, and $y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t))$ for all $t \in I_0$. Additionally, y is $\mathcal{C}^n$ on I_0.

Proof: Define $f_k(t, x_0, x_1, \ldots, x_{n-1}) = x_{k+1}$ for $0 \le k < n-1$ and $f_{n-1} = f$, and apply XI.1.3.2.. ☝

The interval on which the existence of a unique solution is guaranteed can sometimes be shown to be larger than I_0; see Exercise XI.1.10.8..

XI.1.4.2. Corollary. [Existence and Uniqueness Theorem for Higher-Order ODEs, Local Version] Let $\mathcal{X}$ be a Banach space, $x_0, x_0', \ldots, x_0^{(n-1)} \in \mathcal{X}$, $t_0 \in \mathbb{R}$, and f a continuous function from a neighborhood U of $(t_0, x_0, x_0', \ldots, x_0^{(n-1)})$ in $\mathbb{R} \times \mathcal{X}^n$ to $\mathcal{X}$. Suppose f satisfies a uniform Lipschitz condition in the y-variables on U. Then there is a $\delta > 0$ such that the first-order initial-value problem

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y(t_0) = x_0, y'(t_0) = x_0', \ldots, y^{(n-1)}(t_0) = x_0^{(n-1)}$$

has a unique solution on $(t_0 - \delta, t_0 + \delta)$. Additionally, y is $\mathcal{C}^n$ on $(t_0 - \delta, t_0 + \delta)$.

XI.1.4.3. THEOREM. [EXISTENCE AND UNIQUENESS THEOREM FOR HIGHER-ORDER ODES, GLOBAL VERSION] Let $\mathcal{X}$ be a Banach space, I an interval in $\mathbb{R}$, and $t_0 \in I$. Let f be a continuous function from $I \times \mathcal{X}^n$ to $\mathcal{X}$ satisfying a uniform Lipschitz condition in the y-variables on every closed bounded subinterval of I. Then, for any $x_0, x_0', \ldots, x_0^{(n-1)} \in \mathcal{X}$, the initial-value problem

$$y^{(n)} = f(t, y, y', \ldots, y^{(n-1)}), \quad y(t_0) = x_0, y'(t_0) = x_0', \ldots, y^{(n-1)}(t_0) = x_0^{(n-1)}$$

has a unique solution defined on all of I. Additionally, the solution is $\mathcal{C}^n$ on I.

XI.1.4.4. There are similar extensions to systems of higher-order differential equations. The equations in the system do not all need to be of the same order. The extensions are straightforward, but the notation is oppressive, and it is left to the reader to formulate and prove the results.

XI.1.5. Continuous Dependence on Initial Conditions

XI.1.5.1. We now examine what happens when the initial conditions are varied. To do this we will treat the initial value as a parameter. Thus, if $f(t, y)$ is a suitably nice function (e.g. continuous and satisfying a uniform Lipschitz condition in y) defined near (t_0, x_0), the IVP

$$y' = f(t, y), \quad y(t_0) = x_0$$

has a unique solution, and the IVP

$$y' = f(t, y), \quad y(t_0) = x_1$$

even has a unique solution for any x_1 sufficiently close to x_0.

XI.1.5.2. What we would hope is that if x_1 is close to x_0, the solution with initial value x_1 should be uniformly close to the solution for initial value x_0 for t close to t_0, i.e. that the solution for initial value x should "vary continuously" (or even "smoothly") in x.

To be precise, what we might hope for is that there is a continuous (or even smooth) function $u(t, x)$ of two variables, defined near (t_0, x_0), such that if x_1 is close to x_0, the function $y(t) = u(t, x_1)$ is the unique solution to the IVP with initial value x_1. We get continuity:

XI.1.5.3. THEOREM. Let $\mathcal{X}, f, r, K, M$ be as in XI.1.2.2.. Fix ϵ, $0 < \epsilon < \min\left(\frac{r}{2M}, \frac{1}{K}\right)$. Let y_0 be the unique solution to the IVP

$$y' = f(t, y), \quad y(t_0) = x_0$$

in the interval $J = [t_0 - \epsilon, t_0 + \epsilon]$. Fix x_1, $\|x_1 - x_0\| \leq \frac{r}{2}$, and let y_1 be the unique solution to the IVP

$$y' = f(t, y), \quad y(t_0) = x_1$$

on J. Then, for any $t \in J$, we have

$$\|y_1(t) - y_0(t)\| \leq \frac{\|x_1 - x_0\|}{1 - K\epsilon} .$$

XI.1.5.4. Some initial observations:

(i) The ϵ is not the same as the one in the proof of XI.1.2.2. (it is smaller).

(ii) The interval J is shorter than the maximum length interval on which y_0 is guaranteed to exist by XI.1.2.2. and is not the same as the J in that proof, but t_0 is still in the interior of J.

(iii) It is easily checked that XI.1.2.2. guarantees the existence of the solution y_1 as well as y_0 on J.

The reason J is taken shorter in this theorem than in XI.1.2.2. is that we need the T defined in the proof to be an actual strict contraction, not just a strict power contraction, in order to apply VI.2.10.7..

XI.1.5.5. COROLLARY. Under the hypotheses and notation of XI.1.5.3., if $u(t,x)$ is the function from $I \times \bar{B}_{r/2}(x_0)$ to $\mathcal{Y}$ such that, for fixed x, $u(t,x)$ is the unique solution to the initial-value problem

$$y' = f(t,y), \quad y(t_0) = x$$

then u is uniformly continuous on $I \times \bar{B}_{r/2}(x_0)$.

We now give the proof of XI.1.5.3..

PROOF: We proceed exactly as in the proof of XI.1.2.2.. Let Y be the set of continuous functions from J to the closed ball $\bar{B}_r(x_0)$ in $\mathcal{X}$ of radius r centered at x_0. Define a metric ρ on Y by

$$\rho(\phi,\psi) = \max_{t\in J} \|\phi(t) - \psi(t)\| \ .$$

Then Y is a complete metric space under ρ (it is a closed subset of $C(J,\mathcal{X})$ ()). Define maps $T, S : Y \to Y$ by

$$[T(\phi)](t) = x_0 + \int_{t_0}^{t} f(u,\phi(u))\,du$$

$$[S(\phi)](t) = x_1 + \int_{t_0}^{t} f(u,\phi(u))\,du \ .$$

It is obvious from () that $T(\phi)$ and $S(\phi)$ are continuous functions from J to $\mathcal{X}$. We have, for $t \in J$ and $\phi \in Y$,

$$\|[T(\phi)](t) - x_0\| = \left\| \int_{t_0}^{t} f(u,\phi(u))\,du \right\| \le \int_{t_0}^{t} \|f(u,\phi(u)\|\,du \le M|t-t_0| \le M\epsilon = \frac{r}{2}$$

so $T(\phi) \in Y$. Also,

$$\|[S(\phi)](t) - x_0\| \le \|[S(\phi)](t) - x_1\| + \|x_1 - x_0\| = \left\| \int_{t_0}^{t} f(u,\phi(u))\,du \right\| + \|x_1 - x_0\|$$

$$\le \int_{t_0}^{t} \|f(u,\phi(u)\|\,du + \|x_1 - x_0\| \le M|t-t_0| + \|x_1 - x_0\| \le M\epsilon + \frac{r}{2} = \frac{r}{2} + \frac{r}{2} = r$$

so $S(\phi) \in Y$. And we have, for $\phi, \psi \in Y$ and $t \in J$,

$$\|T(\phi)(t) - T(\psi)(t)\| = \left\| \int_{t_0}^{t} [f(u,\phi(u)) - f(u,\psi(u))]\,du \right\|$$

$$\leq \int_{t_0}^{t} \|f(u,\phi(u)) - f(u,\psi(u))\|\, du \leq K \int_{t_0}^{t} \|\phi(u) - \psi(u)\|\, du$$
$$\leq K\epsilon \max_{u\in J} \|\phi(u) - \psi(u)\| = K\epsilon\rho(\phi,\psi)$$

and similarly $\|S(\phi)(t) - S(\psi)(t)\| \leq K\epsilon\rho(\phi,\psi)$. Thus T and S are strict contractions on Y with constant $K\epsilon < 1$. Since $T(\phi) - S(\phi) = x_0 - x_1$ for all ϕ, we have $\|T(\phi) - S(\phi)\| = \|x_1 - x_0\|$ for all ϕ. The fixed points for T and S are y_0 and y_1 respectively. The result now follows from an application of VI.2.10.7.. 👍

XI.1.5.6. The technicalities of the proof only yield that uniform continuity holds on a small neighborhood of each point, but it follows that continuity holds throughout the region in which unique local solutions exist. Uniform continuity can be cobbled together to hold on somewhat larger regions; details are left to the interested reader.

XI.1.5.7. The problem in XI.1.5.2. can be converted to a first-order quasilinear PDE problem (Cauchy problem): we seek a smooth function $u(t,x)$ in a neighborhood of (t_0,x_0) satisfying (writing u_t for $\frac{\partial u}{\partial t}$)

$$u_t = f(t,u)$$

with boundary condition $u(t_0,x) = x$. () then shows the solution $u(t,x)$ of XI.1.5.5. is not only continuous, but even smooth.

XI.1.5.8. We might even want to allow the function f to depend on the x, i.e. to allow the differential equation itself to vary as the initial condition is varied, or in other words to consider the PDE

$$u_t = f(t,x,u)$$

with boundary condition $u(t_0,x) = x$, where f is a suitable function defined in a neighborhood of (t_0,x_0,x_0). Under reasonable hypotheses () shows that there is also a smooth solution in this case.

XI.1.6. Peano's Existence Theorem for First-Order ODEs

If f does not satisfy a uniform Lipschitz condition in y, we cannot expect a locally unique solution to the IVP $y' = f(t,y)$, $y(t_0) = x_0$ in general (Examples XI.1.1.6.(iii)–(iv)). But we can still show existence of a local solution. The first step is to uniformly approximate f by functions which do satisfy a uniform Lipschitz condition in y; this can only done in general if the target Banach space is finite-dimensional. An application of the Arzela-Ascoli Theorem (which also needs finite-dimensionality of the target space) then gives a solution. In fact, the conclusion can fail if the Banach space $\mathcal{X}$ is infinite-dimensional, even if $\mathcal{X}$ is a separable Hilbert space [Yor70]. See [**?**] for other proofs of the theorem and some generalizations.

XI.1.6.1. Theorem. [Peano's Existence Theorem for First-Order ODEs] Let $\mathcal{X}$ be a finite-dimensional Banach space, I a closed bounded interval in $\mathbb{R}$, $t_0 \in I$, $x_0 \in \mathcal{X}$. Let $f : I \times \bar{B}_r(x_0) \to \mathcal{X}$ be a continuous function, and

$$M = \max\{\|f(t,y)\| : (t,y) \in I \times \bar{B}_r(x_0)\}\ .$$

Set

$$I_0 = I \cap \left[t_0 - \frac{r}{M}, t_0 + \frac{r}{M}\right] \ .$$

Then there is a $\mathcal{C}^1$ function $y : I_0 \to \bar{B}_r(y_0) \subseteq \mathcal{X}$ satisfying $y(t_0) = x_0$ and $y'(t) = f(t, y(t))$ for all $t \in I_0$.

PROOF: We may assume $\mathcal{X} = \mathbb{R}^n$ with a norm equivalent to the Euclidean norm. By the Weierstrass Approximation Theorem, we may approximate f uniformly on $I \times \bar{B}_r(y_0)$ by functions f_n whose coordinate functions are polynomials. Thus each f_n satisfies a uniform Lipschitz condition in y on $I \times \bar{B}_r(x_0)$. By scaling f_n slightly, we may also assume $\|f_n(t, y)\| \leq M$ for all n and all $(t, y) \in I \times \bar{B}_r(x_0)$.

Thus by Theorem XI.1.2.2. there is a (unique) solution ϕ_n to the IVP $y' = f_n(t, y)$, $y(t_0) = x_0$ on the closed bounded interval I_0. We have, for all n and all $t \in I_0$,

$$\phi_n(t) = x_0 + \int_{t_0}^{t} f_n(u, \phi_n(u))\, du$$

and thus

$$\|\phi_n(t)\| \leq \|x_0\| + \left| \int_{t_0}^{t} \|f_n(u, \phi_n(u))\|\, du \right| \leq \|x_0\| + M|t - t_0| \leq \|x_0\| + r$$

so the ϕ_n are uniformly bounded on I_0. Also, $\|\phi_n'(t)\| = \|f_n(t, \phi_n(t))\| \leq M$ for all $t \in I_0$, so the set $\{\phi_n : n \in \mathbb{N}\}$ is equicontinuous on I_0 (). Thus there is a uniformly convergent subsequence by the Arzela-Ascoli Theorem (). Passing to a subsequence, we may assume (ϕ_n) converges uniformly on I_0 to a (necessarily continuous) function ϕ. Then, for each $t \in \mathbb{R}$,

$$x_0 + \int_{t_0}^{t} f_n(u, \phi_n(u))\, du \to x_0 + \int_{t_0}^{t} f(u, \phi(u))\, du$$

by (). Since the left side is $\phi_n(t)$, we obtain

$$\phi(t) = x_0 + \int_{t_0}^{t} f(u, \phi(u))\, du$$

for all $t \in I_0$, and thus ϕ is a solution to the original IVP on I_0. The equation $\phi'(t) = f(t, \phi(t))$ shows that ϕ is $\mathcal{C}^1$, since the right side is a continuous function of t. 👍

XI.1.7. Power Series and Analytic Solutions

XI.1.8. Linear ODEs

XI.1.9. Boundary-Value Problems

XI.1.9.1. So far, we have only considered initial-value problems, ones where the value of the solution and its lower-order derivatives at a single point are specified. But an important and much more difficult type of problem, especially for second-order differential equations, is a *boundary-value problem*, where the values of the solution at two points are specified.

XI.1.9.2. Suppose we have a second-order differential equation

$$y'' = f(t, y, y')$$

in standard form. The general solution generically has two arbitrary constants, so is uniquely determined by specifying $y(t_0)$ and $y'(t_0)$ for some t_0. But it is often more important to specify the values $y(t_0)$ and $y(t_1)$ for two distinct values t_0 and t_1 of t. In principle, the two constants should be uniquely determined by these values, at least in good cases. But there are some potential difficulties, which turn out to be real difficulties:

(i) In an initial-value problem, the existence and uniqueness theorems only guarantee a solution in *some* interval around t_0, although in some cases a careful analysis can give the approximate or even exact length of the interval. But (if we take $t_0 < t_1$ without loss of generality) in a boundary-value problem we must have a solution on the entire closed interval $[t_0, t_1]$.

(ii) There may be no values of the constants which give the desired boundary values, so the problem may not have a solution even if the differential equation does on the whole interval $[t_0, t_1]$.

(iii) There may be more than one choice of the constants giving the desired boundary values, so even if there is a solution it may not be unique.

All these difficulties can occur. And there are no really general existence and uniqueness theorems in the boundary-value case.

XI.1.9.3. For a dramatic example of (iii), leading to an example of (ii), consider one of the simplest second-order differential equations:

$$y'' = -y \ .$$

The general solution () is

$$y(t) = A\cos t + B\sin t$$

where A, B are constants. If we consider the boundary-value problem $y(0) = 0$, $y(2\pi) = 0$, then

$$y(t) = B\sin t$$

is a solution for any constant B. More generally, every solution to the differential equation is periodic with period 2π, i.e. $y(0) = y(2\pi)$, so if $x_1 \neq x_0$ the boundary-value problem

$$y'' = -y \ , \quad y(0) = x_0 \ , \quad y(2\pi) = x_1$$

has no solution. And one could not directly predict the periodicity from the form of the equation and anticipate these difficulties without some analysis or experience.

XI.1.10. Exercises

XI.1.10.1. Let I be an interval, $t_0 \in I$, and $f : I \times \mathbb{R} \to \mathbb{R}$ a continuous function. Consider the differential equation

$$y' = f(t, y)$$

and the integral equation

$$y(t) = y(t_0) + \int_{t_0}^{t} f(u, y(u))\, du \ .$$

The aim of this problem is to show that a function on I satisfies the differential equation if and only if it satisfies the integral equation.

(a) Suppose y is a differentiable function on I satisfying the differential equation. Show that y is necessarily $\mathcal{C}^1$ on I. [Use continuity of f.]

(b) Suppose y is a differentiable function on I satisfying the differential equation. Integrate both sides of the differential equation to show that y also satisfies the integral equation. [Use the Fundamental Theorem of Calculus II.]

(c) Conversely, suppose y is a function on I such that the integral in the integral equation exists as a Riemann, Lebesgue, or H-K integral for each $t \in I$, and that y satisfies the integral equation. Show that y is necessarily continuous on I, and hence the integral in the integral equation exists as a Riemann integral for each $t \in I$. [Use ().]

(d) Under the assumptions of (c), show that y is differentiable on I and satisfies the differential equation. [Use the Fundamental Theorem of Calculus I.]

(e) Extend the results to the case where the range space $\mathbb{R}$ is replaced by a Banach space.

XI.1.10.2. Implicit Differential Equations. (a) Prove the following theorem about existence and uniqueness of solutions to a differential equation of the form $F(t, y, y') = 0$.

THEOREM. Let U be an open set in $\mathbb{R}^3$, and $(t_0, x_0, z_0) \in U$. Let $F : U \to \mathbb{R}$ be a continuous function with $F(t_0, x_0, z_0) = 0$. Suppose $\frac{\partial F}{\partial z}$ exists and is continuous on U and that $\frac{\partial F}{\partial z}(t_0, x_0, z_0) \neq 0$. Then

(i) There is an open interval I around t_0 and a $\mathcal{C}^1$ function $y : I \to \mathbb{R}$ such that $y(t_0) = x_0$, $y'(t_0) = z_0$, and $F(t, y(t), y'(t)) = 0$ for all $t \in I$.

(ii) If $\frac{\partial F}{\partial y}$ also exists and is continuous on U, then there is an open interval I around t_0 and a unique differentiable function $y : I \to \mathbb{R}$ such that $y(t_0) = x_0$, $y'(t_0) = z_0$, and $F(t, y(t), y'(t)) = 0$ for all $t \in I$, and additionally y is $\mathcal{C}^1$ on I.

(iii) If F is analytic on U, then there is an open interval I around t_0 and a unique differentiable function $y : I \to \mathbb{R}$ such that $y(t_0) = x_0$, $y'(t_0) = z_0$, and $F(t, y(t), y'(t)) = 0$ for all $t \in I$, and additionally y is analytic on I.

[First use the Implicit Function Theorem to rewrite the equation $F(t, y, z) = 0$ as $z = f(t, y)$. Apply VIII.5.3.4. and XI.1.6.1. in case (i), VIII.5.3.12. and XI.1.2.7. in case (ii), and () and () in case (iii).]

However, such an implicit differential equation need not have any solutions, i.e. there may be no (t_0, x_0, z_0) satisfying the hypotheses: see Exercise XI.1.10.3..

(b) Generalize to systems and to higher-order implicit equations.

XI.1.10.3. Show that the (autonomous) implicit differential equation

$$(y' - 1)^2 + y^2 = 0$$

has no (real) solutions.

XI.1.10.4. Implicit Solutions. (a) Carefully define what it means for a relation $G(t,y)=0$ to be a "solution" to the differential equation $F(t,y,y')=0$ (or a higher-order equation).

(b) [**?**, p. 3] Show that

$$t^2 = 2y^2 \log y$$

is an implicit solution to the differential equation

$$y' = \frac{ty}{t^2+y^2} \ .$$

Implicit solutions are particularly easily obtained for exact differential equations (ones which can be put in the form $g(t,y)y' + h(t,y) = 0$ with $\frac{\partial g}{\partial t} = \frac{\partial h}{\partial y}$), cf. VIII.2.2.8., and in particular for separable differential equations (ones which can be put in the form $g(y)y' = h(t)$).

XI.1.10.5. This is a sort of converse to XI.1.10.4.. Let

$$G(t, y, c_1, \dots, c_n)$$

be a $\mathcal{C}^n$ function with $\frac{\partial G}{\partial y} \neq 0$. Regard t as the independent variable, y the dependent variable, and $c_1, \dots, c_n$ as "constants" (parameters). By the Implicit Function Theorem, the equation

$$G(t, y, c_1, \dots, c_n) = 0$$

implicitly defines y as a function of t and the parameters $c_1, \dots, c_n$, i.e. $y = g(t, c_1, \dots, c_n)$.

(a) Successively differentiate y with respect to t to obtain n additional equations

$$\frac{\partial G}{\partial t} + \frac{\partial G}{\partial y} y' = 0$$

$$\frac{\partial^2 G}{\partial t^2} + 2\frac{\partial^2 G}{\partial t\, \partial y} y' + \frac{\partial^2 G}{\partial y^2}(y')^2 + \frac{\partial G}{\partial y} y'' = 0$$

$$\cdots$$

$$\frac{\partial^n G}{\partial t^n} + \cdots + \frac{\partial G}{\partial y} y^{(n)} = 0$$

(where we really mean $y^{(k)} = \frac{\partial^k g}{\partial t^k}$, but we symbolically regard the $y^{(k)}$ as new variables).

(b) Show that if $\boldsymbol{c} = (c_1, \dots, c_n)$ is independent in G in a suitable sense, then by the Implicit Function Theorem the first n equations can (in principle) be solved for $c_1, \dots, c_n$ as functions of $t, y, y', \dots, y^{(n-1)}$ and substituted into the last equation to obtain a differential equation

$$F(t, y, y', \dots, y^{(n-1)}, y^{(n)}) = 0$$

for which $G(t, y, c_1, \dots, c_n) = 0$ is a solution (in the sense of XI.1.10.4.) for any suitable $\boldsymbol{c}$ (references such as [Gou59, V. 2 Part 2] and [Inc44] state that $c_1, \dots, c_n$ "can be eliminated" from the $n+1$ equations without elaboration; in [**?**] it is at least acknowledged there is a nontrivial issue, but that "a precise statement of the conditions under which this is true is too complicated to be stated here.") Use the existence and uniqueness theorem to show that, under appropriate conditions, every solution is of this form.

(c) Generalize to the case where there is more than one G, y, and/or t (the same number of G's and y's) to obtain a system of ordinary or partial differential equations from an equation or set of equations with an arbitrary (finite) set of constants.

XI.1.10.6. Consider the IVP $y' = y^2$, $y(t_0) = x_0$ (XI.1.1.6.(ii)). Theorem XI.1.2.2. guarantees a unique solution on

$$\left[t_0 - \frac{r}{(|x_0| + r)^2}, t_0 + \frac{r}{(|x_0| + r)^2}\right]$$

for any $r > 0$. If $x_0 = 0$, the supremum of the lengths of these intervals for all r is $+\infty$, and the solution to the IVP $(y = 0)$ is defined on all of $\mathbb{R}$. If $x_0 \neq 0$, show that the maximum of $\frac{r}{(|x_0|+r)^2}$ is $\frac{1}{4|x_0|}$. In this case, the unique solution $y = \frac{1}{c-t}$, $c = t_0 + \frac{1}{x_0}$, has domain $(-\infty, c)$ if $x_0 > 0$ and $(c, +\infty)$ if $x_0 < 0$, so the maximum symmetric interval around t_0 on which the solution exists is $\left(t_0 - \frac{1}{|x_0|}, t_0 + \frac{1}{|x_0|}\right)$.

XI.1.10.7. Show that Theorem XI.1.3.2. can be generalized as follows to give a unique solution on a generally larger interval. Suppose we have a Banach space $\mathcal{X}$, an interval $I \subseteq \mathbb{R}$, a point $t_0 \in I$, n points $x_1^{(0)}, \ldots, x_n^{(0)}$ in $\mathcal{X}$, $r_1, \ldots, r_n > 0$, $M > 0$, $K > 0$, and n continuous functions

$$f_1, \ldots, f_n : I \times \bar{B}_{r_1}(x_1^{(0)}) \times \cdots \times \bar{B}_{r_n}(x_n^{(0)}) \to \mathcal{X}$$

with the property that $\|f_k(t, x_1, \ldots, x_n)\| \leq M_k$ for $t \in I$ and $\|x_j - x_j^{(0)}\| \leq r_j$ $(1 \leq j, k \leq n)$, and

$$\|f_k(t, x_1^{(2)}, \ldots, x_n^{(2)}) - f_k(t, x_1^{(1)}, \ldots, x_n^{(1)})\| \leq K \max_{1 \leq j \leq n} (\|x_j^{(2)} - x_j^{(1)}\|)$$

for $1 \leq k \leq n$ whenever $t \in I$ and $x_j^{(1)}, x_j^{(2)} \in \bar{B}_{r_j}(x_j^{(0)})$ for all j (the f_k satisfy a uniform Lipschitz condition in $x_1, \ldots, x_n$ on I). Slightly adapt the proof of Theorem XI.1.2.2. to show the following:

THEOREM. [EXISTENCE AND UNIQUENESS THEOREM FOR SYSTEMS OF FIRST-ORDER ODES] Using the above notation, set

$$L = \min\left\{\frac{r_1}{M_1}, \ldots, \frac{r_n}{M_n}\right\}$$
$$I_0 = I \cap [t_0 - L, t_0 + L] \ .$$

Then there are unique differentiable functions $y_1, \ldots, y_n : I_0 \to \mathcal{X}$ such that

$$y_k'(t) = f_k(t, y_1(t), \ldots, y_n(t))$$

$(1 \leq k \leq n)$ for $t \in I_0$ and $y_k(t_0) = x_k^{(0)}$. Additionally, the y_k are $\mathcal{C}^1$ on I_0. If $\tilde{I}$ is any subinterval of I_0 containing t_0, the restriction of the y_k to $\tilde{I}$ is the unique solution to the IVP on $\tilde{I}$.

XI.1.10.8. Use the theorem in XI.1.10.7. to obtain the following generalization of Theorem XI.1.4.1. which guarantees a unique solution on a larger interval in general:

THEOREM. [EXISTENCE AND UNIQUENESS THEOREM FOR HIGHER-ORDER ODE'S] Let $\mathcal{X}$ be a Banach space, I an interval in $\mathbb{R}$, $t_0 \in I$, $x_0, x_0', \ldots, x_0^{(n-1)} \in \mathcal{X}$, $r_0, \ldots, r_{n-1} > 0$, and

$$f : I \times \bar{B}_{r_0}(x_0) \times \bar{B}_{r_1}(x_0') \times \cdots \times \bar{B}_{r_{n-1}}(x_0^{(n-1)}) \to \mathcal{X} \ .$$

Suppose f is continuous and bounded by N on $I \times \bar{B}_r(x_0) \times \bar{B}_r(x_0') \times \cdots \times \bar{B}_r(x_0^{(n-1)})$ and satisfies a uniform Lipschitz condition in the y-variables. Set

$$L = \max\left\{ \frac{r_0}{\|x_0'\| + r_1}, \ldots, \frac{r_{n-2}}{\|x_0^{(n-1)}\| + r_{n-1}}, \frac{r_{n-1}}{N} \right\}$$

$$I_0 = I \cap [t_0 - L, t_0 + L] \ .$$

Then there is a unique n-times differentiable function $y : I_0 \to \mathcal{Y}$ with $y(t_0) = x_0$, $y'(t_0) = x_0'$, $\ldots$, $y^{(n-1)}(t_0) = x_0^{(n-1)}$, and $y^{(n)}(t) = f(t, y(t), y'(t), \ldots, y^{(n-1)}(t))$ for all $t \in I_0$. Additionally, y is $\mathcal{C}^n$ on I_0.

XI.1.10.9. Let I be an interval in $\mathbb{R}$, $\mathcal{X}$ a Banach space, and $f : I \to \mathcal{X}$ a continuous function. Then f may be regarded as a continuous function from $I \times \mathcal{X}$ to $\mathcal{X}$ satisfying a uniform Lipschitz condition in $\mathcal{X}$ with Lipschitz constant 0. Show that one step of Picard iteration gives the exact (unique) solution to the IVP $y' = f(t)$, $y(t_0) = x_0$ for any $t_0 \in I$, $x_0 \in \mathcal{X}$.

XI.1.10.10. Let $I = (0, +\infty)$. Define $f : I \times \mathbb{R} \to \mathbb{R}$ by

$$f(t, y) = \frac{y \cos(\log t)}{t} \ .$$

(a) If $(t_0, x_0) \in I \times \mathbb{R}$, show that the hypotheses of XI.1.2.9., (), and () are satisfied for the IVP

$$y' = f(t, y), \quad y(t_0) = x_0$$

(with $\mathcal{X} = \mathbb{R}$), and thus there is a unique solution, which is defined on the entire interval I.

(b) Show that the general solution is $y(t) = Ce^{\sin(\log t)}$ for $C \in \mathbb{R}$.

(c) Show that every solution is bounded and analytic on I, but no solution other than the identically zero one extends to be continuous at 0.

XI.1.10.11. [**?**] This exercise shows how differently even first-order differential equations not in standard form can behave. Define $F : \mathbb{R}^3 \to \mathbb{R}$ by

$$F(t, y, z) = y^2 + z^2 - 1 \ .$$

Then F is $\mathcal{C}^\infty$, even analytic, on all of $\mathbb{R}^3$. Consider the autonomous first-order differential equation

$$F(t, y, y') = 0$$

which can be rewritten

$$(y')^2 = 1 - y^2 \ .$$

(a) Show that

$$y = \sin(t - c)$$

is a solution for any $c \in \mathbb{R}$, and the constant functions $y = 1$ and $y = -1$ are also solutions. [The solutions $y = \cos(t - b)$ can be written in the form $y = \sin(t - c)$ for $c = b - \frac{\pi}{2}$.]

(b) Show that for any $t_0 \in \mathbb{R}$ and $-1 < x_0 < 1$ there are locally two solutions to the IVP

$$F(t, y, y') = 0 \ , \quad y(t_0) = x_0$$

and there are at least two local solutions if $x_0 = \pm 1$.

(c) The solutions in (a) are not the only solutions, however. For example, there is the solution

$$y(t) = \begin{cases} \sin t & \text{if} \quad -\frac{\pi}{2} \leq t \leq \frac{\pi}{2} \\ -1 & \text{if} \quad t < -\frac{\pi}{2} \\ 1 & \text{if} \quad t > \frac{\pi}{2} \end{cases}$$

or, more generally, (continuous) functions whose graph consists of segments of graphs of $y = \sin(t - c)$ for various c interspersed with segments of constant functions with value ± 1. These solutions, of which there are very many, are $\mathcal{C}^1$ but not even twice differentiable everywhere; compare with XI.1.1.2.. Thus the IVP of the form

$$F(t, y, y') = 0 \ , \quad y(t_0) = x_0$$

for $x_0 = \pm 1$ has many solutions on any interval around t_0 for any $t_0 \in \mathbb{R}$.

(d) Show that the solutions in (a) and (c) are all the solutions to the differential equation.

(e) Discuss how the solutions to this differential equation are related to the solutions to the two standard form differential equations

$$y' = \sqrt{1 - y^2} \ , \quad y' = -\sqrt{1 - y^2} \ .$$

XI.1.10.12. (a) Use the Existence and Uniqueness Theorem XI.1.2.7. to give an alternate proof of the one-variable $\mathcal{C}^r$ Implicit Function Theorem VIII.5.3.14. for $r \geq 2$ by considering the IVP (with x the independent variable)

$$y' = -\frac{\frac{\partial F}{\partial x}(x, y)}{\frac{\partial F}{\partial y}(x, y)} \ , \ y(a) = b \ .$$

Use the Chain Rule to show that the unique solution $y = f(x)$ satisfies $F(x, f(x)) = 0$ for all x near a. (One does not obtain uniqueness of the implicitly defined function this way, only uniqueness of a differentiable implicitly defined function.) What can one conclude from this argument if F is only $\mathcal{C}^1$? (cf. XI.1.6.1.)

(b) Generalize to the case of VIII.5.1.7. for $p = 1$ and $r \geq 2$, for general n, by considering a system of differential equations.

(c) Give a similar proof of the general $\mathcal{C}^r$ Implicit Function Theorem VIII.5.1.7. ($r \geq 2$) by considering a system of partial differential equations

$$D\boldsymbol{y} = -D_{\boldsymbol{y}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y})^{-1} D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{y}) \ , \ \boldsymbol{y}(\boldsymbol{a}) = \boldsymbol{b} \ .$$

(d) Use these results to give explicit descriptions of sets on which the implicitly defined functions exist.

XI.1.10.13. For positive functions f and g on $[a, +\infty)$ write $f \approx g$ if

$$0 < \inf_{y \geq a} \frac{f(y)}{g(y)} \leq \sup_{y \geq a} \frac{f(y)}{g(y)} < \infty$$

(f and g are the "same order"; this is stronger than $f(y) = O(g(y))$ as $y \to +\infty$). In this problem, fix a positive function f on $[0, +\infty)$ which is Lipschitz on $[0, n]$ for all n (it suffices that f is $\mathcal{C}^1$), and consider the autonomous IVP $y' = f(y)$, $y(0) = x_0 > 0$.

(a) Show that the IVP has a unique solution $y(t)$ on some maximal interval $[0, b)$, where $0 < b \leq +\infty$.

(b) If $f(y) \approx y^\beta$ on $[a, +\infty)$ for some (hence all) $a > 0$, for some β, $0 \leq \beta < 1$, then $b = +\infty$ and $y(t) \approx t^{1/(1-\beta)}$ as $t \to +\infty$. Thus y has polynomial growth.

(c) If $f(y) \approx y$ on $[a, +\infty)$ for some (hence all) $a > 0$, then $b = +\infty$ and $\log y(t) \approx 1$. Thus y has exponential growth.

(c) If $f(y) \approx y^\alpha$ on $[a, +\infty)$ for some (hence all) $a > 0$, for some $\alpha > 1$, then $b < +\infty$ and $\lim_{t \to b-} y(t) = +\infty$, i.e. y "blows up" at b and the solution cannot be extended beyond b.

XI.2. Partial Differential Equations

> "PDEs constitute one of the most dynamic and varied domains of the mathematical sciences, defying all attempts at unification. They are found in every phenomenon studied by the physics of continuous systems, involving all states of matter (gases, fluids, solids, plasmas) and all physical theories (classical, relativistic, quantum, and so on).
>
> But partial differential equations also lurk behind many geometric problems. Geometric PDEs, as they are called, make it possible to deform geometric objects in accordance with well-established laws. The application of methods of analysis to problems in other fields of mathematics is an example of the sort of cross-fertilization that became increasingly common in the course of the twentieth century."
>
> *C. Villani*[3]

XI.2.1. Elementary Theory

It is common to denote the unknown function in a partial differential equation by u (or v, w), and the independent variables by x, y, z, and/or t. Of course, a PDE can have any number of independent variables, but PDE's arising in physics often describe an unknown function of three (or fewer) spatial coordinates (x, y, z) and/or one time coordinate t. It is also notationally economical to use the subscript notation for partial derivatives, e.g. u_x for $\frac{\partial u}{\partial x}$ or u_{xy} for $\frac{\partial^2 u}{\partial y \partial x}$ (note the order of differentiation, which is often but not always irrelevant; cf. VIII.2.2.2.).

XI.2.1.1. Example. Consider the apparently extremely simple PDE which can be variously written

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial t}$$

$$\frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} = 0$$

$$u_x - u_t = 0$$

where u is an unknown function of x and t. It is easy to check using the Chain Rule that

$$u(x,t) = f(x+t)$$

is a solution on $\Omega = \{(x,t) : x + t \in I\}$, for any real-valued function f which is differentiable on an open interval I in $\mathbb{R}$ (it turns out that these are all the solutions ()). There are several observations to take away from this example:

(i) There is not a clean standard form for PDEs such as in XI.1.1.2. (although see ()).

(ii) Even a homogeneous linear first-order PDE with constant coefficients can have solutions which are not even $\mathcal{C}^1$, much less analytic (cf. ()).

[3] [Vil16, p. 107].

(iii) The general solution to such a PDE does not just contain an arbitrary constant, but an arbitrary (differentiable) function which must be determined by initial or boundary conditions (the fact that there is one equation and two independent variables suggests that there is an arbitrary function of $2 - 1 = 1$ variables which comes into the general solution, which is correct when properly formulated ()).

XI.2.1.2. Here are a couple of worse examples from [**?**]:

(i) The general solution to the first-order PDE

$$u_t = 0$$

where u is an unknown function of x and t is a function of the form

$$u(x,t) = f(x)$$

where f is an arbitrary function, not necessarily even continuous.

(ii) The general solution to the second-order PDE

$$u_{tx} = 0$$

where u is an unknown function of x and t is a function of the form

$$u(x,t) = f(t) + g(x)$$

where f is a differentiable function and g is an arbitrary function, not necessarily even continuous. Thus, for a solution u, the mixed partial u_{xt} is not necessarily even defined, so, according to [**?**], "we are thus led to distinctions which it is embarrassing to have to make."

> "The solving of equations satisfying given boundary conditions is seldom simple and is a task for the skilled mathematician. It cannot be learned in a day. If you want to do this sort of thing as a lifework, start at sixteen or seventeen, not later than nineteen."
>
> *E. T. Bell*[4]

Well, then, I guess some of us are out of luck. (But Bell wasn't always right!)

XI.2.2. Quasilinear Partial Differential Equations

Lectures of S. N. Kruzhkov https://www.math.ntnu.no/conservation/2009/011.pdf

XI.2.3. The Cauchy Problem

A great many (but not all) important PDE problems can be put in the following very general standard form.

[4] [Bel87, p. 334].

XI.2.3.1. DEFINITION. Let N and n_j $(1 \le j \le N)$ be natural numbers, and n a nonnegative integer. A *Cauchy system* is a system of N partial differential equations of the form

$$\frac{\partial^{n_j} u_j}{\partial t^{n_j}} = F_j\left(t, x_1, \ldots, x_n, u_1, \ldots, u_N, \ldots, \cdots \frac{\partial^k u_r}{\partial t^{k_0} \partial x_1^{k_1} \cdots \partial x_n^{k_n}} \cdots\right)$$

of unknown functions $u_1, \ldots, u_N$ of $n+1$ variables $t, x_1, \ldots, x_n$ where N and n_j $(1 \le j \le N)$ are natural numbers, n a nonnegative integer, $1 \le r \le N$, $0 \le k_0 < n_j$, $0 \le k_1, \ldots, k_n \le n_j$ (r, k, and the k_i may vary from entry to entry) with $k_0 + \cdots + k_n = k \le n_j$.

In other words, none of the partials appearing on the right are of order greater than the partial on the left, and none are even of equal order with respect to t. Note that the order n_j can vary from equation to equation, but the u_j are all functions of the same independent variables.

We can specify a set of "initial conditions":

XI.2.3.2. DEFINITION. A *set of Cauchy data* for the Cauchy system in XI.2.3.1. around $(t_0, x_1^0, \ldots, x_n^0)$ is a set of (usually continuous) functions

$$\{\phi_{j,k} : 1 \le j \le N, 0 \le k \le n_j\}$$

defined on an open neighborhood U of $(x_1^0, \ldots, x_n^0)$ in $\mathbb{R}^n$.

A *Cauchy problem* is a Cauchy system with a set of Cauchy data. An *analytic Cauchy problem* is a Cauchy problem in which the F_j and the data functions $\phi_{j,k}$ are analytic.

A *solution* to a Cauchy problem is a set $\{u_1, \ldots, u_N)$ of sufficiently differentiable functions on a set of the form $I \times V$, where I is an interval containing t_0 and V an open neighborhood of $(x_1^0, \ldots, x_n^0)$ contained in U, such that the u_j satisfy the differential equations and

$$\frac{\partial^k u_j}{\partial t^k}(t_0, x_1, \ldots, x_n) = \phi_{j,k}(x_1, \ldots, x_n)$$

for all j and k and all $(x_1, \ldots, x_n) \in V$. If the Cauchy problem is analytic, the solution is *analytic* provided the u_j are all analytic in a neighborhood of $(t_0, x_1^0, \ldots, x_n^0)$.

XI.2.3.3. EXAMPLES. (i) Any ODE in standard (Cauchy) form () is a Cauchy system with $N = 1$ and $n = 0$. A set of Cauchy data is a set of functions of no variables (i.e. constants); thus an ODE IVP is a Cauchy problem, which is analytic if the F is analytic (the $\phi_{j,k}$ are vacuously analytic). More generally, a (finite) system of ODEs with initial conditions is a Cauchy system.

(ii) Any quasilinear PDE () is a Cauchy system provided there is a pure partial of maximal order (put its term on one side and the rest of the terms on the other, and divide by the coefficient function). It is analytic if all the coefficient functions are analytic and the denominator coefficient function is nonzero at $(t_0, x_1^0, \ldots, x_n^0)$.

XI.2.4. The Cauchy-Kovalevskaya Theorem

XI.2.4.1. THEOREM. [CAUCHY-KOVALEVSKAYA THEOREM] An analytic Cauchy problem around a point $(t_0, x_1^0, \ldots, x_n^0)$ has a unique analytic solution in some open neighborhood of $(t_0, x_1^0, \ldots, x_n^0)$.

XI.2.4.2. Note that the uniqueness is only among *analytic* solutions. An analytic Cauchy problem can possibly have nonanalytic solutions too (). The theorem includes () as a special case (the case $n = 0$).

Chapter XII

Fundamentals of Complex Analysis

Complex Analysis is calculus done on the complex numbers. Although at first it bears superficial resemblance to ordinary calculus, it quickly assumes a much different character.

Complex Analysis is one of the most beautiful parts of mathematics. Many of the theorems seem almost too good to be true, especially in comparison with the situation in Real Analysis. Complex Analysis is also extraordinarily important throughout mathematics, with applications ranging from number theory to geometry to mathematical physics (and almost everything else).

It is beyond the scope of this book to give a full treatment of this vast subject. We will only describe some of the most fundamental aspects and their connections with real analysis. There is an overabundant supply of texts on Complex Analysis, which are of varying quality but mostly adequate or better. Four of the best are the classic [Ahl78], [GK06], [Gam01], and [SS03].

Throughout this chapter, we will always use the standard model of the complex numbers $\mathbb{C}$, the complex plane (III.6.2.). Thus we will freely regard $\mathbb{C}$ as $\mathbb{R}^2$ with the additional operation of complex multiplication. Thus any function from $\mathbb{C}$ (or a subset) to $\mathbb{C}$ can be equally well thought of as a function from $\mathbb{R}^2$ to $\mathbb{R}^2$. In discussing limits and continuity in $\mathbb{C}$, we always use the topology on $\mathbb{C}$ obtained by the identification with $\mathbb{R}^2$. This topology is induced by the metric $\rho(z, w) = |z - w|$.

XII.1. Elementary Theory

XII.1.1. Complex Differentiation and Holomorphic Functions

We first discuss the complex analog of ordinary differentiation. Many basic definitions and results bear a close formal resemblance to corresponding notions in usual calculus of one variable.

XII.1.1.1. We will use the following standard notation. The independent variable in complex analysis is usually denoted $z = x + iy$, where $z \in \mathbb{C}$ and x and y are the real and imaginary parts of z. Then z corresponds to the point $(x, y) \in \mathbb{R}^2$. If Ω is a (usually open) subset of $\mathbb{C}$, and $f : \Omega \to \mathbb{C}$ is a function, write

$$f(z) = f(x + iy) = u(x, y) + iv(x, y)$$

where $u, v : \Omega \to \mathbb{R}$ are the real and imaginary parts of f (the coordinate functions of f when f is regarded as a function from Ω to $\mathbb{R}^2$).

XII.1.1.2. DEFINITION. Let Ω be an open subset of $\mathbb{C}$, $f : \Omega \to \mathbb{C}$ a function, and $z_0 = x_0 + iy_0 \in \Omega$. Then f is *differentiable in the complex sense* (or just *complex differentiable*) at z_0 if

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists. If the limit exists, it is called the *complex derivative* of f at z_0, denoted $f'(z_0)$.

The function f is *complex differentiable on* Ω, or *holomorphic on* Ω, if it is complex differentiable at every $z_0 \in \Omega$.

XII.1.1.3. In Complex Analysis, the term *analytic* is often used as a synonym for *holomorphic*. We will not initially use *analytic* in this way since it has a different official definition (V.16.1.1.). One of the major fundamental theorems of Complex Analysis is that the notions of being analytic and holomorphic on an open set Ω coincide (XII.4.4.7.).

Here is the relation between complex differentiation and ordinary differentiation on $\mathbb{R}^2$:

XII.1.1.4. PROPOSITION. Let Ω be an open subset of $\mathbb{C}$, $f : \Omega \to \mathbb{C}$ a function, and $z_0 = x_0 + iy_0 \in \Omega$, $c = a + ib \in \mathbb{C}$. The following are equivalent:

(i) f is complex differentiable at z_0 with $f'(z_0) = c$.

(ii) For $z \in \Omega$, $f(z) = f(z_0) + c(z - z_0) + o(|z - z_0|)$.

(iii) f (regarded as a function from Ω to $\mathbb{R}^2$) is differentiable at $z_0 = (x_0, y_0)$ in the vector calculus sense (VIII.3.1.4.) and

$$Df(x_0, y_0) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

This result gives motivation and "explanation" for the model of $\mathbb{C}$ in Exercise III.6.3.1..

PROOF: The argument is very similar to the calculation motivating the definition of vector calculus differentiation (VIII.3.1.2.). The following statements are equivalent:

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = c$$

$$\lim_{z \to z_0} \frac{f(z) - f(z_0) - c(z - z_0)}{z - z_0} = 0$$

$$\lim_{z \to z_0} \frac{|f(z) - f(z_0) - c(z - z_0)|}{|z - z_0|} = 0$$

$$\lim_{(x,y) \to (x_0,y_0)} \frac{\|f(x, y) - f(x_0, y_0) - T(\boldsymbol{h})\|}{\|\boldsymbol{h}\|} = 0$$

where $\boldsymbol{h} = (x, y) - (x_0, y_0)$ and T is the linear transformation of $\mathbb{R}^2$ given by multiplication by c. The argument is completed by noting that the matrix of T is $\begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. 👍

XII.1.1.5. Corollary. Let Ω be an open subset of $\mathbb{C}$, $f = u + iv : \Omega \to \mathbb{C}$ a function, and $z_0 = x_0 + iy_0 \in \Omega$. Then f is complex differentiable at z_0 if and only if u and v are differentiable in the vector calculus sense at (x_0, y_0) and satisfy the *Cauchy-Riemann equations*:

$$\frac{\partial u}{\partial x}(x_0, y_0) = \frac{\partial v}{\partial y}(x_0, y_0), \quad \frac{\partial u}{\partial y}(x_0, y_0) = -\frac{\partial v}{\partial x}(x_0, y_0) \ .$$

The complex derivative, when it exists, is given by

$$f'(x_0 + iy_0) = \frac{\partial u}{\partial x}(x_0, y_0) + i\frac{\partial v}{\partial x}(x_0, y_0) = \frac{\partial v}{\partial y}(x_0, y_0) - i\frac{\partial u}{\partial y}(x_0, y_0) \ .$$

So f is holomorphic on Ω if and only if f is differentiable in the vector calculus sense on Ω and the Cauchy-Riemann equations hold everywhere on Ω. In particular, if $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ are continuous on Ω and satisfy the Cauchy-Riemann equations, then f is holomorphic on Ω.

XII.1.1.6. It turns out that in XII.1.1.5., the assumption that f be differentiable in the vector calculus sense on Ω can be relaxed. If the first-order partials of u and v exist everywhere on Ω and satisfy the Cauchy-Riemann equations, and if f is continuous on Ω (Looman-Menchoff Theorem) or just locally bounded (Montel-Tolstoff Theorem), then f is automatically holomorphic on Ω. Some additional condition on f besides the Cauchy-Riemann equations is necessary: cf. Exercise XII.9.9.1.. The question of when a function satisfying the Cauchy-Riemann equations is holomorphic is a subtle one, which is discussed thoroughly in [GM78]. (And a strong version of the converse of the last statement of XII.1.1.5. also holds (XII.3.3.6.).)

XII.1.1.7. XII.1.1.4. suggests that it is qualitatively harder for a function from $\mathbb{C}$ to $\mathbb{C}$ to be complex differentiable than it is for a function from $\mathbb{R}$ to $\mathbb{R}$ to be differentiable in the usual sense. This turns out to be correct, and complex differentiability implies all sorts of remarkable properties that do not necessarily hold for usual differentiable functions on $\mathbb{R}$.

There are nonetheless many examples of holomorphic functions, enough to make a rich theory. Here are some basic ones:

XII.1.1.8. Examples. (i) If f is a polynomial in z with complex coefficients, then by a calculation using the product formula in $\mathbb{C}$ the real and imaginary parts of f are polynomials in x and y with real coefficients. Thus f is differentiable (in fact, $\mathcal{C}^\infty$) on $\mathbb{C}$. A rather ugly calculation shows that the Cauchy-Riemann equations are satisfied, so f is holomorphic on $\mathbb{C}$. A simpler argument which shows this, and that f' is given by the usual formula, is in XII.1.2.5..

(ii) Similarly, any rational function in z is holomorphic on its domain (XII.1.2.8.).

(iii) [Complex Exponential Function] For $z = x + iy \in \mathbb{C}$, set

$$e^z = e^{x+iy} = e^x \cos y + ie^x \sin y = e^x(\cos y + i \sin y) \ .$$

Then $f(z) = e^z$ is differentiable (in fact, $\mathcal{C}^\infty$) on $\mathbb{C}$, and it is easily checked that the Cauchy-Riemann equations hold; hence f is holomorphic on $\mathbb{C}$. By XII.1.1.4.,

$$f'(x + iy) = \frac{\partial u}{\partial x}(x, y) + i\frac{\partial v}{\partial x}(x, y) = f(x + iy)$$

i.e. $f' = f$. The complex exponential function also satisfies $e^{z+w} = e^z e^w$, $e^{-z} = \frac{1}{e^z}$, and $e^{z-w} = \frac{e^z}{e^w}$ for all $z, w \in \mathbb{C}$, and agrees with the usual exponential function for z real (and is the only holomorphic function extending the real exponential function, cf. Exercises XII.9.9.6. and XII.9.9.10.).

(iv) [Complex Trigonometric Functions] For $z = x + iy \in \mathbb{C}$, set

$$\sin z = \sin x \cosh y + i \cos x \sinh y$$

$$\cos z = \cos x \cosh y - i \sin x \sinh y$$

Then $f(z) = \sin z$ and $g(z) = \cos z$ are differentiable (in fact, $\mathcal{C}^\infty$) on $\mathbb{C}$, and the Cauchy-Riemann equations hold, so f and g are holomorphic on $\mathbb{C}$. By a calculation as in (iii), $f' = g$ and $g' = -f$. These functions satisfy the usual trigonometric identities, and reduce to the usual sin and cos when z is real. The other trigonometric functions can be defined from these in the usual way, and give holomorphic functions on their domains extending the usual ones and satisfying the same identities and differentiation rules.

(v) [Complex Hyperbolic Functions] For $z = x + iy \in \mathbb{C}$, set

$$\sinh z = \sinh x \cos y + i \cosh x \sin y = \frac{e^z - e^{-z}}{2}$$

$$\cosh z = \cosh x \cos y + i \sinh x \sin y = \frac{e^z + e^{-z}}{2}$$

Then similarly $f(z) = \sinh z$ and $g(z) = \cosh z$ are holomorphic on $\mathbb{C}$, and $f' = g$, $g' = f$. These functions satisfy the usual identities, and reduce to the usual sinh and cosh when z is real. The other hyperbolic functions can be defined from these in the usual way, and give holomorphic functions on their domains extending the usual ones and satisfying the same identities and differentiation rules. We have

$$\cos z = \cosh iz = \frac{e^{iz} + e^{-iz}}{2}$$

$$\sin z = -i \sinh iz = \frac{e^{iz} - e^{-iz}}{2i}$$

for all $z \in \mathbb{C}$.

These examples show the intimate connections between the exponential, trigonometric, and hyperbolic functions which become clear only when the extensions to the complex numbers are considered.

> "... In effect, if one extends these functions by allowing complex values for the arguments, then there arises a harmony and regularity which without it would remain hidden."
>
> *B. Riemann, 1851*[1]

Defining inverse functions (log, arcsin, etc., even $\sqrt{z}$) is a trickier matter, primarily due to complications of domain.

XII.1.1.9. In this case, all these functions are $\mathcal{C}^\infty$ on their domains (as functions from $\mathbb{R}^2$ to $\mathbb{R}^2$), and their derivatives of all orders are also holomorphic. This in fact is true for any holomorphic function on an open set in $\mathbb{C}$ (XII.3.3.6.), one of the remarkable properties of holomorphic functions.

[1] [SS03, p. xv].

Local Behavior and the Conformal Property

XII.1.1.10. The geometric action on $\mathbb{C}$ of multiplying by a nonzero complex number $c = re^{i\theta}$ is scaling by r and rotating by angle θ counterclockwise around the origin. Thus any plane figure has its shape preserved and its dimensions multiplied by r. In particular, two smooth oriented curves in the plane intersecting at z_0 with angle ϕ between them (between their tangent vectors) are sent to oriented curves through cz_0 with the same angle ϕ between them.

XII.1.1.11. If f is holomorphic on a region Ω, $z_0 \in \Omega$, and $f'(z_0) = re^{i\theta} \neq 0$, then near z_0 f looks approximately like the function $g(z) = f(z_0) + f'(z_0)(z - z_0)$, which consists geometrically like rotation around z_0 by angle θ, scaling by r, and translation by $f(z_0)$. Thus two smooth oriented curves in the plane intersecting at z_0 with angle ϕ between them are sent by g to oriented curves through cz_0 with the same angle ϕ between them, and the same is true for f (this can be easily verified using the Chain Rule of vector calculus since the tangent vectors are both transformed by multiplication by $f'(z_0)$ regarded as a linear transformation of $\mathbb{R}^2$).

XII.1.1.12. Thus, if f is holomorphic on Ω and f' is never zero on Ω, then f is a *conformal* map of Ω to $f(\Omega)$: any two smooth oriented curves in Ω which intersect will be sent to smooth oriented curves intersecting at the same angle at the image of the point of intersection. The function f locally looks approximately like a scaling, rotation, and translation; the scaling factor and rotation angle vary smoothly from point to point. This is an extremely important property of holomorphic functions and leads to many applications.

XII.1.1.13. The conformal property does not hold at points of Ω where $f' = 0$. The fundamental example is $f(z) = z^n$ at 0: rays from 0 with angle ϕ between them are sent to rays with angle $n\phi$ between them.

XII.1.1.14. It turns out that these power functions describe the general behavior of holomorphic functions: if f is a nonconstant holomorphic function on a connected open set Ω, then

(i) The points of Ω where $f' = 0$ are isolated ().

(ii) At each point z_0 where $f'(z_0) = 0$ there is an $n > 1$ such that f locally looks like $g(z) = f(z_0)+c(z-z_0)^n$ for some $0 \neq c \in \mathbb{C}$, and hence angles between curves intersecting at z_0 are multiplied by n, and f is locally n-to-one around z_0 ().

(iii) f is conformal and locally one-to-one around all other points of Ω.

(iv) f is an *open mapping*: if U is any open subset of Ω, then $f(U)$ is open ().

XII.1.1.15. There is even a converse: if Ω is an open set in $\mathbb{C}$, and $f : \Omega \to \mathbb{C}$ is differentiable on Ω in the vector calculus sense, and conformal and orientation-preserving, then f is holomorphic on Ω.

XII.1.2. Rules of Differentiation

There are complex analogs of the usual rules of differentiation. These can be proved by an easy adaptation of the arguments in the real case, or by rather tediously using the real results, the formulas for complex multiplication and division, and XII.1.1.4..

XII.1.2.1. Proposition. Let f be a function from an open set $\Omega \subseteq \mathbb{C}$ to $\mathbb{C}$. If f is complex differentiable at $z_0 \in \Omega$, then f is continuous at z_0.

XII.1.2.2. Theorem. [Power Rule] Let $n \in \mathbb{N}$, and $f : \mathbb{C} \to \mathbb{C}$ be defined by $f(z) = z^n$. Then f is complex differentiable at any $z_0 \in \mathbb{C}$, and

$$f'(z_0) = nz_0^{n-1} \ .$$

XII.1.2.3. Proposition. [Sum Rule] Let f and g be functions from an open set $\Omega \subseteq \mathbb{C}$ to $\mathbb{C}$, and $h = f + g : \Omega \to \mathbb{C}$. If f and g are complex differentiable at $z_0 \in \Omega$, then h is complex differentiable at z_0 and

$$h'(z_0) = f'(z_0) + g'(z_0) \ .$$

XII.1.2.4. Proposition. [Constant Multiple Rule] Let f be a function from an open set $\Omega \subseteq \mathbb{C}$ to $\mathbb{C}$, $c \in \mathbb{C}$, and $h = cf : \Omega \to \mathbb{C}$. If f is complex differentiable at $z_0 \in \Omega$, then h is complex differentiable at z_0 and

$$h'(z_0) = cf'(z_0) \ .$$

XII.1.2.5. Corollary. Any sum or constant multiple of functions holomorphic on Ω is holomorphic on Ω. In particular, a polynomial with complex coefficients is holomorphic on $\mathbb{C}$.

XII.1.2.6. Theorem. [Product Rule] Let f and g be functions from an open set $\Omega \subseteq \mathbb{C}$ to $\mathbb{C}$, and $h = fg : \Omega \to \mathbb{C}$. If f and g are complex differentiable at $z_0 \in \Omega$, then h is complex differentiable at z_0 and

$$h'(z_0) = f'(z_0)g(z_0) + f(z_0)g'(z_0) \ .$$

XII.1.2.7. THEOREM. [QUOTIENT RULE] Let f and g be functions from an open set $\Omega \subseteq \mathbb{C}$ to $\mathbb{C}$. If f and g are complex differentiable at $z_0 \in \Omega$ and $g(z_0) \neq 0$, then $h = \frac{f}{g}$ is defined in an open neighborhood of z_0 and differentiable at z_0, and

$$h'(z_0) = \frac{g(z_0)f'(z_0) - f(z_0)g'(z_0)}{[g(z_0)]^2} \ .$$

XII.1.2.8. COROLLARY. If f and g are holomorphic on Ω, then fg is holomorphic on Ω, and if g is nonzero on Ω, then $\frac{f}{g}$ is also holomorphic on Ω. In particular, every rational function with complex coefficients is holomorphic on its domain.

XII.1.2.9. THEOREM. [CHAIN RULE] Let f be defined on a neighborhood $\mathcal{U}$ of z_0 in $\mathbb{C}$ and complex differentiable at z_0, with $f(z_0) = w_0$, and let g be defined on a neighborhood $\mathcal{V}$ of w_0 in $\mathbb{C}$ and complex differentiable at w_0, with $f(\mathcal{U}) \subseteq \mathcal{V}$. Set $h = g \circ f$ on $\mathcal{U}$. Then h is complex differentiable at z_0, and

$$h'(z_0) = (g \circ f)'(z_0) = g'(w_0)f'(z_0) = g'(f(z_0))f'(z_0) \ .$$

XII.1.2.10. THEOREM. [INVERSE FUNCTION THEOREM] Let $\Omega \subseteq \mathbb{C}$ be open, f holomorphic on Ω, and $z_0 \in \Omega$. If $f'(z_0) \neq 0$, then there is an open neighborhood V of z_0 contained in Ω for which f maps bijectively to an open neighborhood W of $f(z_0)$, and the inverse function f^{-1} is holomorphic on W. If $z \in V$, then

$$(f^{-1})'(f(z)) = \frac{1}{f'(z)} \ .$$

PROOF: There is one aspect of this theorem not yet established: it follows automatically that f' is continuous (in fact, holomorphic) on Ω (XII.3.3.6.). The result now follows from the Inverse Function Theorem, since if f is regarded as a function from Ω to $\mathbb{R}^2$, we have $J_f(z) = det(Df(z)) = |f'(z)|^2$. The inverse function is holomorphic, since if $z \in V$ and $f'(z) = a + ib$, then

$$D(f^{-1})(f(z)) = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}^{-1} = \frac{1}{a^2+b^2} \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$$

and thus f^{-1} satisfies the Cauchy-Riemann equations; then

$$(f^{-1})'(f(z)) = \frac{a}{a^2+b^2} - i\frac{b}{a^2+b^2} = \frac{1}{f'(z)} \ .$$

Although we have stated this result in maximum generality, we have so far only given a complete proof of the following slightly weaker version:

XII.1.2.11. Corollary. Let $\Omega \subseteq \mathbb{C}$ be open, f holomorphic on Ω, and $z_0 \in \Omega$. If f' is continuous on Ω and $f'(z_0) \neq 0$, then there is an open neighborhood V of z_0 contained in Ω which f maps bijectively to an open neighborhood W of $f(z_0)$, and the inverse function f^{-1} is holomorphic on W. If $z \in V$, then

$$(f^{-1})'(f(z)) = \frac{1}{f'(z)} \ .$$

XII.1.2.12. In particular, if f is holomorphic on Ω and f' never vanishes, then f is an open mapping. If Ω is connected, this result extends to the case where f is just assumed to be not constant (). In fact, f is locally an open mapping around any point where it is not locally constant ().

At the other extreme, we have the following, which is an immediate consequence of XII.1.1.4. and VIII.2.1.8.:

XII.1.2.13. Proposition. Let Ω be a connected open set in $\mathbb{C}$, and f a holomorphic function on Ω. If f' is identically zero on Ω, then f is constant on Ω.

XII.1.2.14. Corollary. Let Ω be a connected open set in $\mathbb{C}$, and f and g holomorphic functions on Ω. If $f' = g'$ on Ω, then $f - g$ is constant on Ω.

XII.1.3. Branch Cuts and Branches

XII.1.3.1. A holomorphic function which is locally one-to-one (i.e. whose derivative does not vanish) is not globally one-to-one in general. The fundamental example is the complex exponential function: $e^{z+2\pi i} = e^z$ for any z. If we want to define an inverse function for such a function, we must restrict to regions where the function gives a bijection.

The Complex Logarithm

XII.1.3.2. The prototype problem is how to define a complex logarithm. Since we have $e^{x+iy} = e^x e^{iy}$, we have

$$|e^{x+iy}| = e^x, \quad arg(e^{x+iy}) = y$$

so we may try to define the complex logarithm using polar coordinates by

$$\log(re^{i\theta}) = \log r + i\theta \ .$$

The real part is well defined as long as $r \neq 0$ (obviously $\log 0$ cannot make sense since e^z is never zero); but the problem is that the argument θ is only defined up to adding a multiple of 2π.

XII.1.3.3. The standard solution is to restrict the exponential function to an open horizontal strip of width 2π, say the strip

$$S_\alpha = \{x + iy : \alpha < y < \alpha + 2\pi\}$$

for a fixed real number α. The exponential function maps this open strip one-one onto the plane with a ray out from 0 at angle α (measured in the usual way from the positive real axis counterclockwise) removed. An inverse function can then be defined on this open set Ω_α taking values in the strip, which is a reasonable definition of a complex logarithm function. This function is called a *branch* of the complex logarithm function with the ray as *branch cut*. There is thus one branch of the complex logarithm for each real number α. Note that for a given branch cut, there are infinitely many branches: the branch cuts for the branches corresponding to α and β are the same if (and only if) α and β differ by a multiple of 2π.

A branch of the complex logarithm cannot be continuously extended to any point on the branch cut, since the limits as the point is approached from opposite sides of the cut differ by $2\pi i$.

XII.1.3.4. The branch corresponding to $\alpha = -\pi$ is called the *principal branch* of the complex logarithm. The branch cut for the principal branch is the negative real axis, and the values are in

$$S_{-\pi} = \{x + iy : -\pi < y < \pi\} \ .$$

The principal branch of the complex logarithm agrees with the usual (real) logarithm function on the positive real axis.

XII.1.3.5. Although branch cuts are usually taken to be straight rays, any curve beginning at 0 and going to ∞ without crossing itself can be used as a branch cut, and a branch of the complex logarithm (actually infinitely many) can be defined on the complement. (The preimage of such a curve under the exponential function gives a sequence of parallel curves in the plane a vertical distance 2π apart, and the exponential function is bijective on the region between consecutive curves.)

XII.1.3.6. If f is holomorphic on a region Ω, and $f(\Omega)$ is contained in a region on which a branch ϕ of log is defined, set $g = \phi \circ f$. Then g is holomorphic on Ω and $f = e^g$ on Ω. See XII.3.9.2. for a different situation where such a g exists.

Complex Exponentiation

XII.1.3.7. If $a, b \in \mathbb{C}$ $(a \neq 0)$, how do we define a^b? An unambiguous definition cannot be given in general. But a useful way to do it, which is universally used, is to define

$$a^b = e^{b \log a} \ .$$

This is of course not unambiguous in general since $\log a$ is not well defined – it has infinitely many values. By letting $\log a$ run through all possible values, we get (in general) infinitely many values for a^b.

XII.1.3.8. If b is an integer, all these possible values coincide, and give the usual algebraic definition of a^b. If $b = \frac{p}{q}$ (in lowest terms) is a (real) rational number, there are exactly q different values for a^b. But if b is an irrational real number, or a nonreal number, all the values for a^b are distinct.

These statements are only valid if $a \neq 0$; 0^b cannot be reasonably defined unless b is a natural number (in Complex Analysis we usually don't even define things like $0^{1/2}$).

XII.1.3.9. Because of the ambiguity in definition, care must be exercised in stating and applying rules of exponents; see XII.9.9.8..

General Exponential Functions

XII.1.3.10. To define a general exponential function $f(z) = a^z$, where a is a nonzero complex number, write

$$f(z) = e^{z \log a} .$$

This defines an entire function whenever a value of $\log a$ is chosen, but there are infinitely many choices for any a.

If a is a positive real number (even if $a = 1$ or e!) there are technically infinitely many choices for $\log a$ and thus infinitely many corresponding exponential functions. But in this case it is conventional to choose the real value (i.e. the principal branch) for $\log a$, so there is by convention a well-defined exponential function, which extends the real function $f(x) = a^x$.

General Power Functions

XII.1.3.11. If b is a complex number, we can define $f(z) = z^b$ by choosing a branch of $\log z$ and defining

$$f(z) = e^{b \log z}$$

on $\mathbb{C}$ with the corresponding branch cut removed. In general, this produces infinitely many branches of f for any branch cut, and they cannot be extended to be continuous on the cut unless b is an integer. If $b = \frac{p}{q}$ (in lowest terms) is rational, there are exactly q distinct branches for any branch cut.

XII.1.3.12. If $b = \frac{1}{2}$ (or, more generally, any rational number), this process can be viewed somewhat differently without explicit reference to logarithms. Suppose we want to define $f(z) = z^{1/2} = \sqrt{z}$. The problem is that any nonzero z has two square roots, and there is no systematic (i.e. continuous) way to choose one of them over all of $\mathbb{C}$. We can define

$$f(re^{i\theta}) = \sqrt{r}e^{i\theta/2}$$

but to do this carefully we must specify a range of values for θ of length 2π, i.e. require

$$\alpha < \theta < \alpha + 2\pi$$

for some fixed $\alpha \in \mathbb{R}$. The values of the corresponding square root function lie in the open half plane of points with argument between $\alpha/2$ and $\alpha/2 + \pi$.

If α is increased by 2π, the domain of the corresponding square root function is the same, but the range is the opposite half plane. If α is increased by 4π, the function defined is the same. Thus there are two holomorphic square root functions corresponding to any branch cut.

XII.1.3.13. A square root function (actually two such functions) can be more generally defined on any simply connected () open subset of $\mathbb{C}$ not containing 0. The same can be said for the logarithm function or any power function, although there are infinitely many possibilities for these in general.

XII.1.4. The Fundamental Theorem of Algebra

The Fundamental Theorem of Algebra, that every complex polynomial has a root in $\mathbb{C}$, has many essentially different proofs. There is a nice one due to J. Milnor [Mil97], based on the results of XII.1.1., which we describe here. See XII.9.9.3. and XII.9.9.4. for two more proofs using Complex Analysis.

XII.1.4.1. Lemma. Let p be a nonconstant polynomial with complex coefficients,

$$p(z) = a_0 + a_1 z + a_2 z^2 + \cdots + a_n z^n$$

with $n \geq 1$ and $a_n \neq 0$. Then for any $M > 0$ there is a $K > 0$ such that $|p(z)| > M$ for all z with $|z| > K$.

Proof: For all z with $|z| \geq 1$, we have

$$|p(z)| = |z^n|(|a_0 z^{-n} + \cdots + a_{n-1} z^{-1} + a_n|) \geq |z|^n \left(|a_n| - |z|^{-1} \sum_{k=0}^{n-1} |a_k| \right)$$

If $|z|$ is sufficiently large, the second factor is greater than $\frac{|a_n|}{2}$, and $|z|^n > \frac{2M}{|a_n|}$. So

$$K = \max \left(1, \frac{2\sum_{k=0}^{n-1} |a_k|}{|a_n|}, \left[\frac{2M}{|a_n|} \right]^{1/n} \right)$$

will do. 👍

XII.1.4.2. Theorem. [Fundamental Theorem of Algebra] Let p be a nonconstant polynomial with coefficients in $\mathbb{C}$. Then there is at least one $c \in \mathbb{C}$ with $p(c) = 0$.

Proof: Let $R \subseteq \mathbb{C}$ be the range of p, i.e.

$$R = \{p(z) : z \in \mathbb{C}\}$$

and set

$$F = \{p(z) : z \in \mathbb{C},\ p'(z) = 0\}$$

and $S = R \setminus F$. Then F is a finite subset of R since p' is not identically zero. We will show that R is closed and S is open in $\mathbb{C}$. This will prove the theorem since if $\mathbb{C} \setminus R$ is nonempty, then S and $\mathbb{C} \setminus R$ would be disjoint nonempty open sets whose union is the connected open set $\mathbb{C} \setminus F$, a contradiction. (Note that S is nonempty since otherwise $R = F$, and F would have to be a singleton by continuity, so p would be constant.) Thus $R = \mathbb{C}$, and in particular $0 \in R$.

To show that R is closed, it suffices to show that $R \cap \bar{B}_M(0)$ is closed for all $M > 0$. Fix M. There is a K such that $|p(z)| > M$ for all z with $|z| > K$ (Lemma XII.1.4.1.). Thus

$$R \cap \bar{B}_M(0) = p(\bar{B}_K(0)) \cap \bar{B}_M(0)$$

which is compact since p is continuous ().

We now show S is open. Let $w_0 \in S$. Fix $z_0 \in \mathbb{C}$ with $p(z_0) = w_0$, $p'(z_0) \neq 0$. So by XII.1.2.11. there is a neighborhood V of z_0 which p maps homeomorphically onto a neighborhood W of w_0; thus $W \subseteq R$. Since F is finite and $w_0 \notin F$, $W \setminus F$ is a neighborhood of w_0 contained in S. 👍

XII.2. Contour Integration

Contour integration is one of the most fundamental tools of Complex Analysis. Cauchy's Theorem, which roughly says that the contour integral of a holomorphic function around a closed contour is zero, is the key to all the deeper properties of holomorphic functions.

XII.2.1. Integration along a Contour

XII.2.1.1. DEFINITION. Let $p, q \in \mathbb{C}$. A *contour* from p to q is a continuous rectifiable curve () in $\mathbb{C} \cong \mathbb{R}^2$ from p to q.

XII.2.1.2. A contour γ from p to q by definition has a continuous parametrization $\zeta : [a, b] \to \mathbb{C}$ with $\zeta(a) = p$, $\zeta(b) = q$. However, the parametrization is not unique, and is not intrinsic to γ (only the equivalence class of the parametrization is part of the definition). But since γ is rectifiable, it has an intrinsic parametrization by arc length starting at p (). More importantly, γ has an *orientation*, or specified direction of motion, with initial point p and final point q. There is an *opposite contour* $-\gamma$, which is γ traversed from q to p. If $\zeta : [a, b] \to \mathbb{C}$ is a parametrization of γ, then $\tilde{\zeta} : [0, b - a] \to \mathbb{C}$ defined by $\tilde{\zeta}(t) = \zeta(b - t)$ is a parametrization of $-\gamma$. The interval a parametrization is defined on is not important, and can be changed to any other (closed bounded) interval by a linear change of variables.

We will often consider *closed contours*, ones whose initial and final points coincide. A closed contour still has an orientation, and if γ is a simple closed contour, i.e. the image is homeomorphic to a circle, it also makes sense to say how many (net) times γ is traversed in a counterclockwise direction ().

We will almost always concern ourselves only with the special case where γ is piecewise-smooth, i.e. has a piecewise-smooth parametrization (). If γ is piecewise-smooth, parametrization by arc length is a piecewise-smooth parametrization ().

XII.2.1.3. The goal of this section is to show that if γ is a contour in $\mathbb{C}$, and f is a continuous complex-valued function on $\mathbb{C}$, then there is a well-defined complex number called the *(contour) integral of f along γ*, denoted

$$\int_\gamma f = \int_\gamma f(\zeta)\, d\zeta$$

with the usual properties of an integral. The integral does not depend on a parametrization of γ, although it is usually calculated using a parametrization. It does, however, depend on the orientation of γ: we will have

$$\int_{-\gamma} f = -\int_\gamma f \ .$$

Although contour integrals of somewhat more general functions can be defined, we will restrict to the case of continuous functions for simplicity, and because this is the crucial case for applications. In fact, almost all applications involve only contour integrals of holomorphic functions.

Integration on an Interval

Definition of the Contour Integral

Properties of Contour Integrals

The Fundamental Inequality

The next fundamental estimate is an easy consequence of the definition:

XII.2.1.4. PROPOSITION. Let $\gamma : [a, b] \to \mathbb{C}$ be a rectifiable curve, and f a continuous function on γ. If $|f(z)| \leq M$ for all $z \in \gamma$, then

$$\left| \int_\gamma f \right| \leq M \cdot \ell(\gamma) \ .$$

PROOF: Let $\mathcal{P} = \{a = t_0, t_1, \ldots, t_n = b\}$ be a partition of $[a, b]$. Then by the triangle inequality we have

$$\left| \sum_{k=1}^{n} f(\gamma(t_k))(\gamma(t_k) - \gamma(t_{k-1})) \right| \leq \sum_{k=1}^{n} |f(\gamma(t_k))||\gamma(t_k) - \gamma(t_{k-1})| \leq M \sum_{k=1}^{n} |\gamma(t_k) - \gamma(t_{k-1})| \ .$$

As the mesh of $\mathcal{P}$ goes to zero, the left side approaches $\left| \int_\gamma f \right|$ and the right side approaches $M \cdot \ell(\gamma)$. 👍

An important consequence is:

XII.2.1.5. COROLLARY. Let $\gamma : [a, b] \to \mathbb{C}$ be a rectifiable curve, and (f_n) a sequence of continuous functions on γ converging uniformly on γ to a (continuous) function f. Then

$$\int_\gamma f_n \to \int_\gamma f \ .$$

Calculation of Contour Integrals

On piecewise-smooth curves, contour integrals can be computed by an ordinary integral:

XII.2.1.6. THEOREM. If $\gamma : [a, b] \to \mathbb{C}$ is a piecewise-smooth curve in $\mathbb{C}$ and f is a complex-valued function on γ, then

$$\int_\gamma f = \int_a^b f(\gamma(t))\gamma'(t)\, dt \ .$$

XII.2.1.7. EXAMPLE. Here is perhaps the most important example of a contour integral. Let γ be the unit circle in $\mathbb{C}$, parametrized once around counterclockwise. Let $f(z) = \frac{1}{z} = \bar{z}$ for $|z| = 1$. Then

$$\int_\gamma f = \int_0^{2\pi} \frac{1}{e^{it}} ie^{it}\, dt = 2\pi i \ .$$

This integral will appear over and over in many parts of complex analysis, and many of the fundamental theorems and formulas of the subject are based on it.

The Fundamental Theorem of Calculus

The contour integral of a function with an antiderivative is easy to compute:

XII.2.1.8. PROPOSITION. If f is a continuous function on an open set Ω in $\mathbb{C}$ with an antiderivative F on Ω, and $\gamma : [a, b] \to \Omega$ is a piecewise-smooth contour from p to q, then

$$\int_\gamma f = F(q) - F(p) \ .$$

In particular, the contour integral of f along any contour depends only on the endpoints. The contour integral of f around any closed contour is 0.

PROOF: This just looks like the Fundamental Theorem of Calculus, but it does not follow immediately from the real FTC but requires a bit of calculation using the Chain Rule of vector calculus and the Cauchy-Riemann equations. By subdividing $[a, b]$, we may assume γ is smooth. Write

$$F(x + iy) = U(x, y) + iV(x, y)$$

so that

$$f(x + iy) = \frac{\partial U}{\partial x}(x, y) + i\frac{\partial V}{\partial x}(x, y) \ .$$

Write

$$\gamma(t) = \gamma_1(t) + i\gamma_2(t)$$
$$G(t) = U(\gamma_1(t), \gamma_2(t))$$
$$H(t) = V(\gamma_1(t), \gamma_2(t))$$

so that $F(\gamma(t)) = G(t) + iH(t)$. By the vector calculus Chain Rule,

$$G'(t) = \frac{\partial U}{\partial x}(\gamma_1(t), \gamma_2(t))\gamma_1'(t) + \frac{\partial U}{\partial y}(\gamma_1(t), \gamma_2(t))\gamma_2'(t)$$

$$H'(t) = \frac{\partial V}{\partial x}(\gamma_1(t), \gamma_2(t))\gamma_1'(t) + \frac{\partial V}{\partial y}(\gamma_1(t), \gamma_2(t))\gamma_2'(t) \ .$$

Then

$$\int_\gamma f = \int_a^b f(\gamma(t))\gamma'(t)\, dt$$

$$= \int_a^b \left[\frac{\partial U}{\partial x}(\gamma_1(t),\gamma_2(t)) + i\frac{\partial V}{\partial x}(\gamma_1(t),\gamma_2(t))\right][\gamma_1'(t) + i\gamma_2'(t)]\ dt$$

$$= \int_a^b \left[\frac{\partial U}{\partial x}(\gamma_1(t),\gamma_2(t))\gamma_1'(t) - \frac{\partial V}{\partial x}(\gamma_1(t),\gamma_2(t))\gamma_2'(t)\right] dt + i\int_a^b \left[\frac{\partial V}{\partial x}(\gamma_1(t),\gamma_2(t))\gamma_1'(t) + \frac{\partial U}{\partial x}(\gamma_1(t),\gamma_2(t))\gamma_2'(t)\right] dt$$

$$= \int_a^b \left[\frac{\partial U}{\partial x}(\gamma_1(t),\gamma_2(t))\gamma_1'(t) + \frac{\partial U}{\partial y}(\gamma_1(t),\gamma_2(t))\gamma_2'(t)\right] dt + i\int_a^b \left[\frac{\partial V}{\partial x}(\gamma_1(t),\gamma_2(t))\gamma_1'(t) + \frac{\partial V}{\partial y}(\gamma_1(t),\gamma_2(t))\gamma_2'(t)\right] dt$$

$$= \int_a^b G'(t)\,dt + i\int_a^b H'(t)\,dt = (G(b) - G(a)) + i(H(b) - H(a)) = F(q) - F(p)\ .$$



The Complex Mean Value Inequality

The next result is a complex version of the Mean Value Inequality (V.8.2.10.):

XII.2.1.9. PROPOSITION. Let f be a holomorphic function on an open set Ω in $\mathbb{C}$ (with f' continuous), and A a convex subset of Ω. If $|f'(z)| \le M$ for all $z \in A$, then

$$|f(z) - f(w)| \le M|z - w|$$

for all $z, w \in A$.

The requirement that f' be continuous turns out to be automatic (XII.3.3.6.).
PROOF: Let γ be the straight line segment from w to z; then $\gamma \subseteq A$ since A is convex.

$$|f(z) - f(w)| = \left|\int_\gamma f'\right| \le M\ell(\gamma) = M|z - w|\ .$$



Relation with Line Integrals

Green's Theorem for Complex Functions

XII.3. Some Big Theorems

In this section, we use contour integration to prove some of the big theorems of elementary Complex Analysis: Cauchy's Theorem, the Cauchy Integral Formula, Morera's Theorem, Cauchy's Inequalities, and Liouville's Theorem.

These are, of course, not all the theorems of Complex Analysis which can be called "big theorems." A few more more big theorems concerning power series representation and analysis of singularities will be discussed in the next sections. Other big theorems such as the Riemann Mapping Theorem (and many others) are discussed later or are beyond the scope of this book.

XII.3.1. Goursat's Theorem

We begin by proving a special case of Cauchy's Theorem, due to E. GOURSAT in 1883. This result is a considerable technical improvement over CAUCHY'S original theorem, since it assumes nothing about the properties of the derivative of a holomorphic function, and in particular makes no continuity assumption about the derivative (in fact, it will turn out later that the derivative of a holomorphic function is not only automatically continuous, but even itself holomorphic, but we do not want to assume anything about the derivative in developing the theory).

The form of Goursat's Theorem we give (originally due to A. PRINGSHEIM in 1901) uses triangles. It can be equally well stated and proved for rectangles, or for any other polygonal region that can be subdivided into a finite number of congruent pieces of smaller diameter and perimeter (GOURSAT used squares). The use of triangles gives a more flexible result which simplifies succeeding proofs.

XII.3.1.1. THEOREM. [GOURSAT'S THEOREM] Let Ω be an open set in $\mathbb{C}$, and f a holomorphic function on Ω. If T is a (solid) triangle contained in Ω, and γ is the boundary of T parametrized once around counterclockwise, then

$$\int_\gamma f = 0 \ .$$

PROOF: Let d and p be the diameter and perimeter of T respectively. Divide T into four subtriangles $T_1^{(1)}, \ldots, T_1^{(4)}$ whose vertices are the vertices of T and the midpoints of the edges of T. Then each $T_1^{(k)}$ is similar to T. If $\gamma_1^{(k)}$ is the perimeter of $T_1^{(k)}$, parametrized once around counterclockwise, then we have

$$\int_\gamma f = \int_{\gamma_1^{(1)}} f + \int_{\gamma_1^{(2)}} f + \int_{\gamma_1^{(3)}} f + \int_{\gamma_1^{(4)}} f$$

since all interior edges are traversed once in each direction, so those contour integrals cancel. Thus there is a k such that

$$\left| \int_\gamma f \right| \leq 4 \left| \int_{\gamma_1^{(k)}} f \right| \ .$$

Set $T_1 = T_1^{(k)}$ and $\gamma_1 = \gamma_1^{(k)}$ for this k. Repeat the process inductively using T_n in place of T to obtain a decreasing sequence (T_n) of triangles with boundary γ_n, with

$$\left| \int_\gamma f \right| \leq 4^n \left| \int_{\gamma_n} f \right|$$

for all n. We have that the diameter d_n and perimeter p_n of T_n satisfy $d_n = 2^{-n} d$ and $p_n = 2^{-n} p$. We have that

$$\cap_n T_n = \{z_0\}$$

for some $z_0 \in T$.

Since f is differentiable at z_0, for $z \in \Omega \setminus \{z_0\}$ we may set

$$g(z) = \frac{f(z) - f(z_0)}{z - z_0} - f'(z_0) \ .$$

If we set $g(z_0) = 0$, then g is continuous on Ω and we may write

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + g(z)(z - z_0)$$

for $z \in \Omega$. Thus we have

$$\int_{\gamma_n} f(\zeta)\, d\zeta = \int_{\gamma_n} f(z_0)\, d\zeta + \int_{\gamma_n} f'(z_0)(\zeta - z_0)\, d\zeta + \int_{\gamma_n} g(\zeta)(\zeta - z_0)\, d\zeta$$

for any n. The first two integrals on the right are always zero since the integrands have antiderivatives on Ω (they are a constant function and a linear function respectively); thus for any n we have

$$\int_{\gamma_n} f(\zeta)\, d\zeta = \int_{\gamma_n} g(\zeta)(\zeta - z_0)\, d\zeta \ .$$

Let $\epsilon > 0$. Choose $\delta > 0$ such that $|g(z)| < \epsilon$ whenever $|z - z_0| < \delta$. Then for any n large enough that $d_n < \delta$, we have

$$\left| \int_{\gamma_n} f(\zeta)\, d\zeta \right| = \left| \int_{\gamma_n} g(\zeta)(\zeta - z_0)\, d\zeta \right| \leq \epsilon d_n p_n \ .$$

Combining the inequalities, for n sufficiently large we have

$$\left| \int_{\gamma} f \right| \leq 4^n \left| \int_{\gamma_n} f \right| \leq 4^n \epsilon d_n p_n = \epsilon d p \ .$$

Since $\epsilon > 0$ is arbitrary, we obtain $\int_\gamma f = 0$. 👍

It will be convenient to have a technical strengthening of this theorem:

XII.3.1.2. COROLLARY. Let Ω be an open set in $\mathbb{C}$, $\{z_1, \ldots, z_n\}$ a finite set of points in Ω, and f a holomorphic function on $\Omega \setminus \{z_1, \ldots, z_n\}$. Suppose

$$\lim_{z \to z_k} f(z)(z - z_k) = 0$$

for $1 \leq k \leq n$. If T is a (solid) triangle contained in Ω, and the boundary γ of T (parametrized once around counterclockwise) contains no z_k, then

$$\int_\gamma f = 0 \ .$$

PROOF: T can be subdivided into smaller triangles, with some of the subtriangles $T_1, \ldots, T_m$ each containing exactly one of the z_k in its interior, and the rest of the subtriangles contained in $\Omega \setminus \{z_1, \ldots, z_n\}$. Let γ_j be the boundary of T_j $(1 \leq j \leq m)$. Then

$$\int_\gamma f = \sum_{j=1}^{n} \int_{\gamma_j} f$$

since the integrals around the other triangles are 0 and all edges not part of γ or one of the γ_j are traversed once in each direction. Thus it suffices to show that $\int_{\gamma_j} f = 0$ for each j, i.e. it suffices to show the result in the case where T contains exactly one z_k in its interior.

Changing notation, we assume z_j is the only z_k in the interior of T. By a similar subdivision argument, the integral of f around any other triangle in Ω containing only z_j in its interior is the same. Thus we may assume T has arbitrarily small diameter d and circumference p.

Let $\epsilon > 0$, and choose $\delta > 0$ such that $|f(z)(z - z_j)| < \epsilon$ whenever $0 < |z - z_j| < \delta$. Choose T with $p < \delta$ (and hence $d < \delta$ also). If $\zeta \in \gamma$, we then have $|f(\zeta)| < \frac{\epsilon}{\delta}$. We then have

$$\left| \int_\gamma f \right| \leq \frac{\epsilon}{\delta} p < \epsilon \ .$$

Since the integral is independent of which T is chosen, and $\epsilon > 0$ is arbitrary, we have $\int_\gamma f = 0$. 👍

XII.3.2. The Antiderivative Theorem and Cauchy's Theorem for Convex Sets

XII.3.2.1. THEOREM. [ANTIDERIVATIVE THEOREM FOR CONVEX SETS] Let Ω be a convex open set in $\mathbb{C}$. Let f be a complex-valued continuous function on Ω. If

$$\int_\gamma f = 0$$

for every triangular contour γ in Ω, then f has an antiderivative on Ω: there is a holomorphic function F on Ω with $F' = f$ on Ω.

PROOF: Fix $z_0 \in \Omega$. For any $z \in \Omega$, define

$$F(z) = \int_\gamma f$$

where γ is the straight line segment from z_0 to z. If z and w are in Ω, then

$$F(w) - F(z) = \int_\gamma f$$

where γ is the straight line segment from z to w, since the integral of f around the triangle with vertices z_0, z, and w is 0 by assumption.

Now fix $z \in \Omega$, and fix $\eta > 0$ such that $B_\eta(z) \subseteq \Omega$. If $h \in \mathbb{C}$, $|h| < \eta$, then

$$\frac{F(z+h) - F(z)}{h} = \frac{1}{h} \int_{\gamma_h} f$$

where γ_h is the straight line segment from z to $z + h$. We have

$$\int_{\gamma_h} f(\zeta)\, d\zeta = \int_{\gamma_h} [f(\zeta) - f(z)]\, d\zeta + \int_{\gamma_h} f(z)\, d\zeta \ .$$

The second integral is the contour integral of the constant function $f(z)$ (z is fixed), hence equals $f(z)h$. Since f is continuous at z, for any $\epsilon > 0$ there is a $\delta > 0$, $\delta \leq \eta$, such that $|f(\zeta) - f(z)| < \epsilon$ whenever $|\zeta - z| < \delta$. Thus if $|h| < \delta$, we have

$$\left| \int_{\gamma_h} [f(\zeta) - f(z)]\, d\zeta \right| < \epsilon |h| \ .$$

Thus

$$\lim_{h\to 0} \frac{F(z+h) - F(z)}{h} = \lim_{h\to 0} \frac{1}{h} \int_{\gamma_h} f = \lim_{h\to 0} \frac{1}{h} \int_{\gamma_h} [f(\zeta) - f(z)]\, d\zeta + \lim_{h\to 0} \frac{1}{h} \int_{\gamma_h} f(z)\, d\zeta = f(z)$$

since the first limit on the right is 0 and the second expression is exactly $f(z)$ for all h. Thus $F'(z)$ exists and equals $f(z)$ for any $z \in \Omega$. 👍

The proof actually shows the following slightly more general version:

XII.3.2.2. THEOREM. Let Ω be a convex open set in $\mathbb{C}$, $\{z_1, \ldots, z_n\}$ a finite set of points in Ω, and f a continuous function on $\Omega \setminus \{z_1, \ldots, z_n\}$. If

$$\int_\gamma f = 0$$

for every triangular contour γ in Ω not passing through one of the z_k, then f has an antiderivative on $\Omega \setminus \{z_1, \ldots, z_n\}$: there is a holomorphic function F on $\Omega \setminus \{z_1, \ldots, z_n\}$ with $F' = f$ on $\Omega \setminus \{z_1, \ldots, z_n\}$.

PROOF: There are only a few slight modifications needed to the proof. Take $z_0 \in \Omega \setminus \{z_1, \ldots, z_n\}$, and for $z \in \Omega \setminus \{z_1, \ldots, z_n\}$ we take

$$F(z) = \int_\gamma f$$

where γ is a path from z_0 to z consisting of at most two line segments in $\Omega \setminus \{z_1, \ldots, z_n\}$ (by hypothesis, F is well defined; such angled paths will sometimes be necessary since the straight line segment from z_0 to z may pass through a z_k). If $z \in \Omega \setminus \{z_1, \ldots, z_n\}$, choose $\eta > 0$ such that $B_\eta(z) \subseteq \Omega \setminus \{z_1, \ldots, z_n\}$. The rest of the proof is identical. 👍

There is a version for general open sets, whose proof is essentially identical:

XII.3.2.3. THEOREM. [ANTIDERIVATIVE THEOREM] Let Ω be an open set in $\mathbb{C}$. Let f be a complex-valued continuous function on Ω. If

$$\int_\gamma f = 0$$

for every closed contour γ in Ω, then f has an antiderivative on Ω: there is a holomorphic function F on Ω with $F' = f$ on Ω.

PROOF: We may assume Ω is connected since F may be defined separately on each component. Fix $z_0 \in \Omega$, and for $z \in \Omega$ define

$$F(z) = \int_\gamma f$$

where γ is any contour in Ω from z_0 to z. If γ_1 and γ_2 are two such contours, then

$$\int_{\gamma_1} f - \int_{\gamma_2} f = \int_{\gamma_1 - \gamma_2} f = 0$$

by assumption, so F is well defined. If z and w lie in a convex open subset of Ω (e.g. a small disk), then

$$F(w) - F(z) = \int_{\gamma} f$$

where γ is the straight line segment from z to w, since $F(w)$ can be calculated by taking a contour from z_0 to z and appending this line segment.

The rest of the proof is identical to the proof of XII.3.2.1.. 👍

The advantage of XII.3.2.1. on convex sets is that only triangular contours need to be checked. Putting XII.3.2.1. together with Goursat's Theorem and (), we obtain:

XII.3.2.4. Corollary. [Cauchy's Theorem for Convex Sets] Let Ω be a convex open set in $\mathbb{C}$, and f a holomorphic function on Ω. Then f has an antiderivative on Ω, and if γ is any piecewise-smooth closed curve in Ω, then

$$\int_{\gamma} f = 0 \ .$$

XII.3.2.5. Example. This result can fail if Ω is not convex. For example, let $\Omega = \mathbb{C} \setminus \{0\}$, and $f(z) = \frac{1}{z}$. Then f is holomorphic on Ω, but the integral of f around the unit circle is nonzero () and f does not have an antiderivative on Ω (i.e. $\log z$ cannot be defined as a holomorphic function on all of Ω).

Cauchy's Theorem is true if Ω is simply connected ().

We do get a slight extension by using XII.3.2.2. together with XII.3.1.2.:

XII.3.2.6. Theorem. Let Ω be a convex open set in $\mathbb{C}$, $\{z_1, \ldots, z_n\}$ a finite set of points in Ω, and f a holomorphic function on $\Omega \setminus \{z_1, \ldots, z_n\}$. Suppose

$$\lim_{z \to z_k} f(z)(z - z_k) = 0$$

for $1 \leq k \leq n$. Then f has an antiderivative on $\Omega \setminus \{z_1, \ldots, z_n\}$, and if γ is any piecewise-smooth closed curve in $\Omega \setminus \{z_1, \ldots, z_n\}$, then

$$\int_{\gamma} f = 0 \ .$$

The next result can be generalized using the general form of Cauchy's Theorem, but this form is already very useful. Circles can be replaced by convex or, more generally, starlike polygonal arcs such as triangles or rectangles with the same proof (which even simplifies in this case).

XII.3.2.7. THEOREM. Let f be holomorphic on an open set Ω, and γ_1 and γ_2 circles in Ω (not necessarily concentric), parametrized once around counterclockwise, with γ_2 contained in the interior of γ_1. If the region between γ_1 and γ_2 is contained in Ω, then

$$\int_{\gamma_1} f = \int_{\gamma_2} f \ .$$

Note that the integrals around γ_1 and γ_2 are not necessarily 0!

PROOF: The inner circle γ_2 can be subdivided into arcs with consecutive endpoints $z_1, \ldots, z_n$ such that the line segments between pairs of consecutive points, and the region between the segment and the circle, are contained in Ω (cover the circle with finitely many open disks contained in Ω). Thus, if radial line segments are drawn outward from each z_k to γ_1, each wedge is contained in a convex open subset of Ω, and hence the contour integral of f around each, once around counterclockwise, is 0:

The sum of these integrals is

$$\int_{\gamma_1} f - \int_{\gamma_2} f$$

since each radial segment is traversed once in each direction, so those contour integrals cancel. Thus

$$\int_{\gamma_1} f - \int_{\gamma_2} f = 0 \ .$$



XII.3.3. The Cauchy Integral Formula

XII.3.3.1. A remarkable consequence of the theorems obtained so far is that if a function f is holomorphic on and inside a circle (or more general simple closed curve) γ, then the values of f on γ completely determine the values of f at all points inside γ (see XII.5.1.4. for a generalization), and in fact there is a simple formula involving a contour integral.

XII.3.3.2. THEOREM. [CAUCHY INTEGRAL FORMULA] Let f be holomorphic on an open set Ω in $\mathbb{C}$. Then for each $z \in \Omega$ we have

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta$$

where γ is any circle in Ω bounding an open disk contained in Ω and containing z, parametrized once around counterclockwise.

PROOF: Fix z inside γ. If $\delta > 0$ is small enough that the circle γ_δ of radius δ centered at z is in the interior of γ, we have

$$\int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta = \int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta$$

by XII.3.2.7.. Thus it suffices to prove the result for the γ_δ, in fact to show that

$$f(z) = \lim_{\delta \to 0} \frac{1}{2\pi i} \int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta \ .$$

Let $\epsilon > 0$. Since f is continuous, if δ is sufficiently small then $|f(\zeta) - f(z)| < \epsilon$ for $\zeta \in \gamma_\delta$. We then have, for γ_δ inside γ,

$$\int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta = \int_{\gamma_\delta} \frac{f(\zeta) - f(z)}{\zeta - z}\, d\zeta + \int_{\gamma_\delta} \frac{f(z)}{\zeta - z}\, d\zeta \ .$$

If δ is small enough that $|f(\zeta) - f(z)| < \epsilon$ for $\zeta \in \gamma_\delta$, the first integral satisfies

$$\left| \int_{\gamma_\delta} \frac{f(\zeta) - f(z)}{\zeta - z}\, d\zeta \right| \leq \frac{\epsilon}{\delta} \cdot 2\pi\delta = 2\pi\epsilon$$

and thus the first integral approaches 0 as $\delta \to 0$. The second integral can be computed by taking $\zeta(t) = z + \delta e^{it}$, $0 \leq t \leq 2\pi$; thus

$$\int_{\gamma_\delta} \frac{f(z)}{\zeta - z}\, d\zeta = f(z) \int_0^{2\pi} \frac{1}{\delta e^{it}} i\delta e^{it}\, dt = 2\pi i f(z) \ .$$



An interesting reformulation is that if D is a closed disk and f is holomorphic on and inside D, then the value of f at the center of the disk is the average of the values on the boundary circle:

XII.3.3.3. Corollary. Let f be holomorphic on an open set Ω, and D a closed disk in Ω of radius r, centered at z_0. Then

$$f(z_0) = \frac{1}{2\pi} \int_0^{2\pi} f(z_0 + re^{it})\, dt \ .$$

Proof: Parametrize the boundary circle by $\zeta(t) = z_0 + re^{it}$, $0 \leq t \leq 2\pi$. Then

$$f(z_0) = \frac{1}{2\pi i} \int_0^{2\pi} \frac{f(z_0 + re^{it})}{re^{it}} ire^{it}\, dt \ .$$



The Cauchy Integral Formula for Derivatives

XII.3.3.4. We now obtain another remarkable consequence of the Cauchy Integral Formula. The integrand in the formula is a function of z and ζ; since we integrate with respect to ζ, it should be possible to differentiate with respect to z under the integral sign, in analogy with IX.3.5.1., to obtain an integral formula for f'. In fact, since the integrand is infinitely differentiable in z, we could expect that the function defined by the integral, namely f, is also infinitely differentiable and all its derivatives are obtained by successive differentiation under the integral sign. This expectation is valid (XII.3.3.6.). In fact, a more general result along the same lines holds, which is also useful for other purposes:

XII.3.3.5. THEOREM. Let γ be a contour in $\mathbb{C}$ (not necessarily a closed contour), and ϕ a complex-valued continuous function on γ. For $z \in \mathbb{C} \setminus \gamma$, define

$$f(z) = \int_\gamma \frac{\phi(\zeta)}{\zeta - z}\, d\zeta \ .$$

Then f is infinitely differentiable on $\mathbb{C} \setminus \gamma$, and for each $z \in \mathbb{C} \setminus \gamma$ and each n we have

$$f^{(n)}(z) = n! \int_\gamma \frac{\phi(\zeta)}{(\zeta - z)^{n+1}}\, d\zeta \ .$$

PROOF: By induction on n. The case $n = 0$ holds by assumption. Assume the result is true for $n-1$, i.e. for any $z \in \mathbb{C} \setminus \gamma$,

$$f^{(n-1)}(z) = (n-1)! \int_\gamma \frac{\phi(\zeta)}{(\zeta - z)^n}\, d\zeta \ .$$

Fix $z \in \mathbb{C} \setminus \gamma$, and fix $\delta > 0$ such that the closed disk of radius δ around z is contained in $\mathbb{C} \setminus \gamma$. If $|h| < \delta$, then

$$\frac{f^{(n-1)}(z+h) - f^{(n-1)}(z)}{h} = (n-1)! \int_\gamma \frac{\phi(\zeta)}{h} \left[\frac{1}{(\zeta - z - h)^n} - \frac{1}{(\zeta - z)^n} \right] d\zeta \ .$$

Set

$$a = \frac{1}{\zeta - z - h} \ , \quad b = \frac{1}{\zeta - z}$$

and recall the formula

$$a^n - b^n = (a-b)(a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1})$$

(the second factor is 1 if $n = 1$). We have

$$a - b = \frac{1}{\zeta - z - h} - \frac{1}{\zeta - z} = \frac{h}{(\zeta - z - h)(\zeta - z)}$$

so we have

$$\frac{f^{(n-1)}(z+h) - f^{(n-1)}(z)}{h} = (n-1)! \int_\gamma \frac{\phi(\zeta)}{(\zeta - z - h)(\zeta - z)} \left[a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1} \right] d\zeta \ .$$

Since the disk of radius δ around z is bounded away from γ, as $h \to 0$ we have that $\frac{1}{\zeta - z - h}$ converges uniformly in ζ to $\frac{1}{\zeta - z}$ on γ (i.e. $a \to b$ uniformly on γ). Since the other factors $\phi(\zeta)$ and $\frac{1}{\zeta - z}$ are uniformly bounded on γ and do not depend on h, as $h \to 0$ we have

$$\frac{\phi(\zeta)}{(\zeta - z - h)(\zeta - z)} \left[a^{n-1} + a^{n-2}b + \cdots + ab^{n-2} + b^{n-1} \right] \to \frac{\phi(\zeta)}{(\zeta - z)^2} \cdot nb^{n-1} = n \frac{f(\zeta)}{(\zeta - z)^{n+1}}$$

uniformly in ζ on γ. Thus the integrals around γ also converge. So

$$f^{(n)}(z) = \lim_{h \to 0} \frac{f^{(n-1)}(z+h) - f^{(n-1)}(z)}{h}$$

exists and equals

$$\lim_{h\to 0}(n-1)!\int_\gamma \frac{\phi(\zeta)}{(\zeta-z-h)(\zeta-z)}\left[a^{n-1}+a^{n-2}b+\cdots+ab^{n-2}+b^{n-1}\right]\,d\zeta = n!\int_\gamma \frac{\phi(\zeta)}{(\zeta-z)^{n+1}}\,d\zeta\ .$$



XII.3.3.6. Corollary. [Cauchy Integral Formula for Derivatives] Let f be holomorphic on an open set Ω in $\mathbb{C}$. Then f is infinitely differentiable on Ω, and for each $z\in\Omega$ and each n we have

$$f^{(n)}(z) = \frac{n!}{2\pi i}\int_\gamma \frac{f(\zeta)}{(\zeta-z)^{n+1}}\,d\zeta$$

where γ is any circle in Ω bounding an open disk D contained in Ω and containing z, parametrized once around counterclockwise.

Proof: For $\zeta\in\gamma$, set $\phi(\zeta)=\frac{f(\zeta)}{2\pi i}$. Then, for z inside γ, by XII.3.3.2.

$$f(z) = \int_\gamma \frac{\phi(\zeta)}{\zeta-z}\,d\zeta\ .$$



XII.3.3.7. Note that in the special case of XII.3.3.6., for $\zeta\in\gamma$ we have that $\phi(\zeta)$ is, up to a factor of $2\pi i$, the limit of $f(z)$ as $z\to\zeta$ from inside the circle. But in the general situation of XII.3.3.5., there is no such simple relationship (for example, in the situation of XII.3.3.6., the function defined by the integral is identically 0 outside γ by ()). In fact, there is no way to recover ϕ from f in general, although there is in many cases ().

Morera's Theorem

Morera's Theorem is a converse to Cauchy's Theorem:

XII.3.3.8. Theorem. [Morera's Theorem] Let Ω be an open set in $\mathbb{C}$, and f a complex-valued continuous function on Ω. If the contour integral of f around the boundary of every solid triangle contained in Ω is zero, then f is holomorphic on Ω.

Proof: It suffices to show that f is holomorphic in every open disk contained in Ω, i.e. we may assume Ω is a disk. Then f has an antiderivative F on Ω by XII.3.2.1.. But F is infinitely differentiable on Ω, hence so is $F'=f$. 

Cauchy's Inequalities

Cauchy's Inequalities are an immediate consequence of the Cauchy Integral Formula and the usual estimate for the size of a contour integral. They say that the higher derivatives of a holomorphic function do not grow too rapidly at any point.

XII.3.3.9. THEOREM. [CAUCHY'S INEQUALITIES] Let f be holomorphic on an open set Ω in $\mathbb{C}$, and $z_0 \in \Omega$. If the closed disk of radius R centered at z_0 is contained in Ω, then for all n

$$\left|f^{(n)}(z_0)\right| \leq \frac{n!}{R^n} \max_{|\zeta - z_0| = R} |f(\zeta)| \ .$$

Note that the quantity $\max_{|\zeta - z_0| = R} |f(\zeta)|$, the maximum of $|f|$ on the boundary circle, depends on R but not on n.

Liouville's Theorem

An immediate consequence of Cauchy's Inequality for $n = 1$ is the following remarkable result:

XII.3.3.10. THEOREM. [LIOUVILLE'S THEOREM] A bounded entire function is constant.

PROOF: Let f be a bounded entire function, and fix M with $|f(z)| \leq M$ for all $z \in \mathbb{C}$. Fix $z_0 \in \mathbb{C}$. Then for any $R > 0$ we have

$$|f'(z_0)| \leq \frac{M}{R}$$

and thus $f'(z_0) = 0$. Since $z_0 \in \mathbb{C}$ is arbitrary, f' is identically 0. 👍

This can be used, among many other things, to give a proof of the Fundamental Theorem of Algebra (Exercise XII.9.9.3.).

The result can be sharpened:

XII.3.3.11. COROLLARY. Let f be a nonconstant entire function. Then the range of f is dense in $\mathbb{C}$.

PROOF: Suppose the range is not dense. Then there is a $w \in \mathbb{C}$ and an $\epsilon > 0$ such that $|f(z) - w| \geq \epsilon$ for all $z \in \mathbb{C}$. Set

$$g(z) = \frac{1}{f(z) - w} \ .$$

Then g is an entire function, and $|g(z)| \leq \frac{1}{\epsilon}$ for all z. Thus g, and hence f, is constant, a contradiction. 👍

In fact, the following stronger result holds. See e.g. [] for a proof.

XII.3.3.12. THEOREM. [LITTLE PICARD THEOREM] Let f be a nonconstant entire function. Then the range of f is all of $\mathbb{C}$ except possibly one point.

This theorem is the best possible: $f(z) = e^z$ is never 0. See XII.5.2.19. for the "Big Picard Theorem" which is an even stronger statement.

XII.3.4. The Maximum Modulus Theorem

One of the important applications of ordinary calculus is finding maxima and minima of differentiable functions. But a striking difference between the real and complex cases is that if f is a nonconstant holomorphic function, then $|f|$ cannot have any local maxima:

XII.3.4.1. THEOREM. Let f be a holomorphic function on Ω, and let $\mathbb{D}$ be an open disk centered at z_0 with $\bar{\mathbb{D}} \subseteq \Omega$. Then

$$|f(z_0)| \leq \max\{|f(z)| : z \in \partial\mathbb{D}\}$$

with equality if and only if f is constant on $\mathbb{D}$.

PROOF: Multiplying f by a constant of absolute value 1, we may assume $f(z_0) \geq 0$. Let r be the radius of $\mathbb{D}$ and set $M = \max\{|f(z)| : z \in \partial\mathbb{D}\}$. By () we have

$$f(z_0) = |f(z_0)| = \left|\frac{1}{2\pi}\int_0^{2\pi} f(z_0 + re^{it})\,dt\right| \leq \frac{1}{2\pi}\int_0^{2\pi} |f(z_0 + re^{it})|\,dt \leq \frac{1}{2\pi}\int_0^{2\pi} M\,dt = M\ .$$

Suppose $f(z_0) = M$. Since

$$f(z_0) = Re(f(z_0)) = \int_0^{2\pi} Re(f(z_0 + re^{it}))\,dt$$

we have

$$0 = \int_0^{2\pi} [f(z_0) - Re(f(z_0 + re^{it}))]\,dt \geq \int_0^{2\pi} [f(z_0) - |f(z_0 + re^{it})|]\,dt \geq 0$$

because $f(z_0) - Re(f(z_0 + re^{it})) \geq f(z_0) - |f(z_0 + re^{it})| \geq 0$ for all t. Since the integrands are continuous and nonnegative, and the integrals are 0, the integrands are identically zero, i.e. we have $|f(z)| = Re(f(z)) = f(z_0)$ for all $z \in \partial\mathbb{D}$, which implies that $f(z) = f(z_0)$ for all $z \in \partial\mathbb{D}$. 👍

XII.3.4.2. COROLLARY. Let f be a holomorphic function on Ω, and let $z_0 \in \Omega$. If $|f|$ has a local maximum at z_0, then f is constant on a neighborhood of z_0.

PROOF: Let $\mathbb{D}$ be a disk centered at z_0 contained in Ω, such that $|f(z)| \leq |f(z_0)|$ for all $z \in \mathbb{D}$. Let r be less than the radius of $\mathbb{D}$. By the theorem, f is constant and equal to $f(z_0)$ on the circle of radius r centered at z_0. Since this is true for every such r, f is constant on $\mathbb{D}$. 👍

XII.3.4.3. THEOREM. [MAXIMUM MODULUS THEOREM] Let Ω be a connected open set in $\mathbb{C}$, and f a nonconstant holomorphic function on Ω. Then $|f|$ does not attain a maximum at any point of Ω.

PROOF: Suppose $|f|$ attains a maximum at $z_0 \in \Omega$. Then f is constant in a neighborhood of z_0. Let $U = \{z \in \Omega : f(z) = f(z_0)\}$. Then by the same argument f is constant in a neighborhood of any $z \in U$, i.e. U is an open set. But U is relatively closed in Ω since f is continuous. Since Ω is connected, $U = \Omega$. 👍

XII.3.4.4. Actually, under the hypotheses a stronger conclusion holds: $f(\Omega)$ is an open set ().

The Minimum Modulus Theorem

XII.3.4.5. If f is holomorphic, $|f|$ can have a local (even global) minimum even if it is not constant, since $f(z_0)$ can be 0 for some z_0. But this is the only possibility:

XII.3.4.6. Corollary. [Minimum Modulus Theorem] Let Ω be a connected open set in $\mathbb{C}$, and f a nonconstant holomorphic function on Ω which is never zero on Ω. Then $|f|$ does not attain a minimum at any point of Ω.

Proof: Apply the Maximum Modulus Theorem to $g = \frac{1}{f}$. 👍

XII.3.5. The Homotopy Theorem for Contour Integrals

Cauchy's Theorem roughly states that the contour integral of a holomorphic function along a path γ depends only on the endpoints of γ and not on which path is chosen, or alternatively that the contour integral around a closed contour is 0. But this is not literally true in general (XII.2.1.7.).

To prove, or even state, the general version of Cauchy's Theorem, we will show that if f is a holomorphic function on a region Ω and the path γ is "continuously deformed" within Ω leaving the endpoints fixed, the contour integral along γ does not change. This will imply that the integral around any closed curve in Ω which can be "continuously shrunk" to the initial point within Ω is zero.

The careful formulation of a "continuous deformation" is a homotopy:

XII.3.5.1. Definition. Let Ω be an open set in $\mathbb{C}$, $p, q \in \Omega$. Let γ_0 and γ_1 be contours from p to q in Ω with parametrizations $\zeta_0(t)$, $\zeta_1(t)$ defined on the same interval $[a, b]$. A *homotopy* from γ_0 to γ_1 in Ω is a continuous function $h : [0, 1] \times [a, b]$ to Ω such that $h(0, t) = \zeta_0(t)$ and $h(1, t) = \zeta_1(t)$ for all $t \in [a, b]$, and $h(s, 0) = p$ and $h(s, 1) = q$ for all $s \in [0, 1]$. We say γ_0 and γ_1 are *homotopic* (relative to p and q) if there is a homotopy from γ_0 to γ_1.

Technically, what we have defined is a homotopy between the parametrizations ζ_0 and ζ_1 relative to p and q (), but by slight abuse of terminology we will call it a homotopy from γ_0 to γ_1.

XII.3.5.2. If s is fixed, then $\zeta_s(t) = h(s, t)$ defines a parametrized curve γ_s in Ω from p to q, and as s varies these curves "move continuously" from γ_0 to γ_1, and conversely if "continuous deformation" is defined reasonably, the function $h(s, t) = \zeta_s(t)$ is a homotopy from γ_0 to γ_1. Thus γ_0 and γ_1 are homotopic in Ω if and only if γ_0 can be "continuously deformed" to γ_1 within Ω.

XII.3.5.3. A closed contour γ in an open set $\Omega \subseteq \mathbb{C}$ is *contractible* in Ω if it is homotopic in Ω to a constant contour, i.e. if γ is parametrized by $\zeta : [a, b] \to \mathbb{C}$ with $\zeta(a) = \zeta(b) = p$, there is a continuous $h : [0, 1] \times [a, b] \to \Omega$ with $h(0, t) = \zeta(t)$ and $h(1, t) = h(1, 0)$ for all t and $h(s, 0) = h(s, 1)$ for all s. If γ is contractible, there is such a homotopy with $h(s, 0) = h(s, 1) = h(1, t)$ for all t, i.e. the point p is fixed as the base point of each contour γ_s.

XII.3.5.4. THEOREM. [HOMOTOPY THEOREM FOR CONTOUR INTEGRALS] Let f be holomorphic on an open set Ω, and $p, q \in \Omega$. If γ_0 and γ_1 are homotopic contours in Ω from p to q, then

$$\int_{\gamma_0} f = \int_{\gamma_1} f \ .$$

PROOF: We first give a simple argument which "almost" proves the theorem but which has a serious flaw. (This argument is represented as a proof of the theorem in some Complex Analysis texts.)

Suppose ζ_0 and ζ_1 are parametrizations of γ_0 and γ_1 on the same interval, which we may take to be $[0,1]$ to simplify notation, and $h : [0,1]^2 \to \Omega$ a homotopy from ζ_0 to ζ_1 relative to p, q. Fix partitions $\mathcal{P} = \{0 = s_0, s_1, \ldots, s_m = 1\}$ and $\mathcal{Q} = \{0 = t_0, t_1, \ldots, t_n = 1\}$ of $[0,1]$. For each j and k define the curvilinear "rectangle"

$$R_{jk} = \{h(s,t) : s_{j-1} \le s \le s_j, t_{k-1} \le t \le t_k\}$$

(this curvilinear "rectangle" is a curvilinear "triangle" if $k = 1$ or $k = n$). We have that the image of h is compact, so there is an $\epsilon > 0$ such that the ball of radius ϵ around any point of the image of h is contained in Ω (). Since h is uniformly continuous on $[0,1]^2$, there is a $\delta > 0$ such that $|h(s',t') - h(s,t)| < \epsilon$ whenever $s, s', t, t' \in [0,1]$, $|s - s'| < \delta$, and $|t - t'| < \delta$. Thus if $\mathcal{P}$ and $\mathcal{Q}$ are sufficiently fine, for each j and k the "rectangle" R_{jk} is contained within an open disk in Ω. So if we integrate f around the boundary curve γ_{jk} of R_{jk} obtained by setting $s = s_{j-1}$ and letting t vary from t_{k-1} to t_k, and similarly for the other edges, we have that

$$\int_{\gamma_{jk}} f = 0$$

by XII.3.2.4.. But

$$\sum_{j,k} \int_{\gamma_{jk}} f = \int_{\gamma_1} f - \int_{\gamma_0} f$$

since all the other edges are traversed once each direction, so those contour integrals cancel. ✋

The problem with this argument is that while the γ_{jk} are continuous curves, they are not piecewise-smooth or even rectifiable in general, so $\int_{\gamma_{jk}} f$ is not even defined in general. (In fact, this "curve" does not even have to look anything like an ordinary curve topologically; it is just a continuous image of $[0,1]$.) To make this proof work as written, the homotopy h must be "piecewise-smooth," and it must be shown that homotopic piecewise-smooth contours are actually homotopic by a piecewise-smooth homotopy. This can be shown to be true (the argument below can be developed into a proof).

We will proceed differently, avoiding the difficulty by a trick (see [**?**, p. 338] for a different way around the problem). Choose the partitions and define the R_{jk} as above. However, replace γ_{jk} by the curve $\tilde{\gamma}_{jk}$ consisting of the straight line segments from $h(s_{j-1}, t_{k-1})$ to $h(s_j, t_{k-1})$, then to $h(s_j, t_k)$, then to $h(s_{j-1}, t_k)$, then back to $h(s_{j-1}, t_{k-1})$ (some of these line segments may be degenerate). If $j = 1$ or $j = m$, one of the edge curves of R_{jk} is a piece of γ_0 or γ_1, and we use this piece in $\tilde{\gamma}_{jk}$ instead of the straight line segment. Note that since R_{jk} is contained in a disk in Ω, which is convex, these line segments are also contained in the same disk. Then for each j and k, $\tilde{\gamma}_{jk}$ is a piecewise-smooth closed contour contained in a disk inside Ω. We thus have

$$\int_{\tilde{\gamma}_{jk}} f = 0$$

by XII.3.2.4.. But

$$\sum_{j,k} \int_{\tilde{\gamma}_{jk}} f = \int_{\gamma_1} f - \int_{\gamma_0} f$$

since all the other edges are traversed once each direction, so those contour integrals cancel. 👍

The Winding Number of a Closed Curve

If γ is a closed contour and z is a point of $\mathbb{C}$ not on γ, then intuitively γ "winds around z" some net number of times in the counterclockwise sense (this number can be zero or negative as well as positive). If γ is a circle parametrized once around counterclockwise, this winding number should be 1 if z is inside γ and 0 if z is outside. We can make this notion precise:

XII.3.5.5. PROPOSITION. Let γ be a closed piecewise-smooth contour in $\mathbb{C}$, and z a point of $\mathbb{C}$ not on γ. Then

$$\int_\gamma \frac{1}{\zeta - z}\, d\zeta$$

is an integer multiple of $2\pi i$.

PROOF: [Ahl78] Let $\zeta : [a, b] \to \mathbb{C}$ be a piecewise-smooth parametrization of γ with $\zeta(a) = \zeta(b) = p$. Then

$$\int_\gamma \frac{1}{\zeta - z}\, d\zeta = \int_a^b \frac{\zeta'(t)}{\zeta(t) - z}\, dt\ .$$

For $a \le x \le b$, set

$$\phi(x) = \int_a^x \frac{\zeta'(t)}{\zeta(t) - z}\, dt\ .$$

Then ϕ is differentiable at t and

$$\phi'(t) = \frac{\zeta'(t)}{\zeta(t) - z}$$

whenever $a \le t \le b$ and ζ' is continuous at t, i.e. for all but possibly finitely many $t \in [a, b]$. The function

$$\psi(t) = e^{-\phi(t)}(\zeta(t) - z)$$

is continuous on $[a, b]$, and satisfies

$$\psi'(t) = -e^{-\phi(t)}\phi'(t)(\zeta(t) - z) + e^{-\phi(t)}\zeta'(t) = 0$$

wherever ζ' is continuous. Thus ψ is constant on $[a, b]$. We have $\psi(a) = p - z$, so

$$e^{\phi(t)} = \frac{\zeta(t) - z}{p - z}$$

for $a \le t \le b$. We have $e^{\phi(b)} = 1$, so $\phi(b) = 2\pi i n$ for some $n \in \mathbb{Z}$. 👍

XII.3.5.6. DEFINITION. Let γ be a closed piecewise-smooth contour in $\mathbb{C}$, and $z \in \mathbb{C} \setminus \gamma$. The *winding number* of γ around z is

$$w(\gamma, z) = \frac{1}{2\pi i} \int_\gamma \frac{1}{\zeta - z}\, d\zeta \ .$$

By XII.3.5.5., $w(\gamma, z)$ is an integer.

XII.3.5.7. PROPOSITION. Fix a closed piecewise-smooth contour γ. The function $\omega(z) = w(\gamma, z)$ is constant on connected components of $\mathbb{C} \setminus \gamma$ and is zero on the unbounded component.

PROOF: The function ω is clearly a continuous function of z, and

$$\lim_{z \to \infty} |\omega(z)| = 0 \ .$$

Since ω takes only integer values, the result follows. 👍

XII.3.5.8. There is another interpretation of $w(\gamma, z)$. Let $\mathbb{T}$ be the unit circle in $\mathbb{C}$. Given a parametrization ζ of γ, which we may assume is on $[0, 1]$, define

$$\tilde{\gamma}(t) = \frac{\zeta(t) - z}{|\zeta(t) - z|}$$

(this closed contour is essentially the radial projection of γ onto a circle centered at z, translated to 0). Then $\tilde{\gamma}$ is a loop in $\mathbb{T}$ as defined in (), and $w(\gamma, z)$ is precisely the index of this loop.

XII.3.5.9. If γ_0 and γ_1 are closed contours not passing through z, and γ_0 and γ_1 are homotopic in $\mathbb{C} \setminus \{z\}$ (not necessarily with endpoint fixed), then the corresponding loops $\tilde{\gamma}_0$ and $\tilde{\gamma}_1$ ar homotopic as loops in $\mathbb{T}$. Thus their indices are the same (). So we obtain:

XII.3.5.10. PROPOSITION. If γ_0 and γ_1 are closed contours not passing through z, and γ_0 and γ_1 are homotopic in $\mathbb{C} \setminus \{z\}$ (not necessarily with endpoint fixed), then

$$w(\gamma_0, z) = w(\gamma_1, z) \ .$$

We may restate the Cauchy Integral Formula:

XII.3.5.11. THEOREM. Let f be holomorphic on a convex open set Ω, γ a closed piecewise-smooth contour in Ω, and $z \in \Omega$ not on γ. Then

$$f(z) w(\gamma, z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta \ .$$

XII.3.6. Homology of Contours

There is another useful relation on closed contours called *homology*, which is closely related to homotopy (and in fact slightly more general). Although we will primarily use this notion for closed contours, it is convenient to define it more generally for cycles:

XII.3.6.1. Definition. Let Ω be an open set in $\mathbb{C}$. Two cycles γ_0 and γ_1 in Ω are *homologous in* Ω if $w(\gamma_0, z) = w(\gamma_1, z)$ for all $z \in \mathbb{C} \setminus \Omega$. A cycle γ is *homologous to zero in* Ω if $w(\gamma, z) = 0$ for all $z \in \mathbb{C} \setminus \Omega$.

XII.3.6.2. The notion of homology of cycles depends on Ω. Indeed, every cycle is homologous to zero in $\mathbb{C}$. Cycles γ_0 and γ_1 in Ω are homologous in Ω if and only if $\gamma_1 - \gamma_0$ is homologous to zero in Ω. If γ_0 is homologous to zero in Ω, so is $-\gamma_0$; and $\gamma_0 + \gamma_1$ is homologous to γ_1 in Ω for any cycle γ_1 in Ω. If γ_1, $\tilde{\gamma}_1$, γ_2, and $\tilde{\gamma}_2$ are cycles in Ω, and γ_j and $\tilde{\gamma}_j$ are homologous in Ω for $j = 1, 2$, then $\gamma_1 + \gamma_2$ and $\tilde{\gamma}_1 + \tilde{\gamma}_2$ are homologous in Ω.

XII.3.6.3. If two closed contours γ_0 and γ_1 in Ω are homotopic in Ω, then it follows immediately from XII.3.5.10. that γ_0 and γ_1 are homologous in Ω. In particular, if γ is contractible in Ω, then γ is homologous to zero in Ω. The following example shows that the converse is false in general:

XII.3.6.4. Example. Let $\Omega = \mathbb{C} \setminus \{1, -1\}$. Let C_1 be the circle of radius 1 centered at 1, and C_{-1} the circle of radius 1 centered at -1. Let γ_1 be the circle C_1 parametrized once around counterclockwise starting at 0, and γ_2 the circle C_{-1} parametrized once around counterclockwise starting at 0. Set

$$\gamma = \gamma_1 + \gamma_2 - \gamma_1 - \gamma_2 \ .$$

Then γ is a piecewise-smooth closed contour in Ω, and it is easily checked that γ is homologous to zero in Ω. However, γ is not contractible in Ω; this fact, while fairly evident intuitively, is not so easy to prove rigorously (see [GK06, p. 356-358] for a nice proof).

This example shows that addition of closed contours is not commutative up to homotopy. In fact, the closed contours $\gamma_1 + \gamma_2$ and $\gamma_2 + \gamma_1$ are not homotopic in Ω.

The underlying reason for this example is that the fundamental group $\pi_1(\Omega)$ () is not abelian. In fact, it is a free group generated by the classes $[\gamma_1]$ and $[\gamma_2]$.

XII.3.7. Cauchy's Theorem, General Versions

The first general version of Cauchy's Theorem is an immediate corollary of XII.3.5.4.:

XII.3.7.1. Corollary. [Cauchy's Theorem, Version 1] Let f be holomorphic on an open set Ω, and γ a closed contour in Ω which is contractible in Ω. Then

$$\int_\gamma f = 0 \ .$$

Using the Jordan-Schönflies Theorem (), it can be shown that if γ is a simple closed contour in Ω whose interior is also in Ω, then γ is contractible in Ω. Thus we obtain another standard version of Cauchy's Theorem:

XII.3.7.2. Corollary. [Cauchy's Theorem, Version 2] Let f be holomorphic on an open set Ω, and γ a simple closed contour in Ω whose interior is also in Ω. Then

$$\int_\gamma f = 0 \ .$$

XII.3.7.3. Definition. An open set Ω in $\mathbb{C}$ is *simply connected* if it is connected and every closed contour in Ω is contractible in Ω.

The same definition works (path-connectedness is required in general, not just connectedness), and is important, in a general topological space ().

XII.3.7.4. Informally, a connected open set is simply connected if it has no "holes." In particular, any convex or star-shaped open set is simply connected, as is $\mathbb{C}$ with a branch cut removed (this set is actually star-shaped). But a plane or open disk with a point removed, or an annulus, is not simply connected. Any set homeomorphic to a simply connected open set is simply connected (and open).

For open sets in the plane, there are many equivalent characterizations of simple connectedness, such as the following (there are many more):

XII.3.7.5. Theorem. Let Ω be a connected open set in the plane. The following are equivalent:

(i) Ω is simply connected.

(ii) Whenever γ is a simple closed contour in Ω, the interior of γ is also contained in Ω.

(iii) Whenever γ is a closed contour in Ω, the winding number () of γ around any point of $\mathbb{C} \setminus \Omega$ is 0.

If Ω is bounded, these are also equivalent to

(iv) The boundary $\partial\Omega$ is connected.

(v) The complement $\mathbb{C} \setminus \Omega$ is connected.

Condition (iii) may be rephrased:

(iii′) Every cycle in Ω is homologous to zero in Ω.

XII.3.7.6. If Ω is unbounded, (iv) and (v) are still equivalent and imply (i)–(iii), but not conversely in general: consider a horizontal strip in $\mathbb{C}$. (If we work in the Riemann Sphere, the one-point compactification of $\mathbb{C}$, (i)–(v) are equivalent even if Ω is unbounded.) Also, (ii)–(v) can hold if Ω is not connected: consider a figure 8. It is not true that if Ω is simply connected, then the interior of the complement of Ω is necessarily connected, e.g. if Ω is one of the Lakes of Wada (). This example shows that simply connected open sets in $\mathbb{C}$ can be quite complicated. However, it is a (deep) theorem that every nonempty simply connected open set in $\mathbb{C}$ is homeomorphic to a disk (or, equivalently, to $\mathbb{C}$ itself). In fact, much more is true: every nonempty simply connected open proper subset of $\mathbb{C}$ is conformally equivalent to the open unit disk by the Riemann Mapping Theorem ().

We then get the most commonly stated version of Cauchy's Theorem:

XII.3.7.7. Corollary. [Cauchy's Theorem, Simply Connected Version] Let f be holomorphic on a simply connected open set Ω. Then f has an antiderivative on Ω, and if γ is any closed contour in Ω, then

$$\int_\gamma f = 0 \ .$$

The most general version of Cauchy's Theorem which can be reasonably stated is:

XII.3.7.8. Theorem. [Cauchy's Theorem, Homology Version] Let f be holomorphic on an open set Ω. Then

$$\int_\gamma f = 0$$

for every cycle γ in Ω homologous to zero in Ω.

This version is somewhat complicated to prove rigorously. See e.g. [Ahl78] or [**?**].

We get as a special case a general version of XII.3.2.7.:

XII.3.7.9. Corollary. Let f be holomorphic on an open set Ω, and γ_1 and γ_2 simple closed contours in Ω, parametrized once around counterclockwise, with γ_2 contained in the interior of γ_1. If the region between γ_1 and γ_2 is contained in Ω, then

$$\int_{\gamma_1} f = \int_{\gamma_2} f \ .$$

Proof: It is easily checked that $\gamma_1 - \gamma_2$ is homologous to zero in Ω. 👍

Note that the integrals around γ_1 and γ_2 are not necessarily 0!

XII.3.8. The Cauchy Integral Formula, General Versions

XII.3.9. Exercises

XII.3.9.1. Let Ω be an open set in $\mathbb{C}$, and $\mathbb{D}$ a disk with $\bar{\mathbb{D}} \subseteq \Omega$.

(a) Use the Cauchy Integral Formula to show that for every $z_0 \in \mathbb{D}$, there is a constant C (possibly depending on z_0) such that, whenever f is a holomorphic function on Ω,

$$|f(z_0)| \leq C \cdot \max\{|f(z)| : z \in \partial\mathbb{D}\} \ .$$

[Consider the reciprocal of the distance from z_0 to $\partial\mathbb{D}$.]

(b) Show that the inequality holds with $C = 1$ for every z_0. [If f is holomorphic on Ω, apply the inequality to f^n and take n'th roots, and let $n \to \infty$.]

This gives an alternate proof of the first part of XII.3.4.1..

XII.3.9.2. Let Ω be a convex (or, more generally, simply connected) region. Show that if f is a holomorphic function on Ω which is never zero on Ω, then there is a holomorphic function g on Ω with $f = e^g$. [The function $\frac{f'}{f}$ is holomorphic on Ω and has an antiderivative on Ω.] This is especially interesting if $\Omega = \mathbb{C}$, i.e. f is entire.

XII.4. Sequences and Series of Holomorphic Functions; Analytic Functions

In this section, we show the remarkable fact that holomorphic functions can be expanded in power series around any point in the domain.

XII.4.1. Sequences and Series of Holomorphic Functions

We begin by establishing some general results about sequences and series of holomorphic functions.

XII.4.1.1. In complex analysis, u.c. convergence (IV.3.2.1.) of sequences of holomorphic functions on an open set in $\mathbb{C}$ is sometimes called *normal convergence*. We will usually avoid this terminology since it is not universal, and is in fact ambiguous (V.15.1.4.).

The first result is an immediate corollary of Cauchy's Theorem and Morera's Theorem:

XII.4.1.2. PROPOSITION. Let Ω be an open set in $\mathbb{C}$, and (f_n) a sequence of functions from Ω to $\mathbb{C}$ converging u.c. on Ω to a function f. If each f_n is holomorphic on Ω, then f is also holomorphic on Ω.

PROOF: It suffices to show that f is holomorphic on each open disk contained in Ω; so we may assume Ω is a disk. If γ is any closed contour in Ω, then

$$\int_\gamma f_n \to \int_\gamma f$$

by (). But $\int_\gamma f_n = 0$ for all n by Cauchy's Theorem for Convex Sets (XII.3.2.4.), so $\int_\gamma f = 0$. This is true for every γ, so f is holomorphic by Morera's Theorem. 👍

XII.4.1.3. This is one more surprising property of holomorphic functions. The real analog of this statement fails spectacularly: any continuous function on a closed bounded interval, even one not differentiable anywhere (), is a uniform limit of polynomials.

Actually, the situation is even nicer for holomorphic functions, again in stark contrast to the real case ():

XII.4.1.4. THEOREM. Let Ω be an open set in $\mathbb{C}$, and (f_n) a sequence of functions from Ω to $\mathbb{C}$ converging u.c. on Ω to a function f. If each f_n is holomorphic on Ω, then f is also holomorphic on Ω, and $f_n' \to f'$ u.c. on Ω.

In this result, u.c. convergence cannot be replaced by uniform convergence in general: even if $f_n \to f$ uniformly on Ω, f_n' does not converge uniformly on Ω to f' in general, only u.c. (Exercise XII.9.9.2.).

PROOF: This is a fairly easy, although slightly delicate, consequence of the Cauchy Integral Formula for derivatives (XII.3.3.6.); the argument also gives another proof of XII.4.1.2.. By IV.3.2.5., it suffices to show that for any $z_0 \in \Omega$, there is an open disk D centered at z_0 and contained in Ω on which f is holomorphic and such that $f_n' \to f'$ uniformly on D. In fact, let D be any open disk centered at z_0 whose closure $\bar{D}$ is

contained in Ω (there is always such a disk for any $z_0 \in \Omega$). Let R be the radius of D. For sufficiently small $\delta > 0$, the closed disk of radius $R + \delta$ around z_0 is contained in Ω. Fix such a δ, and let γ be the circle centered at z_0 of radius $R + \delta$, parametrized once around counterclockwise. For $z \in \mathbb{C} \setminus \gamma$, set

$$g(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta$$

and note that g is holomorphic on $C \setminus \gamma$ by XII.3.3.5.. By the Cauchy Integral Formula, if z is inside γ,

$$f_n(z) = \frac{1}{2\pi i} \int_\gamma \frac{f_n(\zeta)}{\zeta - z}\, d\zeta$$

for each n. We have that $f_n \to f$ uniformly on γ. As $n \to \infty$, the sequence

$$\frac{f_n(\zeta)}{\zeta - z}$$

converges uniformly on γ (for fixed z inside γ) to

$$\frac{f(\zeta)}{\zeta - z}$$

since the numerators converge uniformly and the absolute value of the denominator is always $\geq \delta$. Thus we have

$$f_n(z) = \frac{1}{2\pi i} \int_\gamma \frac{f_n(\zeta)}{\zeta - z}\, d\zeta \;\to\; \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta = g(z)$$

and since $f_n \to f$ pointwise on Ω, we conclude that $f(z) = g(z)$ for all z inside γ, and f is holomorphic on D. Since z_0 and D are arbitrary, f is holomorphic on Ω.

By a similar but slightly more delicate argument, if z_0, D, and γ are as above, we have that

$$\frac{f_n(\zeta)}{(\zeta - z)^2} \to \frac{f(\zeta)}{(\zeta - z)^2}$$

uniformly for $\zeta \in \gamma$ and $z \in \bar{D}$, since the numerators converge uniformly and the absolute value of the denominator is always $\geq \delta^2$, i.e. for any $\epsilon > 0$ there is an N such that

$$\left| \frac{f_n(\zeta)}{(\zeta - z)^2} - \frac{f(\zeta)}{(\zeta - z)^2} \right| < \epsilon$$

for all $n \geq N$ and all $\zeta \in \gamma$ and $z \in \bar{D}$. Thus, for $z \in \bar{D}$,

$$f_n'(z) = \frac{1}{2\pi i} \int_\gamma \frac{f_n(\zeta)}{(\zeta - z)^2}\, d\zeta \;\to\; \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{(\zeta - z)^2}\, d\zeta = f'(z)$$

uniformly on $\bar{D}$ by the Cauchy Integral Formula for derivatives. 👍

XII.4.1.5. The proof of this theorem is a bit delicate. However, the "converse," the complex analog of V.8.5.5., is simpler:

XII.4.1.6. THEOREM. Let Ω be a connected open set in $\mathbb{C}$, and (f_n) a sequence of functions from Ω to $\mathbb{C}$. If

(1) Each f_n is holomorphic on Ω.

(2) The sequence (f_n') converges u.c. on Ω to a function g.

(3) There is a $z_0 \in \Omega$ such that the sequence $(f_n(z_0))$ converges.

Then the sequence (f_n) converges u.c. on Ω to a function f; f is holomorphic on Ω, and $f' = g$ on Ω.

PROOF: The proof of V.8.5.5. uses the Mean Value Theorem, which is unavailable in $\mathbb{C}$. However, we may mimic the simpler proof of the special case (). We first show that (f_n) converges uniformly on a neighborhood of z_0 (in fact, on any closed disk centered at z_0 contained in Ω) to an antiderivative for g. Fix a closed disk D of radius r centered at z_0 and contained in Ω. Suppose $f_n(z_0) \to c$. If $z \in D$, let γ_z be the line segment from z_0 to z. Then

$$\int_{\gamma_z} f_n' = f_n(z) - f_n(z_0)$$

for any n by (); and if we define

$$f(z) = c + \int_{\gamma_z} g$$

then $f' = g$ on the interior of D. For any $\epsilon > 0$ there is an N such that, for all $n \geq N$, $|f_n'(z) - g(z)| < \frac{\epsilon}{2r}$ for all $z \in D$ and $|f_n(z_0) - c| < \frac{\epsilon}{2}$. Then, for $n \geq N$ and $z \in D$,

$$|f_n(z) - f(z)| = \left| f_n(z_0) + \int_{\gamma_z} f_n' - c - \int_{\gamma_z} g \right| \leq |f_n(z_0) - c| + \left| \int_{\gamma_z} (f_n' - g) \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2r} r = \epsilon \ .$$

Thus $f_n \to f$ uniformly on D.

Now let U be the set of all $z \in \Omega$ such that (f_n) converges uniformly to an antiderivative for g on a neighborhood of z. Then U is obviously open, and $z_0 \in U$. If $w \in U$, then (f_n) converges uniformly on any closed disk centered at w contained in U to an antiderivative for g by the first part of the proof (with w in place of z_0). If $w \in \Omega$ is in the closure of U, there is a sequence (z_n) in U with $z_n \to z$. There is an $\epsilon > 0$ such that the disk of radius ϵ around w is contained in Ω. Fix n with $|z_n - w| < \frac{\epsilon}{3}$; then the closed disk around z_n of radius $\frac{\epsilon}{2}$ is contained in Ω, and is a neighborhood of w on which (f_n) converges uniformly to an antiderivative for g. Thus $w \in U$. So U is relatively closed in Ω, and since Ω is connected, $U = \Omega$. Thus (f_n) converges uniformly to an antiderivative for g on any closed disk contained in Ω; so (f_n) converges pointwise on Ω to a function f which is an antiderivative for g. Since any compact subset of Ω is contained in a finite union of closed disks in Ω, $f_n \to f$ u.c. on Ω. 👍

As usual, the sequence results have series versions, proved by applying the sequence results to the partial sums:

XII.4.1.7. PROPOSITION. Let Ω be an open set in $\mathbb{C}$, and

$$\sum_{k=1}^{\infty} f_k$$

an infinite series of functions from Ω to $\mathbb{C}$ converging u.c. on Ω to a function f. If each f_n is holomorphic on Ω, then f is also holomorphic on Ω.

XII.4.1.8. THEOREM. Let Ω be an open set in $\mathbb{C}$, and

$$\sum_{k=1}^{\infty} f_k$$

an infinite series of functions from Ω to $\mathbb{C}$ converging u.c. on Ω to a function f. If each f_n is holomorphic on Ω, then f is also holomorphic on Ω, and

$$\sum_{k=1}^{\infty} f_k'$$

converges u.c. to f' on Ω.

XII.4.1.9. THEOREM. Let Ω be a connected open set in $\mathbb{C}$, and

$$\sum_{k=1}^{\infty} f_k$$

an infinite series of functions from Ω to $\mathbb{C}$. If

(1) Each f_k is holomorphic on Ω.

(2) The infinite series

$$\sum_{k=1}^{\infty} f_k'$$

converges u.c. on Ω to a function g.

(3) There is a $z_0 \in \Omega$ such that the series

$$\sum_{k=1}^{\infty} f_k(z_0)$$

converges.

Then the infinite series

$$\sum_{k=1}^{\infty} f_k$$

converges u.c. on Ω to a function f; f is holomorphic on Ω, and $f' = g$ on Ω.

XII.4.2. Power Series

We recall some of the important general facts about power series from V.15.3. in the complex case.

XII.4.2.1. DEFINITION. If z_0 is a complex number, a *(complex) power series centered at* z_0 is an expression of the form

$$\sum_{k=0}^{\infty} a_k (z - z_0)^k$$

where the a_k are complex numbers.

XII.4.2.2. A power series is regarded as an infinite series of functions. It is conventional when working with power series to regard $(z - z_0)^0 = 1$ no matter how $z - z_0$ is interpreted; thus the $k = 0$ term of the power series is just the constant function a_0.

XII.4.2.3. A power series is, at first, only a formal expression, but it can be regarded as defining a function f of z, whose domain is the set of all z for which the series converges. Because of our convention, z_0 is always in the domain of the defined function f, with $f(z_0) = a_0$. It can happen that the domain of f consists only of $\{z_0\}$ (such power series are not very interesting as functions).

The most important basic property of power series is XII.4.2.5., which says that the domain of the defined function is a disk, or degenerate disk, centered at z_0.

XII.4.2.4. DEFINITION. Let $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ be a power series, and define

$$L = \limsup_{k \to \infty} |a_k|^{1/k} = \limsup_{k \to \infty} \sqrt[k]{|a_k|}$$

and let $R = \frac{1}{L}$ if $0 < L < +\infty$, $R = 0$ if $L = +\infty$, and $R = +\infty$ if $L = 0$. R is called the *radius of convergence* of the power series.

Note that the radius of convergence of a power series depends only on the coefficients a_k, not on the center z_0, of the power series. The slower the coefficients grow in size, or the faster they approach zero, the larger the radius of convergence.

The term "radius of convergence" is justified by the next result:

XII.4.2.5. THEOREM. Let $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ be a power series, with radius of convergence R. Then

(i) If $z \in \mathbb{C}$ and $|z - z_0| < R$, then the series $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ converges absolutely.

(ii) If $z \in \mathbb{C}$ and $|z - z_0| > R$, then the series $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ diverges.

(iii) If $r \in \mathbb{R}$ and $0 < r < R$, then the series $\sum_{k=0}^{\infty} a_k(z-z_0)^k$ converges uniformly on the closed disk $\{z : |z-z_0| \le r\}$.
(Of course, (i) and (iii) are vacuous if $R = 0$ and (ii) is vacuous if $R = +\infty$.)

PROOF: In cases (i) and (ii), apply the Root Test () to the series $\sum_{k=0}^{\infty} |a_k(z-z_0)^k|$. We have

$$\limsup |a_k(z-z_0)^k|^{1/k} = |z-z_0| \limsup |a_k|^{1/k} = \frac{|z-z_0|}{R}$$

(it is $+\infty$ if $z \neq z_0$ and $R = 0$, and 0 if $R = +\infty$ or $z = z_0$). Thus the series converges absolutely if $|z-z_0| < R$ or $z = z_0$ and diverges since the terms do not go to zero if $|z-z_0| > R$.
For (iii), use the Weierstrass M-test (). Set $M_k = |a_k| r^k$. Then $|a_k(z-z_0)^k| \le M_k$ for all k and for all x, $|z-z_0| \le r$. We have

$$\limsup M_k^{1/k} = r \limsup |a_k|^{1/k} = \frac{r}{R} < 1$$

if $0 < R < \infty$, and $\limsup M_k^{1/k} = 0$ if $R = +\infty$. Thus $\sum_{k=0}^{\infty} M_k$ converges by the Root Test. 👍

XII.4.2.6. Thus the set on which a power series centered at z_0 converges is a disk centered at z_0, possibly of radius zero, i.e. the degenerate disk $\{z_0\}$, or of infinite radius, i.e. all of $\mathbb{C}$, possibly including some or all of the points on the boundary, called the *disk of convergence* of the power series.

XII.4.2.7. Convergence behavior of a complex power series on the boundary circle of the disk of convergence is quite subtle. The series can diverge everywhere on this circle (e.g. $\sum_{k=0}^{\infty}(z-z_0)^k$) or it can converge absolutely at all points of the circle (e.g. $\sum_{k=1}^{\infty} \frac{1}{k^2}(z-z_0)^k$), or it can converge conditionally on some subset S of the boundary circle. But S cannot be an arbitrary subset of the circle (it must be an $F_{\sigma\delta}$ (), and not even all of these occur; see [Kör83]). The study of boundary behavior of analytic functions is a big topic in complex analysis, with important applications.

For the purposes of basic complex analysis, however, we want to restrict attention to open sets, so we generally consider only the *open disk of convergence* $\{z : |z-z_0| < R\}$ (provided $R > 0$).

Combining XII.4.2.5. with XII.4.1.7., we obtain:

XII.4.2.8. COROLLARY. Let $\sum_{k=0}^{\infty} a_k(z-z_0)^k$ be a power series, with radius of convergence $R > 0$. Then the series converges u.c. on the open disk of convergence, and the sum is holomorphic on this disk.

Using XII.4.1.8., we can extend this:

XII.4.2.9. COROLLARY. Let $\sum_{k=0}^{\infty} a_k(z-z_0)^k$ be a power series, with radius of convergence $R>0$. Then the series converges u.c. on the open disk of convergence $\{z : |z-z_0|<R\}$, and the sum f is holomorphic on this disk. The power series

$$\sum_{k=1}^{\infty} k a_k (z-z_0)^{k-1}$$

of term-by-term derivatives converges u.c. on $\{z : |z-z_0|<R\}$ to f'.

This can be also be proved by noting that the derived series has the same radius of convergence as the original series (V.15.3.21.) and applying XII.4.1.9..

By iterating, we obtain, for each n, that the n'th derived power series

$$\sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1)a_k(z-z_0)^{k-n} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} a_k (z-z_0)^{k-n}$$

converges u.c. to $f^{(n)}$ on $\{z : |z-z_0|<R\}$.

The coefficients a_k are thus uniquely determined by the function f defined by the power series:

XII.4.2.10. COROLLARY. Let $\sum_{k=0}^{\infty} a_k(z-z_0)^k$ be a power series, with radius of convergence $R>0$, and let f be the holomorphic function defined by the series on the open disk of convergence. Then, for any $n \geq 0$,

$$a_n = \frac{f^{(n)}(z_0)}{n!} \ .$$

The coefficients a_k can also be computed by contour integration:

XII.4.2.11. PROPOSITION. Let

$$\sum_{k=0}^{\infty} a_k (z-z_0)^k$$

be a power series with radius of convergence $R>0$, and f the function defined by the series on $\{z : |z-z_0|<R\}$. Then, for any $k \geq 0$,

$$a_k = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{(\zeta-z_0)^{k+1}}\, d\zeta$$

where γ is any circle centered at z_0 of radius less than R.

PROOF: By (), we have

$$\int_\gamma \frac{f(\zeta)}{(\zeta-z_0)^{k+1}}\, d\zeta = \int_\gamma (\zeta-z_0)^{-k-1} \left[\sum_{j=0}^{\infty} a_j (\zeta-z_0)^j \right] d\zeta$$

$$= \sum_{j=0}^{\infty} \left[\int_\gamma a_j (\zeta - z_0)^{j-k-1} \, d\zeta \right] .$$

All the terms in this sum are 0 except the term with $j = k$, which is $2\pi i a_k$ (). 👍

This result can also be concluded from XII.4.2.10. and the Cauchy Integral Formula for Derivatives (XII.3.3.6.).

XII.4.2.12. The radius of convergence of a power series $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ can always be computed as

$$R = \frac{1}{\limsup |a_k|^{1/k}}$$

(with the usual conventions if the lim sup is 0 or $+\infty$), but it is not always easy to evaluate this expression. Sometimes the radius of convergence can be computed more simply by the following:

XII.4.2.13. PROPOSITION. Let $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ be a power series, with all a_k nonzero. If

$$\lim_{k \to \infty} \left| \frac{a_k}{a_{k+1}} \right|$$

exists (in the usual or extended sense), it equals the radius of convergence of the power series.

PROOF: Let S be this limit. Apply the Ratio Test () to show that $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ converges absolutely if $|z - z_0| < S$ and diverges if $|z - z_0| > S$. Thus S must be exactly the radius of convergence. 👍

Note that in the limit of XII.4.2.13., the a_k and a_{k+1} are reversed from the way they appear in the Ratio Test ().

Another convenient characterization of the radius of convergence is the following:

XII.4.2.14. PROPOSITION. Let $\sum_{k=0}^{\infty} a_k (z - z_0)^k$ be a power series. The power series has radius of convergence $R > 0$ if and only if, for all r, $0 < r < R$, there is a constant M such that $|a_k| \leq \frac{M}{r^k}$ for all k, and there is no such M if $r > R$.

PROOF: Suppose the series has radius of convergence $R > 0$, and $0 < r < R$. Then

$$\limsup |a_k|^{1/k} < \frac{1}{r}$$

so there is a k_0 such that $|a_k| < \frac{1}{r^k}$ for all $k \geq k_0$. There is an $M \geq 1$ such that $|a_k| \leq \frac{M}{r^k}$ for $0 \leq k < k_0$. This M works for all k.

Conversely, if the condition holds for all r with $0 < r < R$ and no $r > R$, we have, for each such r with $0 < r < R$,

$$\limsup |a_k|^{1/k} \leq \limsup \frac{M^{1/k}}{r} = \frac{1}{r}$$

for some constant M, so the radius of convergence R' is $\geq R$. If $R' > R$, choose r so that $R < r < R'$; then the condition holds for r by the first half of the proof, a contradiction. 👍

XII.4.3. Inverse Power Series

It will be convenient to extend the results about power series to closely related series we call inverse power series:

XII.4.3.1. DEFINITION. An *inverse power series* centered at $z_0 \in \mathbb{C}$ is an infinite series of the form

$$\sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

where the $b_k \in \mathbb{C}$. (We could also allow a constant term b_0.)

XII.4.3.2. An inverse power series defines a complex-valued function whose domain is the set of z for which the series converges. This set is (almost) nicely describable, as in the case of power series:

XII.4.3.3. THEOREM. Let

$$\sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

be an inverse power series, and set

$$S = \limsup_{k \to \infty} |b_k|^{1/k} \ .$$

Then

(i) If $S < \infty$ and $|z - z_0| > S$, then the series converges absolutely.

(ii) If $S > 0$ and $|z - z_0| < S$, then the series diverges.

(iii) If $S < \infty$ and $s > S$, then the series converges uniformly on

$$\{z : |z - z_0| \geq s\}$$

and in particular the series converges u.c. on

$$\{z : |z - z_0| > S\} \ .$$

PROOF: This follows almost immediately from XII.4.2.5. by considering the power series

$$\sum_{k=1}^{\infty} b_k w^k$$

which has radius of convergence $R = \frac{1}{S}$. If $w = \frac{1}{z-z_0}$, then $|w| < R$ if and only if $|z - z_0| > S$, etc. 👍

XII.4.3.4. The number S is called the *inverse radius of convergence* and the set

$$\{z : |z - z_0| > S\}$$

is called the *open disk complement of convergence* of the inverse power series. If $S = 0$, this set is the entire complex plane with z_0 removed. (The series never converges for $z = z_0$. If $S = \infty$, the series does not converge for any $z \in \mathbb{C}$. If $0 < S < \infty$, the series may converge for some or all z with $|z - z_0| = S$, but not necessarily; cf. XII.4.2.7..)

XII.4.3.5. PROPOSITION. Let

$$\sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

be an inverse power series with finite inverse radius of convergence. Then the series defines a holomorphic function f on its open disk complement of convergence Ω, and the series

$$\sum_{k=1}^{\infty} \frac{-kb_k}{(z - z_0)^{k+1}}$$

of term-by-term derivatives converges u.c. to f' on Ω.

The fact that f is holomorphic is immediate from XII.4.1.7., and that term-by-term differentiation is valid follows from XII.4.1.8.. Alternatively, both parts follow from XII.4.1.9. and the fact that the derived series has the same inverse radius of convergence (cf. V.15.3.21.) and thus converges u.c. on Ω.

The coefficients b_k can be computed by contour integration:

XII.4.3.6. PROPOSITION. Let

$$\sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

be an inverse power series with finite inverse radius of convergence S, and f the function defined by the series on $\Omega = \{z : |z - z_0| > S\}$. Then, for any $k \geq 1$,

$$b_k = \frac{1}{2\pi i} \int_{\gamma} f(\zeta)(\zeta - z_0)^{k-1}\, d\zeta$$

where γ is any circle centered at z_0 of radius greater than S.

PROOF: By (), we have

$$\int_\gamma f(\zeta)(\zeta - z_0)^{k-1}\,d\zeta = \int_\gamma (\zeta - z_0)^{k-1}\left[\sum_{j=1}^{\infty}\frac{b_j}{(\zeta - z_0)^j}\right]d\zeta$$

$$= \sum_{j=1}^{\infty}\left[\int_\gamma b_j(\zeta - z_0)^{k-1-j}\,d\zeta\right] .$$

All the terms in this sum are 0 except the term with $j = k$, which is $2\pi i b_k$ (). 

XII.4.4. Analytic Functions

XII.4.4.1. DEFINITION. Let Ω be an open set in $\mathbb{C}$, and f a complex-valued function on Ω. Then f is *analytic* on Ω if for every $z_0 \in \Omega$ there is a power series centered at z_0 which converges (pointwise, hence necessarily u.c.) to f on some open disk centered at z_0.

Note that the power series representation of f around z_0 need not converge to f on all of Ω, only on *some* open disk centered at z_0. Thus analyticity is a local property: if for each $z_0 \in \Omega$ there is some neighborhood of z_0 in Ω on which f is analytic, then f is analytic on Ω. 

The following facts are immediate from XII.4.2.9.:

XII.4.4.2. COROLLARY. Let Ω be an open set in $\mathbb{C}$, and f a complex-valued function on Ω. If f is analytic on Ω, then f is holomorphic on Ω and f' is also analytic on Ω.

XII.4.4.3. Conversely, it follows easily from XII.4.1.9. that if f is holomorphic on Ω and f' is analytic on Ω, then f is analytic on Ω. We omit details since we will presently prove a much stronger result (XII.4.4.7.).

We also obtain as a consequence of XII.4.2.10.:

XII.4.4.4. COROLLARY. Let Ω be an open set in $\mathbb{C}$, and f an analytic function on Ω. For any $z_0 \in \Omega$, the power series centered at z_0 representing f in an open disk around z_0 is unique, and is

$$\sum_{k=0}^{\infty}\frac{f^{(k)}(z_0)}{k!}(z - z_0)^k$$

which is called the *Taylor series for f around z_0*.

XII.4.4.5. For real functions on an interval I, the following classes of functions are successively considerably larger:

(i) The class of analytic functions on I ().

(ii) The class of $\mathcal{C}^\infty$ functions on I ().

(iii) The class of $\mathcal{C}^1$ functions on I ().

(iv) The class of differentiable functions on I ().

The following spectacular theorem, which is the main result of this section, implies that for complex functions on an open set Ω, the above four classes of functions coincide.

XII.4.4.6. THEOREM. Let Ω be an open set in $\mathbb{C}$ and f a holomorphic function on Ω. If $z_0 \in \Omega$ and D is an open disk centered at z_0 contained in Ω, then the Taylor series for f around z_0 converges to $f(z)$ for every $z \in D$; in fact, the Taylor series converges u.c. to f on D.

Combining this with XII.4.4.2., we obtain:

XII.4.4.7. COROLLARY. Let Ω be an open set in $\mathbb{C}$ and f a complex-valued function on Ω. Then f is analytic on Ω if and only if it is holomorphic on Ω.

PROOF: Suppose f is holomorphic on Ω. Fix $z_0 \in \Omega$ and an open disk D of radius R centered at z_0 contained in Ω. Fix $z \in D$, and let γ be a circle centered at z_0 with radius r for some r, $|z - z_0| < r < R$, parametrized once around counterclockwise. Thus

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta$$

by the Cauchy Integral Formula ().

Since $\left|\dfrac{z - z_0}{\zeta - z_0}\right| = \dfrac{|z - z_0|}{r} < 1$ for all $\zeta \in \gamma$, the infinite series

$$\sum_{k=0}^{\infty} \left(\frac{z - z_0}{\zeta - z_0}\right)^k$$

converges uniformly on γ to

$$\frac{1}{1 - \frac{z - z_0}{\zeta - z_0}} = \frac{\zeta - z_0}{\zeta - z}$$

i.e. for $\zeta \in \gamma$ we have

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0} \sum_{k=0}^{\infty} \left(\frac{z - z_0}{\zeta - z_0}\right)^k = \sum_{k=0}^{\infty} \frac{(z - z_0)^k}{(\zeta - z_0)^{k+1}}$$

where the series converges uniformly on γ. Thus the series

$$\sum_{k=0}^{\infty} \left[(z - z_0)^k \int_\gamma \frac{f(\zeta)}{(\zeta - z_0)^{k+1}}\, d\zeta\right] = \int_\gamma f(\zeta) \left[\sum_{k=0}^{\infty} \frac{(z - z_0)^k}{(\zeta - z_0)^{k+1}}\right] d\zeta$$

converges to

$$\int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta = 2\pi i f(z) \ .$$

In other words, the power series

$$\sum_{k=0}^{\infty} a_k (z - z_0)^k$$

converges to $f(z)$, where

$$a_k = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{(\zeta - z_0)^{k+1}}\, d\zeta$$

for each k. So the open disk of convergence D' of the power series contains D; since the power series converges u.c. on D' and pointwise to f on D, it converges u.c. to f on D. 👍

XII.4.4.8. The proof of the theorem as stated is then completed by either identifying a_k with $\frac{f^{(k)}(z_0)}{k!}$ by the Cauchy Integral Formula for Derivatives (XII.3.3.6.) or by using XII.4.2.10.. In fact, by combining the above proof with XII.4.2.10., we can obtain an alternate proof of the Cauchy Integral Formula for Derivatives. Note that the proof of XII.4.4.6. makes crucial use of the ordinary Cauchy Integral Formula.

XII.4.4.9. Note that XII.4.4.6. says more than just that f is analytic on Ω: if $z_0 \in \Omega$, then the Taylor series for f around z_0 does not merely converge u.c. to f on *some* open disk centered at z_0, but actually converges u.c. to f on the *largest* open disk centered at z_0 contained in Ω (if $\Omega \neq \mathbb{C}$). If $\Omega = \mathbb{C}$, i.e. if f is an entire function, this Taylor series converges u.c. to f on all of $\mathbb{C}$.

This result is again in stark contrast to the real case (cf. ()).

XII.4.4.10. As a result, from now on we will consider *analytic* and *holomorphic* as synonyms for complex functions, and will use the terms interchangeably.

Holomorphic Functions on an Annulus

We can also obtain a similar series representation for a holomorphic function on an annulus, involving both a power series and an inverse power series in general.

XII.4.4.11. Definition. Let $z_0 \in \mathbb{C}$, and fix r and R with $0 \leq r < R \leq +\infty$. The *(open) annulus* centered at z_0 with inner radius r and outer radius R is the set

$$\{z \in \mathbb{C} : r < |z - z_0| < R\} \ .$$

If $0 < r < R < +\infty$, this annulus is the open region between the circles of radius r and R centered at z_0. If $r = 0$ and $R < +\infty$, the annulus is the punctured open disk of radius R centered at z_0 with z_0 removed. If $r = 0$ and $R = +\infty$, it is the punctured plane with z_0 removed; if $r > 0$ and $R = +\infty$, it is the plane with the closed disk of radius r centered at z_0 removed. (It is convenient to include these degenerate cases.)

XII.4.4.12. THEOREM. Let Ω be an open annulus centered at z_0 with inner radius r and outer radius R, and f a holomorphic function on Ω. Then there are a unique power series

$$\sum_{k=0}^{\infty} a_k (z - z_0)^k$$

and a unique inverse power series

$$\sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

which both converge u.c. on Ω, and for which

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

for all $z \in \Omega$. For each $k \geq 0$ we have

$$a_k = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{(\zeta - z_0)^{k+1}} \, d\zeta$$

and for each $k \geq 1$ we have

$$b_k = \frac{1}{2\pi i} \int_\gamma f(\zeta)(\zeta - z_0)^{k-1} \, d\zeta$$

where γ is any circle centered at z_0 whose radius ρ satisfies $r < \rho < R$.

PROOF: The argument is similar to the previous ones. Fix $z \in \Omega$, and let γ_1 and γ_0 be circles centered at z_0 of radius S and s respectively, where $0 < s < |z - z_0| < S < R$, parametrized once around counterclockwise. Then

$$f(z) = \frac{1}{2\pi i} \int_{\gamma_1 - \gamma_0} \frac{f(\zeta)}{\zeta - z} \, d\zeta = \frac{1}{2\pi i} \int_{\gamma_1} \frac{f(\zeta)}{\zeta - z} \, d\zeta - \frac{1}{2\pi i} \int_{\gamma_0} \frac{f(\zeta)}{\zeta - z} \, d\zeta$$

by (). We analyze the two integrals on the right separately.

Since $\left| \dfrac{z - z_0}{\zeta - z_0} \right| = \dfrac{|z - z_0|}{S} < 1$ for all $\zeta \in \gamma_1$, the infinite series

$$\sum_{k=0}^{\infty} \left(\frac{z - z_0}{\zeta - z_0} \right)^k$$

converges uniformly on γ_1 to

$$\frac{1}{1 - \frac{z - z_0}{\zeta - z_0}} = \frac{\zeta - z_0}{\zeta - z}$$

i.e. for $\zeta \in \gamma_1$ we have

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0} \sum_{k=0}^{\infty} \left(\frac{z - z_0}{\zeta - z_0} \right)^k = \sum_{k=0}^{\infty} \frac{(z - z_0)^k}{(\zeta - z_0)^{k+1}}$$

where the series converges uniformly on γ_1. Thus the series

$$\sum_{k=0}^{\infty}\left[(z-z_0)^k\int_{\gamma_1}\frac{f(\zeta)}{(\zeta-z_0)^{k+1}}\,d\zeta\right]=\int_{\gamma_1}f(\zeta)\left[\sum_{k=0}^{\infty}\frac{(z-z_0)^k}{(\zeta-z_0)^{k+1}}\right]d\zeta$$

converges to

$$\int_{\gamma_1}\frac{f(\zeta)}{\zeta-z}\,d\zeta\ .$$

Similarly, since $\left|\dfrac{\zeta-z_0}{z-z_0}\right|=\dfrac{s}{|z-z_0|}<1$ for all $\zeta\in\gamma_0$, the infinite series

$$\sum_{j=0}^{\infty}\left(\frac{\zeta-z_0}{z-z_0}\right)^j$$

converges uniformly on γ_0 to

$$\frac{1}{1-\frac{\zeta-z_0}{z-z_0}}=-\frac{z-z_0}{\zeta-z}$$

i.e. for $\zeta\in\gamma_0$ we have

$$\frac{1}{\zeta-z}=-\frac{1}{z-z_0}\sum_{j=0}^{\infty}\left(\frac{\zeta-z_0}{z-z_0}\right)^j=-\sum_{j=0}^{\infty}\frac{(\zeta-z_0)^j}{(z-z_0)^{j+1}}$$

where the series converges uniformly on γ_0. Thus the series

$$\sum_{k=1}^{\infty}\left[\frac{1}{(z-z_0)^k}\int_{\gamma_0}f(\zeta)(\zeta-z_0)^{k-1}\,d\zeta\right]$$

$$=\sum_{j=0}^{\infty}\left[\frac{1}{(z-z_0)^{j+1}}\int_{\gamma_0}f(\zeta)(\zeta-z_0)^j\,d\zeta\right]=\int_{\gamma_0}f(\zeta)\left[\sum_{j=0}^{\infty}\frac{(\zeta-z_0)^j}{(z-z_0)^{j+1}}\right]d\zeta$$

converges to

$$-\int_{\gamma_0}\frac{f(\zeta)}{\zeta-z}\,d\zeta\ .$$

So we have that both series converge and that

$$f(z)=\sum_{k=0}^{\infty}a_k(z-z_0)^k+\sum_{k=1}^{\infty}\frac{b_k}{(z-z_0)^k}$$

where

$$a_k=\frac{1}{2\pi i}\int_{\gamma_1}\frac{f(\zeta)}{(\zeta-z_0)^{k+1}}\,d\zeta$$

$$b_k=\frac{1}{2\pi i}\int_{\gamma_0}f(\zeta)(\zeta-z_0)^{k-1}\,d\zeta\ .$$

The circles γ_1 and γ_0 depend on z. But note that if γ is any circle centered at z_0 contained in Ω, then, for any k,

$$\int_\gamma \frac{f(\zeta)}{(\zeta - z_0)^{k+1}}\, d\zeta = \int_{\gamma_1} \frac{f(\zeta)}{(\zeta - z_0)^{k+1}}\, d\zeta$$

$$\int_\gamma f(\zeta)(\zeta - z_0)^{k-1}\, d\zeta = \int_{\gamma_0} f(\zeta)(\zeta - z_0)^{k-1}\, d\zeta$$

so the a_k and b_k are independent of the choice of γ_1 and γ_0, i.e. the same series are obtained for all $z \in \Omega$, and these series converge pointwise on Ω.

Since the power series converges pointwise on Ω, its radius of convergence is at least R, hence it converges u.c. on Ω. Similarly, the inverse radius of convergence of the inverse power series is at most r, so it converges u.c. on Ω.

It remains to show uniqueness. If there are series with coefficients c_k and d_k, converging u.c. on Ω with

$$f(z) = \sum_{k=0}^{\infty} c_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{d_k}{(z - z_0)^k}$$

then for any $k \geq 0$ and any circle γ in Ω centered at z_0 we have

$$2\pi i a_k = \int_\gamma \frac{f(\zeta)}{(\zeta - z_0)^{k+1}}\, d\zeta = \sum_{j=0}^{\infty} \left[\int_\gamma c_j (\zeta - z_0)^{j-k-1}\, d\zeta \right] + \sum_{j=1}^{\infty} \left[\int_\gamma \frac{d_k}{((\zeta - z_0)^{j+k+1}}\, d\zeta \right]$$

where term-by-term integration is justified by u.c. convergence. But all terms in the second sum are zero, and all terms in the first are also zero except the term with $j = k$, which is $2\pi i c_k$. Thus $c_k = a_k$. Similarly, if $k \geq 1$ we have

$$2\pi i b_k = \int_\gamma f(\zeta)(\zeta - z_0)^{k-1}\, d\zeta = \sum_{j=0}^{\infty} \left[\int_\gamma c_j (\zeta - z_0)^{j+k-1}\, d\zeta \right] + \sum_{j=1}^{\infty} \left[\int_\gamma \frac{d_k}{((\zeta - z_0)^{j-k+1}}\, d\zeta \right]$$

and the term in the second sum with $j = k$ is $2\pi i d_k$ and all other terms are zero. 👍

The uniqueness statement in this theorem can be extended:

XII.4.4.13. Corollary. Let Ω be an open annulus centered at z_0 with inner radius r and outer radius R, and f a holomorphic function on Ω, with expansion

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

as in XII.4.4.12.. Suppose there are a power series

$$\sum_{k=0}^{\infty} c_k (z - z_0)^k$$

and an inverse power series

$$\sum_{k=1}^{\infty} \frac{d_k}{(z-z_0)^k}$$

converging pointwise on Ω with

$$f(z) = \sum_{k=0}^{\infty} c_k (z-z_0)^k + \sum_{k=1}^{\infty} \frac{d_k}{(z-z_0)^k}$$

for all $z \in \Omega$. Then $c_k = a_k$ and $d_k = b_k$ for all k.

PROOF: Since

$$\sum_{k=0}^{\infty} c_k (z-z_0)^k$$

converges pointwise on Ω, its radius of convergence is at least R, so the series converges u.c. on Ω. Similarly, the inverse power series

$$\sum_{k=1}^{\infty} \frac{d_k}{(z-z_0)^k}$$

converges pointwise on Ω, so its inverse radius of convergence is at most r, so this series converges u.c. on Ω. The result now follows from XII.4.4.12.. 👍

XII.4.4.14. If we set $a_{-k} = b_k$ for $k \geq 1$, we can more compactly write the expansion in XII.4.4.12. as

$$f(z) = \sum_{-\infty}^{\infty} a_k (z-z_0)^k$$

which is called the *Laurent expansion* of f around z_0 (it is technically only a Laurent series (III.5.3.3.) if only finitely many b_k are nonzero, which turns out to often be the case if $r = 0$). But the proper phrasing of convergence of such a series is somewhat tricky, so it is better to consider the positive and negative powers separately in general.

XII.4.4.15. One important difference between XII.4.4.12. and XII.4.4.6. is that there is no obvious interpretation of the coefficients a_k (or b_k) as values of derivatives of f at z_0, which of course do not make sense if f is defined only on an annulus around z_0. However, in at least some cases (e.g. if only finitely many b_k are nonzero) there is an analogous interpretation ().

XII.4.5. Normal Families and Montel's Theorem

There is an important compactness-like property for families of holomorphic functions which we discuss in this section.

Normal Families

XII.4.5.1. Definition. Let Ω be an open set in $\mathbb{C}$, and $\mathcal{F}$ a family of holomorphic functions on Ω. Then $\mathcal{F}$ is a *normal family* if every sequence in $\mathcal{F}$ has a subsequence which converges u.c. ("normally") on Ω.

XII.4.5.2. Normal families turn out to exist in abundance in Complex Analysis, and this fact is the key to many advanced results: they are used to construct holomorphic functions with an exact set of specified properties from families of functions approximately satisfying the properties.

Here is a nice technical simplification used in proving normality for families of holomorphic functions:

XII.4.5.3. Proposition. Let Ω be an open set in $\mathbb{C}$, and $\mathcal{F}$ a family of holomorphic functions on Ω. Suppose there is a sequence (D_j) of compact subsets of Ω such that

(i) Every compact subset of Ω is contained in the union of finitely many D_j.

(ii) For every j, every sequence in $\mathcal{F}$ has a subsequence converging uniformly on D_j.

Then $\mathcal{F}$ is a normal family.

Proof: If (f_n) is a sequence in $\mathcal{F}$, then there is a subsequence $(f_{n_{k,1}})$ which converges uniformly on D_1. This subsequence has a subsubsequence $(f_{n_{k,2}})$ which converges uniformly on D_2. Continue inductively to get, for each j, a sub$\cdots$sequence $f_{n_{k,j}}$ converging uniformly on D_j. The diagonal subsequence $(f_{n_{k,k}})$ then converges uniformly on D_j for each j, and hence on any finite union of the D_j; thus it converges u.c. on Ω. 👍

XII.4.5.4. Corollary. Let Ω be an open set in $\mathbb{C}$, and $\mathcal{F}$ a family of holomorphic functions on Ω. Then $\mathcal{F}$ is a normal family if and only if, for every closed disk D in Ω, any sequence in $\mathcal{F}$ has a subsequence which converges uniformly on D.

The proof of this Corollary really involves no Complex Analysis, only topology. The key is the following topological lemma:

XII.4.5.5. Lemma. Let Ω be an open set in $\mathbb{C}$. Then there is a sequence (D_j) of closed disks in Ω such that every compact subset of Ω is contained in the union of finitely many of the D_j.

Proof: First note that there is a sequence (K_n) of compact sets whose union is Ω: for example, set

$$K_n = \left\{ z \in \Omega : d(z, \mathbb{C} \setminus \Omega) \geq \frac{1}{n}, |z| \leq n \right\} .$$

For each $z \in \Omega$, there is an open disk B_z centered at z whose closure is contained in Ω. For each n, finitely many of the B_z cover K_n; thus countably many of the B_z cover Ω, say $\{B_{z_1}, B_{z_2}, \dots\}$. If K is any compact subset of Ω, finitely many of the B_{z_j} cover K. Take $D_j = \bar{B}_{z_j}$. 👍

This proof works essentially verbatim for any open set in any locally compact, σ-compact metric space ().

Montel's Theorem

XII.4.5.6. THEOREM. [MONTEL'S THEOREM] Let Ω be an open set in $\mathbb{C}$, and $\mathcal{F}$ a family of holomorphic functions on Ω. If $\mathcal{F}$ is uniformly bounded on every closed disk in Ω, then $\mathcal{F}$ is a normal family.

PROOF: We will use XII.4.5.3. for a carefully chosen sequence (D_j) of closed disks. The result will follow immediately from the Arzela-Ascoli Theorem () if $\mathcal{F}$ is equicontinuous on each D_j. We proceed similarly to the proof of XII.4.5.5.: for each $z \in \Omega$ there is an $r > 0$ (depending on z) such that the closed disk of radius $2r$ around z is contained in Ω. Let B_z be the open disk of radius r centered at z. As in the proof of XII.4.5.5., there is a sequence (B_{z_j}) such that every compact subset of Ω is covered by finitely many B_{z_j}. Set $D_j = \bar{B}_{z_j}$.

Equicontinuity of $\mathcal{F}$ on D_j follows from the first Cauchy Inequality, as follows. Fix j. If r is the radius of D_j, then the closed disk D of radius $2r$ centered at z_j is contained in Ω, and if $z \in D_j$, then the closed disk of radius r around z is contained in D. So if M is a bound for $\mathcal{F}$ on D, we have $|f'(z)| \leq \frac{M}{r}$ for all $f \in \mathcal{F}$ and $z \in D_j$. Then, since D_j is convex, for any $z, w \in D_j$ we have

$$|f(z) - f(w)| \leq \frac{M}{r}|z - w|$$

for all $f \in \mathcal{F}$ by XII.2.1.9., and equicontinuity of $\mathcal{F}$ on D_j follows. 👍

The most commonly used special case is called Montel's Theorem in many references:

XII.4.5.7. COROLLARY. Let Ω be an open set in $\mathbb{C}$, and $\mathcal{F}$ a family of holomorphic functions on Ω. If $\mathcal{F}$ is uniformly bounded on Ω, then $\mathcal{F}$ is a normal family.

XII.4.5.8. These are not the most general theorems proved by MONTEL. In [Gam01], XII.4.5.7. is called the "thesis grade" Montel Theorem since MONTEL proved it in his Ph.D. thesis. He later proved a spectacular generalization of XII.4.5.7. (not quite a generalization of XII.4.5.6.) related to the Picard Theorems, which is beyond the scope of this book; see e.g. [Gam01].

XII.5. Zeroes and Singularities

The existence of local power series representations for holomorphic functions has some surprising and pleasant rigidity consequences, and the closely related existence of Laurent expansions around isolated singularities yields more consequences which shed great light on the global properties of holomorphic functions. In fact, it was the philosophy of CAUCHY and, especially, RIEMANN that a holomorphic (analytic) function should be primarily understood through its zeroes and singularities.

XII.5.1. Zeroes of Holomorphic Functions

The first result is immediate from XII.4.4.6., and underlies the rest of the results of this section:

XII.5.1.1. COROLLARY. Let Ω be an open set in $\mathbb{C}$, f holomorphic on Ω, and $z_0 \in \Omega$ with $f(z_0) = 0$. Then either

(i) f is identically zero in a neighborhood of z_0

or

(ii) There is a unique $n \in \mathbb{N}$ and a unique function ϕ which is holomorphic on Ω, with $\phi(z_0) \neq 0$, for which

$$f(z) = (z - z_0)^n \phi(z)$$

for all $z \in \Omega$; and there is a neighborhood U of z_0 containing no other zeroes of f.

PROOF: Let

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k$$

be the Taylor series for f around z_0 and D an open disk centered at z_0 on which this series converges to f. Either $f^{(k)}(z_0) = 0$ for all k, in which case this series is identically zero and f is identically zero on D, or else there is a unique $n \geq 1$ such that $f^{(n)}(z_0) \neq 0$ and $f^{(k)}(z_0) = 0$ for all $k < n$. In this case, the power series

$$\sum_{k=n}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^{k-n}$$

has the same radius of convergence as the Taylor series, and thus converges to a holomorphic function ϕ on D which satisfies $\phi(z_0) = \frac{f^{(n)}(z_0)}{n!} \neq 0$ and

$$\phi(z) = \frac{f(z)}{(z - z_0)^n}$$

for all $z \in D \setminus \{z_0\}$. We may thus extend ϕ to a function holomorphic on Ω with the same properties by setting $\phi(z) = \frac{f(z)}{(z-z_0)^n}$ for $z \in \Omega \setminus D$. Uniqueness of n and ϕ is clear.

For the last statement, since ϕ is continuous and $\phi(z_0) \neq 0$, there is a neighborhood U of z_0 on which $\phi(z) \neq 0$ for all $z \in U$. Then

$$f(z) = (z - z_0)^n \phi(z) \neq 0$$

for all $z \in U \setminus \{z_0\}$ since both factors are nonzero. 👍

A crucial property of holomorphic functions follows, that the zeroes of a nonconstant holomorphic function are *isolated*:

XII.5.1.2. COROLLARY. Let Ω be a connected open set in $\mathbb{C}$, and f a holomorphic function on Ω. Suppose there is a sequence $(z_n) \in \Omega$ of distinct points such that $f(z_n) = 0$ for all n, and that $z_n \to z_0$ for some $z_0 \in \Omega$. Then f is identically zero on Ω.

PROOF: We have that $f(z_0) = 0$ by continuity, and thus f is identically zero in a neighborhood of z_0 by XII.5.1.1.. Let U be the set of all $z \in \Omega$ such that f is identically zero in a neighborhood of z. Then U is open and contains z_0. If $w \in \bar{U} \cap \Omega$, then there is a sequence $(w_n) \in U$ with $w_n \to w$, and thus $w \in U$ as in the first statement of the proof. So U is relatively closed in Ω. Since Ω is connected, $U = \Omega$. 👍

XII.5.1.3. This result says that the zeroes of a nonconstant holomorphic function on a connected Ω cannot have an accumulation point in Ω. Note that they can have an accumulation point outside Ω: set $\Omega = \mathbb{C} \setminus \{0\}$ and $f(z) = \sin \frac{1}{z}$. Then f is holomorphic and nonconstant on Ω and f has a sequence of zeroes (at $\frac{1}{\pi n}$) converging to 0.

Connectedness of Ω is essential: a holomorphic function on a nonconnected Ω could be identically zero on a component of Ω and not on another component.

The statement of XII.5.1.2. becomes seemingly much more powerful by the simple trick of considering $f - g$:

XII.5.1.4. THEOREM. [RIGIDITY THEOREM] Let Ω be a connected open set in $\mathbb{C}$, and f and g holomorphic functions on Ω. Suppose there is a sequence $(z_n) \in \Omega$ of distinct points such that $f(z_n) = g(z_n)$ for all n, and that $z_n \to z_0$ for some $z_0 \in \Omega$. Then f and g are identically equal on Ω.

XII.5.1.5. This result is called the Rigidity Theorem because it means that if f is holomorphic on Ω, the values of f on any small subregion of Ω (or even subset of Ω large enough to have an accumulation point in Ω) completely determine the values of f throughout all of Ω (if Ω is connected).

XII.5.1.6. If a holomorphic function f has an isolated zero at z_0, the number n in XII.5.1.1. is called the *order* of the zero at z_0. This language extends the usual notion of order of a zero of a polynomial. Part of the content of XII.5.1.1. is that an isolated zero of a holomorphic function has finite order.

Limits of Sequences of Holomorphic Functions

Using Montel's Theorem (), we can obtain a considerable extension of the Rigidity Theorem:

XII.5.1.7. THEOREM. Let Ω be a connected open set in $\mathbb{C}$, (f_k) a sequence of holomorphic functions on Ω, and (z_n) a sequence of distinct points in Ω converging to $z_0 \in \Omega$. Suppose the f_k are uniformly bounded on compact subsets of Ω, and that $w_n := \lim_{k\to\infty} f_k(z_n)$ exists for all n. Then there is a unique holomorphic function f on Ω such that $f(z_n) = w_n$ for all n, and $f_n \to f$ u.c. on Ω.

PROOF: By Montel's Theorem, there is a subsequence of (f_k) converging u.c. to a holomorphic function f on Ω, and we must have $f(z_n) = w_n$ for all n. Similarly, if (f_{m_k}) is any subsequence of (f_k), then (f_{m_k}) has a subsequence converging u.c. on Ω to a holomorphic function g with $g(z_n) = w_n$ for all n. By the Rigidity

Theorem, $g = f$ on all of Ω. Thus, if K is a compact subset of Ω containing all the z_n, every subsequence of (f_n) has a subsubsequence converging uniformly to f on K; so $f_n \to f$ uniformly on K (). 👍

This is a powerful result since it shows that to prove that a sequence of holomorphic functions converges u.c. on Ω, it suffices to show the sequence is uniformly bounded on compact subsets of Ω (a necessary condition for u.c. convergence) and converges pointwise on a subset of Ω with a limit point in Ω (e.g. a small subregion of Ω). Note that the Rigidity Theorem is the uniqueness part applied to a constant sequence. But the Rigidity Theorem cannot be considered a corollary of this result since it is used in the proof.

XII.5.2. Singularities

XII.5.2.1. DEFINITION. Let f be holomorphic on an open set Ω. Then f has an *isolated singularity* at $z_0 \in \mathbb{C} \setminus \Omega$ if there is a deleted neighborhood of z_0 contained in Ω.

In other words, f has an isolated singularity at z_0 if it is holomorphic in a neighborhood of z_0 except at z_0.

XII.5.2.2. EXAMPLES. (i) The function

$$f(z) = \frac{z-1}{z^2-1}$$

has isolated singularities at 1 and -1.

(ii) The function $f(z) = e^{1/z}$ has an isolated singularity at 0.

(iii) The function $f(z) = \tan \frac{1}{z}$ has isolated singularities at $\frac{2}{n\pi}$ for all odd n, and a "singularity" at 0 which is not an isolated singularity.

(iv) The functions $\log z$ and z^c, c not an integer, have a "singularity" at 0. But these functions cannot be defined as holomorphic functions on any deleted neighborhood of 0, so no version of them has an isolated singularity at 0. (If c is a negative integer, z^c is a well-defined holomorphic function on $\mathbb{C} \setminus \{0\}$ which has an isolated singularity at 0.)

XII.5.2.3. We will not consider "singularities" such as the ones at 0 in (iii) or (iv) which are not isolated singularities: we will use the term *singularity* to mean "isolated singularity."

The three singularities in (i) and (ii) are examples of the three types of singularity discussed below.

XII.5.2.4. If f has a singularity at z_0, then by XII.4.4.12. there is a unique Laurent expansion

$$f(z) = \sum_{k=0}^{\infty} a_k (z-z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z-z_0)^k}$$

which converges u.c. in a deleted neighborhood of z_0.

Removable Singularities

A singularity such as the one at 1 in XII.5.2.2.(i) is "artificial" since the function can be defined there in such a way that the extended function is holomorphic in a whole neighborhood of 1.

XII.5.2.5. DEFINITION. If f has a singularity (isolated singularity) at z_0, the singularity is a *removable singularity* if there is a function g, holomorphic in a neighborhood of z_0, such that $f = g$ on a deleted neighborhood of z_0.

XII.5.2.6. EXAMPLES. (i) The function $f(z) = \frac{z-1}{z^2-1}$ has singularities at 1 and -1. The singularity at 1 is removable, since $f(z) = g(z) = \frac{1}{z+1}$ in a deleted neighborhood of 1, and g is holomorphic in a neighborhood of 1. The singularity of f at -1 is not removable.

(ii) The function

$$f(z) = \frac{\sin z}{z}$$

has a singularity at 0. It is not so easy to prove directly that

$$\lim_{z\to 0} \frac{\sin z}{z}$$

exists (it does, and equals 1). But on $\mathbb{C} \setminus \{0\}$ f agrees with the entire function g defined by the power series

$$g(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k}$$

and thus the singularity is removable.

If f has a removable singularity at z_0, it is obvious that

$$\lim_{z\to z_0} f(z)$$

exists. It is remarkable that this condition is also sufficient for the singularity to be removable; in fact, apparently much weaker conditions are also sufficient:

XII.5.2.7. THEOREM. If f has a singularity at z_0, the following are equivalent:

(i) The singularity at z_0 is removable.

(ii) $\lim_{z\to z_0} f(z)$ exists.

(iii) f is bounded in a deleted neighborhood of z_0.

(iv) $\lim_{z\to z_0} f(z)(z - z_0) = 0$.

(v) In the Laurent expansion of f around z_0, all b_k are zero.

PROOF: (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv) are obvious.

(iv) $\Rightarrow$ (i): Let D be an open disk centered at z_0 such that f is holomorphic on $D \setminus \{z_0\}$, and let γ be a circle of radius r centered at z_0 contained in $D \setminus \{z_0\}$. Then, for each z inside γ, $z \neq z_0$,

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta - \frac{1}{2\pi i} \int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta$$

where γ_δ is any circle centered at z_0 of radius $\delta < |z - z_0|$ (). Note that, for fixed z, the second integral is independent of δ, as long as $0 < \delta < |z - z_0|$.

Fix z inside γ, and let $\epsilon > 0$. If $\delta > 0$ is sufficiently small, then $|f(\zeta)(\zeta - z_0)| < \epsilon$ whenever $0 < |\zeta - z_0| \leq \delta$. We may assume $\delta \leq \frac{|z-z_0|}{2}$. Then

$$|f(\zeta)| < \frac{\epsilon}{\delta}$$

$$\left|\frac{1}{\zeta - z}\right| \leq \frac{1}{|z - z_0| - \delta} \leq \frac{2}{|z - z_0|}$$

for $\zeta \in \gamma_\delta$, so

$$\left|\int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta\right| < \frac{2\epsilon}{\delta|z - z_0|} \cdot 2\pi\delta = \frac{4\pi\epsilon}{|z - z_0|} \ .$$

Since the integral is independent of δ, and $\epsilon > 0$ is arbitrary, we have

$$\int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z}\, d\zeta = 0$$

$$f(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta \ .$$

This is true for each z inside γ, $z \neq z_0$. But the function

$$g(z) = \frac{1}{2\pi i} \int_\gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta$$

is holomorphic on the entire disk inside γ by ().

(v) $\iff$ (i): Let D be an open disk centered at z_0 such that f is holomorphic on $D \setminus \{z_0\}$, of radius r. If all $b_k = 0$, the Laurent series for f around z_0 is a power series which converges to f on $D \setminus \{z_0\}$; thus its radius of convergence is at least r and it defines a function g which is holomorphic on D and agrees with f on $D \setminus \{z_0\}$. Conversely, if g is a holomorphic function on an open disk D centered at z_0 such that $f = g$ on $D \setminus \{z_0\}$, the Taylor series for g around z_0 converges to f on $D \setminus \{z_0\}$, so must be the Laurent series for f around z_0 by uniqueness of Laurent expansions. 👍

There is a simpler proof that (iii) $\Rightarrow$ (i) (Exercise XII.9.9.12.), and an even simpler proof that (ii) $\Rightarrow$ (i) (Exercise XII.9.9.13.).

XII.5.2.8. The implication (ii) $\Rightarrow$ (i) means that a continuous function from $\mathbb{C}$ (or an open subset) to $\mathbb{C}$ cannot have an isolated point where it is not differentiable (cf. Exercise XII.9.9.13.). This is in contrast to the real case, e.g. for the function $f(x) = |x|$. The theorem also rules out complex analogs of $g(x) = \sin\frac{1}{x}$ ((iii) $\Rightarrow$ (i)) and $h(x) = \frac{1}{\sqrt{|x|}}$ ((iv) $\Rightarrow$ (i)). (The function $g(z) = \sin\frac{1}{z}$ is holomorphic with an isolated singularity at 0, but g is not bounded near 0; in fact the singularity is an essential singularity.)

Poles

XII.5.2.9. Definition. If f has a singularity (isolated singularity) at z_0, the singularity is a *pole of order n* if $\frac{1}{f}$ has a removable singularity at z_0 which is a zero of order n. A pole of order 1 is called a *simple pole*.

Poles can be simply characterized:

XII.5.2.10. PROPOSITION. Suppose f has a singularity at z_0. Then f has a pole at z_0 if and only if

$$\lim_{z\to z_0} |f(z)| = +\infty$$

i.e. for any $M > 0$ there is a $\delta > 0$ such that $|f(z)| > M$ whenever $0 < |z - z_0| < \delta$.

PROOF: If f has a pole at z_0 of order n, and $\frac{1}{f} = g$ in a deleted neighborhood of z_0 and g has a zero of order n at z_0, then there is a deleted neighborhood U of z_0 such that g has no zeroes in U, and we have

$$\lim_{z\to z_0} g(z) = 0$$

and it follows that

$$\lim_{z\to z_0} |f(z)| = +\infty \ .$$

Conversely, if

$$\lim_{z\to z_0} |f(z)| = +\infty$$

there is a deleted neighborhood of z_0 on which f is nonzero, so $h = \frac{1}{f}$ is defined in a deleted neighborhood of z_0. We have

$$\lim_{z\to z_0} h(z) = 0$$

so h has a removable singularity at z_0 and the extended function g has an isolated zero at z_0. 👍

XII.5.2.11. THEOREM. If f has a singularity at z_0, the following are equivalent:

(i) f has a pole of order n at z_0.

(ii) There is a function ϕ which is holomorphic in a neighborhood of z_0 with $\phi(z_0) \neq 0$, such that

$$f(z) = \frac{\phi(z)}{(z - z_0)^n}$$

for all z in a deleted neighborhood of z_0.

(iii) In the Laurent expansion for f around z_0 (XII.5.2.4.), $b_n \neq 0$ but $b_k = 0$ for all $k > n$.

PROOF: (i) ⇒ (ii): There is a function g, holomorphic in a neighborhood of z_0, such that $\frac{1}{f(z)} = g(z)$ for z in a deleted neighborhood of z_0, and there is a function ψ, holomorphic in a neighborhood of z_0 with $\psi(z_0) \neq 0$, such that $g(z) = (z - z_0)^n \psi(z)$ in a neighborhood of z_0. Then $\psi(z) \neq 0$ in a neighborhood U of z_0, and for all $z \in U \setminus \{z_0\}$,

$$f(z) = \frac{1}{(z - z_0)^n} \cdot \frac{1}{\psi(z)} \ .$$

(ii) ⇒ (i): There is a neighborhood U of z_0 such that $\phi(z) \neq 0$ for all $z \in U$ and $f(z) = \frac{\phi(z)}{(z-z_0)^n}$ for all $z \in U \setminus \{z_0\}$. Then

$$g(z) = (z - z_0)^n \frac{1}{\phi(z)}$$

is holomorphic on U, has a zero of order n at z_0, and agrees with $\frac{1}{f}$ on $U \setminus \{z_0\}$.
(ii) $\Rightarrow$ (iii): Suppose the equation holds on $D \setminus \{z_0\}$, where D is an open disk centered at z_0. Let

$$\sum_{k=0}^{\infty} c_k (z - z_0)^k$$

be the Taylor expansion of ϕ around z_0, which converges to ϕ on D. Then $c_0 = \phi(z_0) \neq 0$. Then, for $z \in D \setminus \{z_0\}$,

$$f(z) = \frac{c_0}{(z - z_0)^n} + \cdots + \frac{c_{n-1}}{z - z_0} + \sum_{k=n}^{\infty} c_k (z - z_0)^{k-n}$$

where the power series converges u.c. on D. This must be the Laurent expansion for f around z_0 by uniqueness of Laurent expansions.
(iii) $\Rightarrow$ (ii): Suppose

$$f(z) = \frac{b_n}{(z - z_0)^n} + \cdots + \frac{b_1}{z - z_0} + \sum_{k=0}^{\infty} a_k (z - z_0)^k$$

is the Laurent expansion of f around z_0, with $b_n \neq 0$, converging u.c. on $D \setminus \{z_0\}$, for an open disk D centered at z_0. Set $c_k = b_{n-k}$ for $0 \leq k \leq n-1$ and $c_k = a_{k-n}$ for $k \geq n$. Then the power series

$$\sum_{k=0}^{\infty} c_k (z - z_0)^k$$

converges u.c. on D to a holomorphic function ϕ, with the required properties. 👍

XII.5.2.12. There is thus a symmetry between poles and zeroes: a pole of order n can be thought of as a zero of order $-n$, or a zero of order n can be considered a pole of order $-n$. Indeed, if f is holomorphic and not identically zero in a deleted neighborhood of z_0 and has a pole or removable singularity at z_0, or is already defined and differentiable also at z_0, then the Laurent expansion of f around z_0 is of the form

$$f(z) = \sum_{k=n}^{\infty} a_k (z - z_0)^k$$

where $n \in \mathbb{Z}$ may be positive, negative, or 0 and $a_n \neq 0$. If $n > 0$, f has a zero of order n at z_0 (perhaps after a singularity at z_0 has been removed). If $n < 0$, f has a pole of order $-n$ at z_0. If $n = 0$, then f is holomorphic around z_0 and $f(z_0) \neq 0$, and we may regard f as having a zero or pole at z_0 of order 0.

XII.5.2.13. EXAMPLES. The function $f(z) = \frac{1}{(z-z_0)^n}$ has a pole of order n at z_0, as does the function

$$g(z) = \frac{e^z}{(z - z_0)^n} \ .$$

Since the Taylor series for e^z around z_0 is

$$e^z = \sum_{k=0}^{\infty} \frac{e^{z_0}}{k!} (z - z_0)^k$$

we have that the Laurent series of g around z_0 is

$$\frac{e^{z_0}}{(z-z_0)^n} + \cdots + \frac{e^{z_0}}{(n-1)!(z-z_0)} + \sum_{k=n}^{\infty} \frac{e^{z_0}}{k!}(z-z_0)^{k-n} \ .$$

Essential Singularities

XII.5.2.14. DEFINITION. If f has a singularity (isolated singularity) at z_0, the singularity is an *essential singularity* if it is not a removable singularity or a pole.

XII.5.2.15. We should emphasize that the term *essential singularity* is used only for isolated singularities. Thus the "singularities" at 0 in XII.5.2.2.(iii)–(iv) are not essential singularities (these "singularities" are even more "essential" than essential singularities!)

Essential singularities can be characterized from the Laurent expansion by a process of elimination:

XII.5.2.16. PROPOSITION. If f has a singularity at z_0, and

$$f(z) = \sum_{k=0}^{\infty} a_k(z-z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z-z_0)^k}$$

is its Laurent expansion around z_0, then the singularity at z_0 is an essential singularity if and only if infinitely many b_k are nonzero.

PROOF: The singularity is removable if and only if all b_k are zero, and a pole if and only if at least one b_k, but only finitely many, is/are nonzero.

XII.5.2.17. EXAMPLES. (i) Let $f(z) = e^{1/z}$. Then f has a singularity at 0. As $z \to 0$ along the positive real axis, $|f(z)| \to +\infty$, so the singularity at 0 is not removable. And as $z \to 0$ along the negative real axis, $f(z) \to 0$, so the singularity is not a pole. Thus it is an essential singularity.

This can be seen another way. If $z \neq 0$, the inverse power series

$$1 + \sum_{k=1}^{\infty} \frac{1}{k!}\frac{1}{z^k}$$

converges to $e^{1/z}$. Thus this is the Laurent expansion of f around 0. Since infinitely many negative power coefficients are nonzero, the singularity is essential.

(ii) A closely related example is Cauchy's example (V.17.2.2.) of a $\mathcal{C}^\infty$ function from $\mathbb{R}$ to $\mathbb{R}$ which is not (real) analytic. Let $f(z) = e^{-1/z^2}$. This function is holomorphic on $\mathbb{C} \setminus \{0\}$. As $z \to 0$ along the real axis, $f(z) \to 0$, and if we define $f(0) = 0$, it becomes $\mathcal{C}^\infty$ as a function from $\mathbb{R}$ to $\mathbb{R}$. But if $z \to 0$ along the imaginary axis, $|f(z)| \to +\infty$. Thus f, as a function from $\mathbb{C} \setminus \{0\}$ to $\mathbb{C}$, has an essential singularity at 0.

The next result shows that the behavior of a holomorphic function near an essential singularity is very complicated.

XII.5.2.18. THEOREM. [CASORATI-WEIERSTRASS] Let f have an essential singularity at z_0, and U a deleted neighborhood of z_0 contained in the domain of f. Then the image of U under f is dense in $\mathbb{C}$. If $w \in \mathbb{C}$, then there is a sequence (z_n) in U with $z_n \to z_0$ and $f(z_n) \to w$.

PROOF: The two statements are obviously equivalent (the second follows from the first by taking a decreasing sequence of such neighborhoods), so we prove the first. The proof is very similar to the proof of XII.3.3.11.. Suppose $f(U)$ is not dense in $\mathbb{C}$. Then there is a $w \in \mathbb{C}$ and an $\epsilon > 0$ such that $|f(z) - w| \geq \epsilon$ for all $z \in U$. Set

$$g(z) = \frac{1}{f(z) - w}$$

for $z \in U$; then g is holomorphic on U and has a singularity at z_0. We have $|g(z)| \leq \frac{1}{\epsilon}$ for all $z \in U$, so by XII.5.2.7. this singularity is removable, i.e.

$$c = \lim_{z \to z_0} g(z)$$

exists. Then, if $c \neq 0$, we have

$$\lim_{z \to z_0} f(z) = w + \frac{1}{c}$$

so the singularity of f at z_0 is removable, a contradiction. And if $c = 0$, we have

$$\lim_{z \to z_0} |f(z)| = +\infty$$

so the singularity of f at z_0 is a pole, another contradiction. 👍

Actually, a much stronger result is true:

XII.5.2.19. THEOREM. [BIG PICARD THEOREM] Let f have an essential singularity at z_0, and U a deleted neighborhood of z_0 contained in the domain of f. Then, with at most one exception, f takes every complex value infinitely many times in U.

This result is the best possible: $e^{1/z}$ never takes the value 0. The proof is beyond the scope of this book; see e.g. []. The Big Picard Theorem implies the Little Picard Theorem XII.3.3.12.; cf. Exercise XII.9.9.14..

XII.5.3. Behavior at Infinity

It also makes sense to say that certain functions have singularities at infinity:

XII.5.3.1. DEFINITION. Let f be holomorphic on a set $\{z : |z| > R\}$ for some R. Then f has a *singularity at infinity*. We say f has a *removable singularity* [resp. *pole, essential singularity*] at infinity if the function $g(z) = f\left(\frac{1}{z}\right)$ has a removable singularity [resp. pole, essential singularity] at 0.

We can characterize the types of singularities at infinity:

XII.5.3.2. PROPOSITION. Let f have a singularity at infinity. Then

(i) The singularity at infinity is removable if and only if $\lim_{z\to\infty} f(z)$ exists, i.e. there is a number $c \in \mathbb{C}$ such that, for every $\epsilon > 0$, there is an M such that $|f(z) - c| < \epsilon$ whenever $|z| > M$.

(ii) The singularity at infinity is a pole if and only if $\lim_{z\to\infty} |f(z)| = +\infty$, i.e. for every N there is an M such that $|f(z)| > N$ whenever $|z| > M$.

(iii) Otherwise, the singularity at infinity is an essential singularity.

XII.5.3.3. If f has a singularity at infinity, i.e. it is holomorphic on an annulus $\Omega = \{z : |z| > R\}$, then it has a Laurent expansion

$$f(z) = \sum_{k=0}^{\infty} a_k z^k + \sum_{k=1}^{\infty} \frac{b_k}{z^k}$$

around 0 which converges on Ω. If $g(z) = f\left(\frac{1}{z}\right)$, then the Laurent expansion of g around 0 converging on $\{z : 0 < |z| < 1/R\}$ is obtained by replacing z by z^{-1} in this expansion, effectively interchanging the a_k and b_k. Thus we obtain a classification of singularities at infinity from the Laurent expansion:

XII.5.3.4. PROPOSITION. If f is holomorphic on the annulus $\Omega = \{z : |z| > R\}$, with Laurent expansion

$$f(z) = \sum_{k=0}^{\infty} a_k z^k + \sum_{k=1}^{\infty} \frac{b_k}{z^k}$$

around 0 which converges on Ω, then the singularity at infinity is

(i) Removable if (and only if) all a_k $(k \geq 1)$ are 0.

(ii) A pole if (and only if) some a_k $(k \geq 1)$, but only finitely many, are nonzero.

(iii) Essential if (and only if) infinitely many a_k are nonzero.

As a corollary, we get an important property of entire functions:

XII.5.3.5. COROLLARY. Let f be an entire function. Then the singularity at infinity is

(i) Removable if f is constant.

(ii) A pole if f is a nonconstant polynomial.

(iii) Essential otherwise.

PROOF: The Laurent expansion of f on any annulus centered at 0 is just the Taylor series of f around 0. 👍

The fact that a nonconstant polynomial has a pole at infinity is the essential content of Lemma XII.1.4.1..

XII.5.3.6. We can also define the order of a pole of f at infinity to be the order of the pole of $g(z) = f\left(\frac{1}{z}\right)$ at 0. And we can consider a zero at infinity: f has a zero at infinity if g has a removable singularity at 0 which, when removed, is a zero. The order of the zero of f at infinity is the order of this zero of g. These orders can be read from the Laurent expansion of f in an unbounded annulus in the obvious way, as the largest n for which $a_n \neq 0$ for a pole, and similarly for the b_n for a zero.

XII.5.3.7. We have used expansions around 0 converging in an unbounded annulus centered at 0 merely for convenience. The same results hold if we use unbounded annuli centered at any z_0 by using $g(z) = f\left(\frac{1}{z-z_0} + z_0\right)$, i.e. $g(z - z_0) = f\left(\frac{1}{z-z_0}\right)$. This is occasionally useful when f is naturally expanded around some $z_0 \neq 0$.

XII.5.4. Meromorphic Functions

XII.5.4.1. Definition. Let Ω be a region in $\mathbb{C}$. A *meromorphic function* on Ω is a complex-valued function which is holomorphic on Ω except for (isolated) singularities, none of which is an essential singularity.

We do not exclude the case where the set of singularities is empty. Thus a holomorphic function on Ω is meromorphic on Ω.

XII.5.4.2. By our definition a meromorphic function can have removable singularities (some authors do not allow this). But any meromorphic function can be canonically extended to a meromorphic function with no removable singularities.

XII.5.4.3. Proposition. Let f and g be meromorphic functions on a region Ω. Then

(i) $f + g$, fg, and cf for any $c \in \mathbb{C}$ are meromorphic on Ω.

(ii) If g (extended) is not identically 0 on any component of Ω, then $\frac{f}{g}$ is meromorphic on Ω.

XII.5.4.4. In particular, a quotient of two holomorphic functions (e.g. polynomials) is meromorphic. (Such a quotient can have removable singularities, so this statement would be false if we did not allow removable singularities in meromorphic functions. Even XII.5.4.3.(i) would be false.)

XII.5.4.5. Say two meromorphic functions are equivalent if they are equal except where one has a removable singularity. It is easily checked that this is an equivalence relation. On the set of equivalence classes there are well-defined addition, multiplication, scalar multiplication, and division (by classes not identically 0 on a component). If Ω is connected, these operations make the set of equivalence classes into a field () which is an extension of $\mathbb{C}$ (identified with constant functions).

XII.5.4.6. A meromorphic function (by our definition) on a region Ω can have infinitely many singularities in Ω, but they can cluster only on the boundary of Ω. Some authors allow a meromorphic function to have only finitely many poles in Ω. Equivalence classes of such functions form a subring of the field described in XII.5.4.5., but it is not a subfield (e.g. a nonzero holomorphic function can have infinitely many zeroes in Ω, so its reciprocal has infinitely many poles).

XII.5.5. Exercises

XII.5.5.1. Schwarz Reflection Principle. If Ω is a region in $\mathbb{C}$, set $\bar{\Omega} = \{\bar{z} : z \in \Omega\}$. Then $\bar{\Omega}$ is also a region in $\mathbb{C}$.

(a) If f is holomorphic on Ω, for $z \in \bar{\Omega}$ set $f^*(z) = \overline{f(\bar{z})}$. Show that f^* is holomorphic on $\bar{\Omega}$. [Use the Cauchy-Riemann Equations.]

(b) Let Ω be a connected region containing an interval I on the real axis. Let f be holomorphic on Ω, with $f(z) \in \mathbb{R}$ for $z \in I$. Show that $f = f^*$ on $\Omega \cap \bar{\Omega}$ (which contains $\Omega \cap \mathbb{R}$.) [Use the Rigidity Theorem.]

(c) Let Ω be a region, and f be a function from Ω to $\mathbb{C}$ which is continuous on Ω and holomorphic on $\Omega \setminus \mathbb{R}$. Show that f is holomorphic on Ω. [Use Morera's Theorem.]

(d) Let Ω be a region in the upper half plane whose boundary contains an interval $I \subseteq \mathbb{R}$. Suppose f extends to be continuous on $\Omega \cup I$ and real-valued on I. Let Λ be the interior of $\Omega \cup I \cup \bar{\Omega}$ and define $g : \Lambda \to \mathbb{C}$ by

$$g(z) = \begin{cases} f(z) & \text{if} \quad z \in \Omega \cup I \\ f^*(z) & \text{if} \quad z \in \bar{\Omega} \end{cases} .$$

Show that g is holomorphic on Λ. We say f has been *reflected across* I.

(e) Use fractional linear transformations to extend to the case where I is replaced by an arc A of a circle (or interval on a line) and f takes values on A in another circle arc or line interval B.

XII.6. Mapping Properties of Analytic Functions

XII.6.1. The Complex Morse Lemma

There are complex analogs of V.10.5.2. and V.10.5.6..

XII.6.1.1. If f is analytic on an open disk D around z_0, then for any $n \in \mathbb{N}$

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \cdots + \frac{f^{(n-1)}(z_0)}{(n-1)!}(z - z_0)^{n-1} + \sum_{k=n}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k$$

for $z \in D$, so if we set

$$g_n(z) = \sum_{k=n}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^{k-n}$$

for $z \in D$ (the power series converges u.c. to an analytic function on D), then

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \cdots + \frac{f^{(n-1)}(z_0)}{(n-1)!}(z - z_0)^{n-1} + (z - z_0)^n g_n(z)$$

for $z \in D$, and g_n is analytic on D.

XII.6.1.2. THEOREM. [COMPLEX GENERALIZED ONE-DIMENSIONAL MORSE LEMMA] Let $m \in \mathbb{N}$ and let f be analytic on an open disk D around z_0, with $f(z_0) = c$ and $f'(z_0) = \cdots = f^{(m-1)}(z_0) = 0$ but $f^{(m)}(z_0) \neq 0$. Then there is a conformal bijection ϕ from an open set Ω around 0 onto an open disk $\tilde{D} \subseteq D$ around z_0, with $\phi(0) = z_0$, such that

$$[f \circ \phi](w) = c + w^m$$

for all $w \in \Omega$.

PROOF: By V.10.5.2., for $z \in D$ f can be written $f(z) = c + (z - z_0)^m g_m(z)$ with g_m analytic on D and $g_m(z_0) \neq 0$. Thus there is an open disk $D_0 \subseteq D$ around z_0 in which $g_m(z)$ lies in an open disk U around $g_m(z_0)$ not containing 0. On U there is an analytic branch ρ of $z^{1/m}$. For $z \in D_0$ set

$$\psi(z) = (z - z_0)h(z)$$

where $h = \rho \circ g_m$. Then ψ is an analytic function on D_0, $\psi(z_0) = 0$, and

$$\psi'(z) = h(z) + (z - z_0)h'(z) \ .$$

Since $h(z_0) \neq 0$, $\psi'(z_0) \neq 0$. So ψ is a conformal bijection of a smaller disk $\tilde{D}$ around z_0 onto an open set Ω containing 0. Let $\phi : \Omega \to \tilde{D}$ be the inverse function. Then ϕ is also analytic by the Inverse Function Theorem (XII.1.2.10.). If $w = \psi(z) \in \Omega$, then

$$[f \circ \phi](w) = f(z) = c + (z - z_0)^m g_m(z) = c + [(z - z_0)h(z)]^m = c + \psi(z)^m = c + w^m \ .$$



XII.6.2. Local Mapping of an Analytic Function

Here is an important consequence of the Complex Morse Lemma:

XII.6.2.1. THEOREM. Let f be analytic on a region Ω, $z_0 \in \Omega$, $f(z_0) = c$. Suppose $f(z) - c$ has a zero of order n at z_0, i.e. $f'(z_0) = \cdots = f^{(n-1)}(z_0) = 0$, $f^{(n)}(z_0) \neq 0$. Then for every neighborhood U of z_0 there is a deleted neighborhood V of z_0 and a deleted neighborhood W of c, such that for each $w \in W$, there are exactly n solutions to the equation $f(z) = w$ in V, each of which is a simple zero of the function $f(z) - w$.

PROOF: By the Complex Morse Lemma, we may assume $z_0 = 0$ and f is of the form $f(z) = c + z^n$ on a neighborhood of 0, where it is obvious since every nonzero complex number w has exactly n n'th roots, each of which is a simple zero of $g(z) = z^n - w$. 

XII.6.2.2. In XII.6.2.1., we do not exclude the case $n = 1$, although this has already been proved ().

XII.6.2.3. COROLLARY. Let Ω be a connected region, and f a nonconstant analytic function on Ω. Then f is an open mapping: if U is an open subset of Ω, then $f(U)$ is open.

PROOF: Since Ω is connected, f is not constant in a neighborhood of any point. Thus if c is in the range of f, say $f(z_0) = c$, then f, z_0, and c satisfy the hypotheses of XII.6.2.1. for some n, and thus if U is any neighborhood of z_0 contained in Ω, then every w sufficiently close to c is the image of an element of U.

XII.6.2.4. COROLLARY. Let f be analytic and one-to-one on a region Ω. Then f' is never zero on Ω.

XII.6.2.5. XII.6.2.3. gives a quick alternate proof of the Maximum Modulus Theorem ().

XII.6.2.6. See XII.7.5.1. for another proof (the standard proof in texts) of XII.6.2.1..

XII.7. Residues and the Principle of the Argument

XII.7.1. The Residue Theorem

XII.7.1.1. DEFINITION. Let f be an analytic function with an isolated singularity at z_0. Write

$$f(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_0)^k}$$

converging u.c. on a deleted neighborhood of z_0. The number b_1 is the *residue* of f at z_0, denoted $Res(f, z_0)$.

XII.7.1.2. Because the series representing f near z_0 is unique, the residue is well defined. Note that the residue exists at any type of singularity, not just a pole, although the residue at a removable singularity is always 0. The residue at a simple pole is always nonzero, and the residue at a higher-order pole or essential singularity may or may not be zero.

XII.7.1.3. Why single out b_1? The underlying reason is that the $\frac{b_1}{z-z_0}$ is the only term in the series which does not have an antiderivative on a deleted neighborhood of z_0 (if $b_1 \neq 0$).

XII.7.1.4. If f and g have singularities at z_0, so does $f+g$, and $Res(f+g, z_0) = Res(f, z_0) + Res(g, z_0)$ since the series can be added term-by-term. In particular, if f has a singularity at z_0 and g is analytic in a neighborhood of z_0 (or has a removable singularity at z_0), then $Res(f + g, z_0) = Res(f, z_0)$.

The main theorem of this section, one of the important theorems of Complex Analysis, is:

XII.7.1.5. THEOREM. [RESIDUE THEOREM] Let f be analytic on a region Ω except for isolated singularities. Let γ be a simple closed curve in Ω not passing through any of the singularities of f, such that the interior of γ is contained in Ω. Let $z_1, \ldots, z_n$ be the singularities of f inside γ. Then

$$\int_\gamma f = 2\pi i \sum_{k=1}^{n} Res(f, z_k) \ .$$

(There can be only finitely many singularities inside γ by compactness since the singularities are isolated.)

We give two versions of the proof (they are not really different).

XII.7.1.6. The first step is to note that if $\gamma_1, \ldots, \gamma_n$ are boundary circles of disjoint disks around the z_k contained in the interior of γ, then

$$\int_{\gamma-\gamma_1-\cdots-\gamma_n} f = 0$$

by (), and thus

$$\int_\gamma f = \sum_{k=1}^{n} \int_{\gamma_k} f \ .$$

Thus we may assume that γ is a circle in Ω centered at some z_j whose interior contains no other singularities of f (and not passing through any singularities). Thus the series representation

$$\sum_{k=0}^{\infty} a_k (z - z_j)^k + \sum_{k=1}^{\infty} \frac{b_k}{(z - z_j)^k}$$

for f around z_j converges u.c. to f on the deleted disk bounded by γ, and in particular converges uniformly to f on γ.

XII.7.1.7. The first way to finish the proof is to note that since the series converges uniformly on γ, the series

$$\sum_{k=0}^{\infty} \left[\int_\gamma a_k (\zeta - z_j)^k \, d\zeta \right] + \sum_{k=1}^{\infty} \left[\int_\gamma \frac{b_k}{(\zeta - z_j)^k} \, d\zeta \right]$$

converges to $\int_\gamma f$. Each of the terms has an antiderivative on $\mathbb{C} \setminus \{z_j\}$, thus the integral around γ is 0, except the $\frac{b_1}{\zeta - z_j}$ term. By (),

$$\int_\gamma \frac{b_1}{\zeta - z_j} \, d\zeta = 2\pi i b_1 \ .$$

XII.7.1.8. Alternatively, $\sum_{k=0}^{\infty} a_k (z - z_j)^k$ converges uniformly on and inside γ to an analytic function g, whose integral around γ is 0. And the series $\sum_{k=2}^{\infty} \frac{b_k}{(z-z_j)^k}$ converges u.c. on $\mathbb{C} \setminus \{z_j\}$ to an analytic function h. The function h has an antiderivative on $\mathbb{C} \setminus \{z_j\}$ by (), so its integral around γ is 0. We have $f(z) = g(z) + h(z) + \frac{b_1}{z - z_j}$ for z on and inside γ ($z \neq z_j$). So

$$\int_\gamma f = \int_\gamma g + \int_\gamma h + \int_\gamma \frac{b_1}{\zeta - z_j} \, d\zeta = 0 + 0 + 2\pi i b_1 \ .$$

This completes the proof of the Residue Theorem. 

XII.7.1.9. Note that from the second proof (or just from XII.3.2.3.), if f has a singularity at z_0, then $Res(f, z_0) = 0$ if and only if f has an antiderivative on a deleted neighborhood of z_0.

XII.7.1.10. It has been suggested that CAUCHY got the idea for this theorem from his dog, which left a residue at every pole. [Actually CAUCHY did not use the term *pole*, although he did introduce and use *résidue*.]

XII.7.2. Principle of the Argument

A clever application of the Residue Theorem has far-reaching implications.

XII.7.2.1. Let f be meromorphic on a region Ω, and γ a simple closed curve in Ω (this can be relaxed, cf. ()) not passing through any zero or pole of f (so f is not identically zero inside γ). Then f has only finitely many zeroes inside γ, say at $a_1, \ldots, a_r$ (some may be multiple zeroes), and finitely many poles inside γ, say at $b_1, \ldots, b_s$, since zeroes and poles are isolated in Ω. We will apply the Residue Theorem to the function $\frac{f'}{f}$ to count the number (including multiplicity) of zeroes and poles of f inside γ.

One technicality often overlooked: a meromorphic function (by our definition) can have removable singularities, some of which may become zeroes when filled in. We will assume all removable singularities of f, or at least all which become zeroes (the others do not cause problems), have been removed.

XII.7.2.2. LEMMA. If f is meromorphic on Ω, and not identically zero on any connected component, then

(i) The function $\frac{f'}{f}$ is also meromorphic on Ω, with poles precisely at the zeroes and poles of f, all of which are simple poles.

(ii) If f has a zero at $a \in \Omega$ of order n, then $Res\left(\frac{f'}{f}, a\right) = n$.

(iii) If f has a pole at $b \in \Omega$ of order m, then $Res\left(\frac{f'}{f}, b\right) = -m$.

((ii) and (iii) can be combined into one statement if a pole of order m is considered to be a zero of order $-m$.)

PROOF: The function $\frac{f'}{f}$ is analytic except at the zeroes and poles of f. If f has a zero of order n at a, then $f(z) = (z-a)^n\phi(z)$ on a neighborhood U of a for some analytic function ϕ with $\phi(z) \neq 0$ for $z \in U$. Then $f'(z) = n(z-a)^{n-1}\phi(z) + (z-a)^n\phi'(z)$ on U, so

$$\frac{f'(z)}{f(z)} = \frac{n(z-a)^{n-1}\phi(z)}{(z-a)^n\phi(z)} + \frac{(z-a)^n\phi'(z)}{(z-a)^n\phi(z)} = \frac{n}{z-a} + \frac{\phi'(z)}{\phi(z)}$$

for $z \in U \setminus a$. Since $\frac{\phi'}{\phi}$ is analytic on U, $\frac{f'}{f}$ has a simple pole at a, with $Res\left(\frac{f'}{f}, a\right) = n$.

Similarly, if f has a pole at b of order m, then $f(z) = (z-b)^{-m}\psi(z)$ on a deleted neighborhood V of b for some analytic function ψ on $W := V \cup \{b\}$ with $\psi(z) \neq 0$ for $z \in W$. Then $f'(z) = -m(z-b)^{-m-1}\psi(z) + (z-b)^{-m}\psi'(z)$ on V, so

$$\frac{f'(z)}{f(z)} = \frac{-m(z-b)^{-m-1}\psi(z)}{(z-b)^{-m}\psi(z)} + \frac{(z-b)^{-m}\psi'(z)}{(z-b)^{-m}\psi(z)} = \frac{-m}{z-b} + \frac{\psi'(z)}{\psi(z)}$$

for $z \in V$. Since $\frac{\psi'}{\psi}$ is analytic on W, $\frac{f'}{f}$ has a simple pole at b, with $Res\left(\frac{f'}{f}, b\right) = -m$.

XII.7.2.3. THEOREM. Let f be meromorphic of a region Ω, and γ a simple closed curve in Ω not passing through any zero or pole of f, and such that the interior of γ is contained in Ω. Let Z_f be the number of zeroes of f inside γ and P_f the number of poles of f inside γ, both counted with multiplicities. Then

$$\int_\gamma \frac{f'}{f} = 2\pi i(Z_f - P_f) \ .$$

In particular, if f is analytic on and inside γ, so $P_f = 0$, then

$$Z_f = \frac{1}{2\pi i}\int_\gamma \frac{f'}{f} \ .$$

PROOF: Apply the Residue Theorem to $\frac{f'}{f}$ and use Lemma XII.7.2.2.. 👍

XII.7.2.4. Note that the Residue Theorem also holds if f has some essential singularities inside γ, but in XII.7.2.3. f cannot have essential singularities inside γ. (But see XII.7.5.2..) ✋

XII.7.2.5. The contour integral in the theorem is sometimes called the *logarithmic integral* of f around γ, since the integrand $\frac{f'}{f}$ is the derivative of $\log f$ (although $\log f$ is not well defined on a large enough region in general; indeed, if it were, the contour integral would be 0).

XII.7.2.6. This leads to a very nice geometric interpretation of the logarithmic integral. The curve $\gamma_f(t) = f(\gamma(t))$ is a closed curve in $\mathbb{C}$ (not a simple closed curve in general) not passing through 0. It thus has a well-defined winding number $\omega(f \circ \gamma, 0)$ () around 0. The quantity $2\pi\omega(f \circ \gamma, 0)$ is often denoted $\Delta Arg(f \circ \gamma)$, the *change in argument* as $f \circ \gamma$ is traversed.

XII.7.2.7. PROPOSITION. Under the hypotheses of XII.7.2.3.,

$$2\pi i\omega(f \circ \gamma, 0) = i\Delta Arg(f \circ \gamma) = \int_\gamma \frac{f'}{f} \ .$$

Thus XII.7.2.3. can be rephrased:

XII.7.2.8. THEOREM. [PRINCIPLE OF THE ARGUMENT] Let f be meromorphic of a region Ω, and γ a simple closed curve in Ω not passing through any zero or pole of f, and such that the interior of γ is contained in Ω. Let Z_f be the number of zeroes of f inside γ and P_f the number of poles of f inside γ, both counted with multiplicities. Then

$$\Delta Arg(f \circ \gamma) = 2\pi(Z_f - P_f) \ .$$

In particular, if f is analytic on and inside γ, so $P_f = 0$, then

$$Z_f = \omega(f \circ \gamma, 0) \ .$$

XII.7.2.9. This can be considered a surprising result, since the winding number of $f \circ \gamma$ around 0 has no obvious *a priori* connection with the number of zeroes and poles of f.

XII.7.2.10. The best way to visualize and interpret this is to consider $f \circ \gamma$ as a contour on the Riemann surface () of $\log z$. It is not a closed contour in general: its endpoints lie on different sheets, and the value of the logarithm differs by an integer multiple of $2\pi i$ which is exactly the change of argument. The derivative of $\log f$ is $\frac{f'}{f}$, so if γ is parametrized on $[0,1]$, by XII.2.1.8. we have

$$\int_\gamma \frac{f'}{f} = \log(f(\gamma(1)) - \log(f(\gamma(0)) = i\Delta Arg(f \circ \gamma) \ .$$

(This argument is slightly informal but can be carefully justified, e.g. by analytic continuation, breaking up the path $f \circ \gamma$ into segments, each of which lie in a branch of $\log z$ agreeing at the cut points.)

XII.7.3. Rouché's Theorem

Informally, the number of zeroes of an analytic function inside a simple closed curve is "invariant under small perturbations of the function." Rouché's Theorem is a precise version:

XII.7.3.1. Theorem. [Rouché's Theorem] Let f and h be meromorphic on Ω, and γ a simple closed curve in Ω not passing through any zero or pole of f, and such that the interior of γ is contained in Ω. Suppose $|h(z)| < |f(z)|$ for all $z \in \gamma$ (so γ does not pass through any poles of h). Then $Z_{f+h} - P_{f+h} = Z_f - P_f$. If f and h are analytic on the interior of γ, then f and $f+h$ have the same number of zeroes inside γ.

Proof: By XII.7.2.8. the statement is equivalent to the statement that $\omega((f+h) \circ \gamma) = \omega(f \circ \gamma)$. For $0 \le t \le 1$ set

$$\tilde{\gamma}_t = (f + th) \circ \gamma \ .$$

Then $(\tilde{\gamma}_t)$ is a homotopy of closed paths in $\mathbb{C}$, and $\tilde{\gamma}_0 = f \circ \gamma$, $\tilde{\gamma}_1 = (f+h) \circ \gamma$. For each t and each $z \in \gamma$ we have

$$f(z) + th(z) = f(z)\left(1 + \frac{th(z)}{f(z)}\right)$$

and the second factor lies in the open ball of radius 1 around 1, hence is never zero. Since the first factor is also never zero, the homotopy $(\tilde{\gamma}_t)$ lies in $\mathbb{C} \setminus \{0\}$. Thus $\omega(\tilde{\gamma}_t, 0)$ is constant by (). 👍

XII.7.3.2. Rouché's Theorem is sometimes called the "dog-walking theorem." If a person walks a dog around $\mathbb{C}$ on a leash of variable length, and $f(\gamma(t))$ denotes the person's position at time t, and $f(\gamma(t)) + h(\gamma(t))$ the position of the dog at time t, if there is a pole at the origin and the path $f(\gamma(t))$ of the person winds around the pole some number of times, the path of the dog will wind around the same number of times and the leash will not get tangled around the pole, as long as the length $|h(\gamma(t))|$ of the leash is always kept less than the distance $|f(\gamma(t))|$ of the person from the pole. (It appears that dogs have played a significant role in the history of Complex Analysis; cf. XII.7.1.10..)

XII.7.4. Hurwitz's Theorem

The following theorem has several interesting and useful corollaries, any of which can be called "Hurwitz's Theorem" (and are in various references).

XII.7.4.1. THEOREM. [HURWITZ'S THEOREM] Let Ω be an open set in $\mathbb{C}$, and (f_n) a sequence of holomorphic functions on Ω converging u.c. on Ω to a (holomorphic) function f. Suppose f has a zero of order m at $z_0 \in \Omega$, and let γ be a circle centered at z_0 bounding a closed disk contained in Ω and containing no other zeroes of f. Then, for all sufficiently large n, f_n has exactly m zeroes inside γ, counted with multiplicity.

PROOF: Since $f_n \to f$ uniformly on γ, f_n has no zeroes on γ for sufficiently large n. Thus, passing to a tail of the sequence (f_n), we may assume no f_n has zeroes on γ. Then $\frac{f_n'}{f_n} \to \frac{f'}{f}$ uniformly on γ, so if Z_{f_n} is the number of zeroes of f_n inside γ, we have

$$Z_{f_n} = \frac{1}{2\pi i}\int_\gamma \frac{f_n'}{f_n} \to \frac{1}{2\pi i}\int_\gamma \frac{f'}{f} = m \ .$$

XII.7.4.2. COROLLARY. Let Ω be a connected open set in $\mathbb{C}$, and (f_n) a sequence of holomorphic functions on Ω converging u.c. on Ω to a (holomorphic) function f which is not identically 0. If $f(z_0) = 0$, then there is a sequence (k_n) in $\mathbb{N}$ and a sequence (z_n) in Ω with $z_n \to z_0$ and $f_{k_n}(z_n) = 0$ (informally, every zero of f is a limit of zeroes of the f_n).

PROOF: Let (γ_n) be a sequence of circles centered at z_0 of radius decreasing to 0, within a closed disk contained in Ω containing no other zeroes of f.

XII.7.4.3. The restriction that f be not identically 0 is necessary: consider the sequence $f_n(z) = \frac{e^z}{n}$.

XII.7.4.4. COROLLARY. Let Ω be a connected open set in $\mathbb{C}$, and (f_n) a sequence of holomorphic functions on Ω converging u.c. on Ω to a (holomorphic) function f which is not identically 0. If each f_n has at most m zeroes in Ω (counted with multiplicity), then f also has at most m zeroes in Ω (counted with multiplicity).

PROOF: Suppose f has zeroes at $z_1, \ldots, z_r$ of multiplicity $m_1, \ldots, m_r$ respectively, with $m_1 + \cdots + m_r > m$. For each k, let γ_k be a small circle centered at z_k so that the circles do not overlap and the bounded closed disks are contained in Ω and contain no other zeroes of f. Then, for sufficiently large n, f_n has m_k zeroes inside γ_k for each k, and hence more than m zeroes in all, a contradiction.

XII.7.4.5. Under the hypotheses, even if all f_n have exactly m zeroes in Ω, the limit function can have fewer than m if the zeroes of the f_n cluster on the boundary of Ω. As a very simple case, let Ω be the open right half plane and $f_n(z) = z - \frac{1}{n}$ for each n.

The special case $m = 0$ is often stated as Hurwitz's Theorem:

XII.7.4.6. Corollary. Let Ω be a connected open set in $\mathbb{C}$, and (f_n) a sequence of holomorphic functions on Ω converging u.c. on Ω to a (holomorphic) function f which is not identically 0. If each f_n is never zero in Ω, then f is also never zero in Ω.

Again, the example $f_n(z) = \frac{e^z}{n}$ shows that the assumption that f is not identically 0 is necessary.

The following consequence of a special case of XII.7.4.4. is especially important:

XII.7.4.7. Corollary. Let Ω be a connected open set in $\mathbb{C}$, and (f_n) a sequence of holomorphic functions on Ω converging u.c. on Ω to a (holomorphic) function f which is not constant. If each f_n is one-to-one, then f is also one-to-one.

Proof: For any $w \in \mathbb{C}$, the function $g_n = f_n - w$ has at most one zero in Ω, and hence the limit function $f - w$ also has at most one zero in Ω. 👍

XII.7.4.8. This result can be generalized: if each f_n is at most m-to-one, and f is not constant, then it is at most m-to-one too.

XII.7.5. Exercises

XII.7.5.1. Use Rouché's Theorem to give an alternate proof of XII.6.2.1. (the proof found in most Complex Analysis texts). Suppose f is analytic on Ω, $z_0 \in \Omega$, $f(z_0) = c$, and $g(z) = f(z) - c$ has a zero of order n at z_0.

(a) There is a neighborhood Λ of z_0 such that neither g nor f' has a zero in $\Lambda \setminus \{z_0\}$. Let γ be the boundary circle of a closed disk centered at z_0 contained in Λ.

(b) Let ϵ be less than the minimum of g on γ. If $|w - c| \leq \epsilon$, then the constant function $h(z) = c - w$ satisfies $|h(z)| < |g(z)|$ on γ.

(c) Apply Rouché's Theorem to conclude that $f(z) - w = g(z) + h(z)$ has the same number of zeroes inside γ as $g(z)$.

(d) Give an alternate argument for (c) using XII.7.2.3. but not Rouché's Theorem: for $|w - c| \leq \epsilon$ set

$$\phi(w) = \frac{1}{2\pi i} \int_\gamma \frac{f'(\zeta)}{f(\zeta) - w}\, d\zeta \ .$$

Show that ϕ is continuous, and use XII.7.2.3. to show that ϕ is integer-valued; thus ϕ is constant. By XII.7.2.3., $\phi(c) = n$.

XII.7.5.2. (a) If f has an essential singularity at z_0, then by the Big Picard Theorem () $\frac{f'}{f}$ almost always has a singularity at z_0 which is not isolated. The only exception is when 0 is the excluded value of f near z_0.

(b) The standard example of the exception in (a) is when $f = e^g$, where g has a pole or essential singularity at z_0. Show that in this case $Res\left(\frac{f'}{f}, z_0\right) = 0$. [$\frac{f'}{f} = g'$ near z_0.]

(c) Is it true that whenever f has an essential singularity at z_0, and is nonzero in a deleted neighborhood of z_0, then $Res\left(\frac{f'}{f}, z_0\right) = 0$ (i.e. $\frac{f'}{f}$ has an antiderivative on a deleted neighborhood of z_0, cf. XII.7.1.9.)? This seems unlikely, but if true XII.7.2.3. could be extended to allow essential singularities, as long as f has only finitely many zeroes in any compact subset of Ω.

XII.8. Conformal Equivalence and the Riemann Mapping Theorem

Throughout this section, we will let $\mathbb{D}$ denote the open unit disk in $\mathbb{C}$.

XII.8.1. The Schwarz Lemma

The Schwarz Lemma concerns analytic functions from $\mathbb{D}$ to $\mathbb{D}$. It is a simple consequence of XII.5.1.1. and the Maximum Modulus Theorem XII.3.4.3., but has remarkable consequences.

XII.8.1.1. Theorem. [Schwarz Lemma] Let f be an analytic function from $\mathbb{D}$ to $\mathbb{D}$ with $f(0) = 0$. Then

(i) $|f(z)| \leq |z|$ for all $z \in \mathbb{D}$, and $|f'(0)| \leq 1$.

(ii) If $|f(z)| = |z|$ for some $z \in \mathbb{D}$, $z \neq 0$, or if $|f'(0)| = 1$, then there is a $c \in \mathbb{C}$, $|c| = 1$, with $f(z) = cz$ for all $z \in \mathbb{D}$ (i.e. f is geometrically just a rotation around 0). In fact, in either case $c = f'(0)$.

Proof: There is an analytic function g on $\mathbb{D}$ with $f(z) = zg(z)$ by XII.5.1.1. (or XII.5.2.7.), and

$$g(0) = \lim_{z \to 0} \frac{f(z)}{z} = f'(0) \ .$$

If $0 < r < 1$, for $|z| = r$ we have

$$|g(z)| = \frac{|f(z)|}{r} < \frac{1}{r}$$

and thus by XII.3.4.2. we have $|g(z)| < \frac{1}{r}$ for all z, $|z| \leq r$. Letting $r \to 1$, we obtain $|g(z)| \leq 1$ for all $z \in \mathbb{D}$, and thus $|f(z)| \leq |z|$ for $z \in \mathbb{D}$. If $|g(z)| = 1$ for some $z \in \mathbb{D}$, then g is constant by XII.3.4.2., of absolute value 1, and equal to $g(0) = f'(0)$. 👍

XII.8.1.2. If f is an analytic bijection from $\mathbb{D}$ to $\mathbb{D}$ (so f^{-1} is too, cf. XII.8.2.1.), with $f(0) = 0$, applying the Schwarz Lemma to f and f^{-1} we obtain that $|f(z)| = |z|$ for all $z \in \mathbb{D}$, and hence $f(z) = cz$ for some $c \in \mathbb{C}$, $|c| = 1$. Thus analytic bijections from $\mathbb{D}$ to $\mathbb{D}$ are quite rigid.

XII.8.2. Conformal Equivalence

The following result is a combination of XII.1.2.10., XII.1.1.12., and XII.6.2.4.:

XII.8.2.1. Proposition. Let Ω and Λ be open sets in $\mathbb{C}$, and f an analytic bijection from Ω to Λ. Then

(i) f' is never zero on Ω.

(ii) f is a conformal mapping.

(iii) f^{-1} is an analytic, hence conformal, bijection from Λ to Ω.

XII.8.2.2. DEFINITION. Let Ω and Λ be open sets in $\mathbb{C}$. An analytic bijection from Ω to Λ is called a *conformal bijection* or *conformal equivalence*. Ω and Λ are *conformally equivalent* if there is a conformal bijection between them.

XII.8.2.3. A composition of conformal equivalences is a conformal equivalence, and conformal equivalence is an equivalence relation on open subsets of $\mathbb{C}$. Conformally equivalent open sets are obviously homeomorphic, even diffeomorphic, but diffeomorphic open sets are not necessarily conformally equivalent (XII.8.2.4.). It is a remarkable fact that any two simply connected open sets in $\mathbb{C}$ are not only homeomorphic, but even conformally equivalent (if not all of $\mathbb{C}$) by the Riemann Mapping Theorem.

XII.8.2.4. EXAMPLE. Let C be the punctured plane $\mathbb{C} \setminus \{0\}$, D the punctured disk $\mathbb{D} \setminus \{0\}$, and A a proper annulus $\{z : r < |z| < R\}$, where $0 < r < R < \infty$. Then C, D, and A are all homeomorphic, even diffeomorphic. But no two are conformally equivalent. For any analytic function from C or D to a bounded set has a removable singularity at 0 by XII.5.2.7., so extends to an analytic function on $\mathbb{C}$ or $\mathbb{D}$ respectively. If f is an analytic function from C to a bounded set, then f extends to a bounded entire function, which must be constant by Liouville's Theorem. Thus C cannot be conformally equivalent to A or D (or indeed any bounded set).

It is a little harder to show that A and D are not conformally equivalent. Suppose f is a conformal equivalence from D to A. Then f extends to an analytic function g from $\mathbb{D}$ to $\bar{A}$. The function g cannot send 0 to a point of A: if $g(0) = f(z_0)$ for some $z_0 \in D$, then a small neighborhood of z_0 and small deleted neighborhood of 0 could not have disjoint images under f. Thus $g(0)$ lies on one of the boundary circles of A. But this is impossible since $g(\mathbb{D})$ is an open set in $\mathbb{C}$.

By a more delicate argument (see e.g. [Rud87]), if A_1 and A_2 are two proper annuli, $A_1 = \{z : r_1 < |z - z_1| < R_1\}$ and $A_2 = \{z : r_2 < |z - z_2| < R_2\}$ for $z_1, z_2 \in \mathbb{C}$, $0 < r_1 < R_1 < \infty$, $0 < r_2 < R_2 < \infty$, then A_1 and A_2 are always homeomorphic (diffeomorphic), but are conformally equivalent if and only if $\frac{R_2}{r_2} = \frac{R_1}{r_1}$ (i.e. A_1 and A_2 are similar). Thus conformal equivalence of annuli is quite rigid.

XII.8.3. Simply Connected Sets in $\mathbb{C}$

It is a remarkable fact, partly topological and partly analytic, that the open sets in $\mathbb{C}$ which are conformally equivalent to $\mathbb{C}$ or $\mathbb{D}$ can be cleanly characterized in a number of ways. We first recall a special case of a definition from topology ():

XII.8.3.1. DEFINITION. Let Ω be a nonempty open set in $\mathbb{C}$. Then Ω is *simply connected* if it is connected and every closed contour in Ω is homotopic in Ω to a constant.

Some authors do not require a simply connected space to be connected; but this seems a little silly, especially for open subsets of the plane. In fact, a simply connected space should be path-connected.

XII.8.3.2. THEOREM. Let Ω be a nonempty connected open set in $\mathbb{C}$ ($\mathbb{R}^2$). The following are equivalent:

Topological conditions:

(i) Ω is simply connected.

(ii) For any $p, q \in \Omega$, any two paths from p to q are homotopic in Ω.

(iii) Ω is contractible.

(iv) Ω is homeomorphic to $\mathbb{C}$ and $\mathbb{D}$.

Intermediate condition:

(v) For every closed contour γ in Ω and $z \in \mathbb{C} \setminus \Omega$, the winding number $w(\gamma, z)$ () of γ around z is 0.

Analytic conditions:

(vi) The contour integral of any holomorphic function on Ω around any closed curve in Ω is 0.

(vii) The contour integral of any holomorphic function on Ω along any contour γ in Ω depends only on the endpoints of γ.

(viii) Every holomorphic function on Ω has an antiderivative on Ω.

(ix) Ω is conformally equivalent to $\mathbb{C}$ or $\mathbb{D}$.

If Ω is bounded, these are also equivalent to

(x) $\partial\Omega$ is connected.

(xi) $\mathbb{C} \setminus \Omega$ is connected.

XII.8.3.3. The implications (iii) $\Rightarrow$ (i) $\Longleftrightarrow$ (ii) hold in any topological space (), and (ix) $\Rightarrow$ (iv) $\Rightarrow$ (iii) $\Rightarrow$ (i) are trivial. The implication (x) $\Rightarrow$ (xi) is true in great generality (XIII.13.1.11.), but (xi) $\Rightarrow$ (x) is much more special (cf. XII.8.3.10.). Conditions (i)–(iv) and (x) make sense for open sets in $\mathbb{R}^n$ for any n (and indeed much more generally; (iv) means Ω is homeomorphic to $\mathbb{R}^n$), but are not equivalent for $n \neq 2$: the interior of a solid torus in $\mathbb{R}^3$ satisfies (x) but not (i), and the open region between two concentric spheres in $\mathbb{R}^3$ satisfies (i) but not (iii) or (x). And a contractible open set in $\mathbb{R}^3$ is not homeomorphic to $\mathbb{R}^3$ in general (). (Note also that (i) $\Rightarrow$ (x) fails for bounded open sets in $\mathbb{R}$; (i)–(iv) are equivalent in $\mathbb{R}$ since a simply connected open set in $\mathbb{R}$ is just an open interval.) The implications (xi) $\Rightarrow$ (i) and (iv) $\Rightarrow$ (xi) do not hold for unbounded open sets in $\mathbb{R}^2$, e.g. $\mathbb{C} \setminus \{0\}$ and an infinite open strip. (If we work in the Riemann sphere, (i), (x), and (xi) are equivalent.)

XII.8.3.4. Condition (v) is called an "intermediate condition" since it is essentially topological, but only makes sense in $\mathbb{R}^2$ and is closely connected with contour integration. We have (v) $\Longleftrightarrow$ (vi) $\Longleftrightarrow$ (vii) $\Longleftrightarrow$ (viii) by () and (). If Ω is bounded, (xi) $\Rightarrow$ (v) since winding number is constant on connected subsets of $\mathbb{C} \setminus \Omega$ () and is 0 on the unbounded component. Also, (i) $\Rightarrow$ (vi) is ().

XII.8.3.5. Thus, we have the following implications:

$$\text{(ix)} \Rightarrow \text{(iv)} \Rightarrow \text{(iii)} \Rightarrow \text{(i)} \Longleftrightarrow \text{(ii)} \Rightarrow \text{(v)} \Longleftrightarrow \text{(vi)} \Longleftrightarrow \text{(vii)} \Longleftrightarrow \text{(viii)}$$

and, if Ω is bounded, (x) $\Rightarrow$ (xi) $\Rightarrow$ (v).

XII.8.3.6. To finish the proof of the theorem, we will show (viii) $\Rightarrow$ (ix) and, if Ω is bounded, (v) $\Rightarrow$ (x). The implication (viii) $\Rightarrow$ (ix) is the most spectacular result, called the Riemann Mapping Theorem; the proof is complicated and is given in XII.8.4.. In [GK06], an open set satisfying (viii) is called *holomorphically simply connected*, which seems like good terminology.

XII.8.3.7. Even the implication (i) $\Rightarrow$ (iv), a purely topological result, is highly nontrivial. The simplest proof of this implication seems to be via (ix) using the Riemann Mapping Theorem, a much stronger result! As noted above, in Euclidean spaces this result is special to $\mathbb{R}^2$.

The Bounded Case

To prove that (v) $\iff$ (x) $\iff$ (xi) it is convenient to prove a more general result:

XII.8.3.8. Theorem. Let Ω be a connected open set in S^1. Then (v) $\iff$ (x) $\iff$ (xi).

We first show (v) and (xi) are equivalent in a more general setting.

XII.8.3.9. Lemma. Let U be a proper open subset of S^1, not necessarily connected. Then the following are equivalent:

(v) Whenever γ is a simple piecewise-smooth closed curve in U and $p \in S^1 \setminus U$, and $S^1 \setminus \{p\}$ is smoothly identified with $\mathbb{C}$, the winding number $w(\gamma, z)$ () of γ around z is 0 for every $z \in \mathbb{C} \setminus U$.

(xi) $S^1 \setminus U$ is connected.

(In this case (v) means that every connected component of U is simply connected.)

Proof: (xi) $\Rightarrow$ (v): Take $p \in S^1 \setminus U$, and identify $S^1 \setminus \{p\}$ with $\mathbb{C}$. Then $\mathbb{C} \setminus U$ has no bounded component. If γ is a simple piecewise-smooth curve in U, then $f(z) = w(\gamma, z)$ is constant on connected components of $\mathbb{C} \setminus U$ and zero on any unbounded component, thus is 0 everywhere on $\mathbb{C} \setminus U$.

(v) $\Rightarrow$ (xi): Suppose $S^1 \setminus U$ is not connected. Then there are nonempty disjoint closed sets E and F with $E \cup F = S^1 \setminus U$. Fix $p \in F$ and identify $S^1 \setminus \{p\}$ with $\mathbb{C}$. Then E is a compact subset of $\mathbb{C}$ and $W := \mathbb{C} \setminus F = U \cup E$ is open. By XIII.25.1.36. there is a simple closed polygonal curve γ in W with E in the inside of γ. If γ is parametrized once around counterclockwise, then $w(\gamma, z) = 1$, and thus (v) does not hold. 👍

We now give the proof of Theorem XII.8.3.8..

Proof: We already have (x) $\Rightarrow$ (xi) $\iff$ (v). Suppose U is connected and $S^1 \setminus U$ is connected. Then U satisfies (v). Let $V = S^1 \setminus \bar{U}$. Then $S^1 \setminus V = \bar{U}$ is connected since U is connected (XIII.13.1.9.). Thus V also satisfies (v). Since U and V are disjoint and satisfy (v), so does $W = U \cup V$. Therefore W satisfies (xi), i.e. $S^1 \setminus W = \partial U$ is connected. 👍

XII.8.3.10. The implication (xi) $\Rightarrow$ (x) is false in general if S^1 is replaced by another closed surface. Let X be a torus in $\mathbb{R}^3$ which is a surface of rotation of a circle in the xz-plane around the z-axis. Then the xz-plane intersects the torus in two circles C_1 and C_2. These two circles divide X into two open cylinders U and V. Then $X \setminus U$ and $X \setminus V$ are connected, but $\partial U = \partial V = C_1 \cup C_2$ is not connected.

XII.8.4. The Riemann Mapping Theorem

The Riemann Mapping Theorem is one of the major theorems of Complex Analysis:

XII.8.4.1. Theorem. [Riemann Mapping Theorem] Let Ω be a holomorphically simply connected open set in $\mathbb{C}$, i.e. a nonempty open set satisfying (viii) of XII.8.3.2., with $\Omega \neq \mathbb{C}$. Then Ω is conformally equivalent to $\mathbb{D}$. More precisely, if $z_0 \in \Omega$, then there is a unique conformal equivalence $f : \Omega \to \mathbb{D}$ with $f(z_0) = 0$ and $f'(z_0)$ a positive real number.

XII.8.4.2. Note that the restriction $\Omega \neq \mathbb{C}$ is necessary: a conformal bijection from $\mathbb{C}$ to $\mathbb{D}$ would be a bounded nonconstant entire function, contradicting Liouville's Theorem ().

XII.8.4.3. We first show uniqueness of f, assuming existence; we will first relax the last condition and show near uniqueness in that setting. Suppose f and g are conformal equivalences from Ω to $\mathbb{D}$ with $f(z_0) = g(z_0) = 0$. Set $h = f \circ g^{-1}$; then h is a conformal bijection from $\mathbb{D}$ to $\mathbb{D}$, and $h(0) = 0$. Thus there is an $\alpha \in \mathbb{C}$, $|\alpha| = 1$, with $h(z) = \alpha z$ (). As a consequence, $f(z) = \alpha g(z)$, i.e. f is just g followed by a rotation (and vice versa). By the Chain Rule we have

$$\alpha = h'(0) = f'(g^{-1}(0))(g^{-1})'(0) = f'(z_0)\frac{1}{g'(z_0)}$$

(note that $g'(z_0) \neq 0$ by ()). If $f'(z_0)$ and $g'(z_0)$ are positive real numbers, then α is a positive real number of absolute value 1, i.e. $\alpha = 1$, $f = g$.

XII.8.4.4. To show existence, we roughly follow the exposition of [GK06]. Let $\mathcal{F}$ be the set of all one-to-one holomorphic functions $f : \Omega \to \mathbb{D}$ with $f(z_0) = 0$. We will show:

1. $\mathcal{F} \neq \emptyset$.
2. There is an $f \in \mathcal{F}$ for which $|f'(z_0)|$ is maximum for all functions in $\mathcal{F}$.
3. A function $f \in \mathcal{F}$ for which $|f'(z_0)|$ is maximum is a conformal bijection.
4. Such an f can be adjusted to make $f'(z_0)$ a positive real number.

As pointed out in [GK06], this outline can be motivated by the case where $\Omega = \mathbb{D}$ and $z_0 = 0$: a holomorphic map f from $\mathbb{D}$ to $\mathbb{D}$ with $f(0) = 0$ is a conformal bijection if and only if $|f'(0)| = 1$, which is the largest possible value of $|f'(0)|$ for holomorphic functions from $\mathbb{D}$ to $\mathbb{D}$ sending 0 to 0 by the Schwarz Lemma.

For both 1 and 3, we need the following lemma:

XII.8.4.5. LEMMA. Let Ω be a holomorphically simply connected open set in $\mathbb{C}$, and f holomorphic on Ω. Suppose f is never 0 on Ω. Then

(i) There is a holomorphic function g on Ω with $e^g = f$ (i.e. g is a "logarithm" for f).

(ii) There is a holomorphic function h on Ω with $h^2 = f$ (i.e. h is a "square root" for f).

PROOF: (i): Since f is never 0 on Ω, $\frac{f'}{f}$ is holomorphic on Ω. Let ϕ be an antiderivative for $\frac{f'}{f}$ on Ω. Fix a $z_0 \in \Omega$ and α a value for $\log(f(z_0))$, i.e. so that $e^\alpha = f(z_0)$. Let $g(z) = \phi(z) + (\alpha - \phi(z_0))$. Then g is an antiderivative for $\frac{f'}{f}$, and $g(z_0) = \alpha$, so $e^{g(z_0)} = f(z_0)$. Then, if $\psi = \frac{e^g}{f}$, we have, for all $z \in \Omega$,

$$\psi'(z) = \frac{f(z)e^{g(z)}g'(z) - e^{g(z)}f'(z)}{[f(z)]^2} = \frac{f(z)e^{g(z)}\frac{f'(z)}{f(z)} - e^{g(z)}f'(z)}{[f(z)]^2} = 0$$

so ψ is constant since Ω is connected. Since $\psi(z_0) = 1$, we have $e^{g(z)} = f(z)$ for all $z \in \Omega$.
(ii): Let g be as in (i), and let $h = e^{g/2}$. 👍

XII.8.4.6. We now prove 1. If Ω is bounded, we could just scale Ω by a small positive number and translate to get an element of $\mathcal{F}$. But Ω is not necessarily bounded, so we proceed as follows. $\Omega \neq \mathbb{C}$; fix $w_0 \in \mathbb{C}\setminus\Omega$, and let ϕ be the restriction to Ω of the function $z \mapsto \frac{1}{z-w_0}$. Then ϕ is holomorphic and one-to-one on Ω, and never takes the value 0. By XII.8.4.5., there is a holomorphic function h on Ω with $h^2 = \phi$. Then not only is h one-to-one, but if it takes the value c it cannot also take the value $-c$. Suppose it takes the value c. Since h is an open mapping, there is an $\epsilon > 0$ such that it takes all values in $B_\epsilon(c)$, and thus takes no values in $B_\epsilon(-c)$. Set

$$g(z) = \frac{\epsilon}{h(z) + c}$$

and note that if $z \in \Omega$, $|h(z) + c| \geq \epsilon$ and thus $|g(z)| \leq 1$; in fact, by the Maximum Modulus Theorem $|g(z)| < 1$. Also, g is one-to-one, i.e. g is a one-to-one holomorphic function from Ω to $\mathbb{D}$. We can correct g by composing with a Möbius transformation (XII.8.6.3.) to get a one-to-one holomorphic function $f : \Omega \to \mathbb{D}$ with $f(z_0) = 0$. Thus $f \in \mathcal{F}$.

XII.8.4.7. The proof of 2 is easy (given some previous results) except for one technicality. There is an $r > 0$ such that $\bar{B}_r(z_0) \subseteq \Omega$; thus by the first Cauchy Inequality we have $|f'(z_0)| \leq \frac{1}{r}$ for all $f \in \mathcal{F}$. Thus $d := \sup\{|f'(z_0)| : f \in \mathcal{F}\}$ is finite (and positive by ()). Choose a sequence (f_n) in $\mathcal{F}$ with $|f_n'(z_0)| > d - \frac{1}{n}$. By Montel's Theorem (XII.4.5.7.), (f_n) has a subsequence (f_{k_n}) which converges u.c. on Ω to a holomorphic function f. Then (f_{k_n}') converges u.c. to f' on Ω by XII.4.1.4., and thus $|f'(z_0)| = d$. So f is not constant. We have that $f : \Omega \to \bar{\mathbb{D}}$, and by the Maximum Modulus Theorem f maps Ω into $\mathbb{D}$ (in fact we will prove later that f maps Ω onto $\mathbb{D}$).

The technicality is that it is not obvious that f is one-to-one. But this follows immediately from XII.7.4.7.. Thus $f \in \mathcal{F}$.

XII.8.4.8. To prove 3, we prove the contrapositive: if $g \in \mathcal{F}$ is not a conformal equivalence, i.e. not onto $\mathbb{D}$, we will show that $|g(z_0)| < d$. Suppose $w_0 \in \mathbb{D}$ is not in $g(\Omega)$. Let ϕ be a Möbius transformation with $\phi(w_0) = 0$. Then $\phi \circ g$ is a one-to-one holomorphic function from Ω to $\mathbb{D}$, and does not take the value 0. By Lemma XII.8.4.5., there is an h, holomorphic on Ω, with $h^2 = \phi \circ g$. Then h is one-to-one and maps Ω into $\mathbb{D}$. Set $w_1 = h(z_0)$, and let ψ be a Möbius transformation with $\psi(w_1) = 0$. Thus $\theta = \psi \circ h$ is in $\mathcal{F}$. If $s(z) = z^2$, then s maps $\mathbb{D}$ to $\mathbb{D}$ and we have

$$g = \phi^{-1} \circ s \circ \psi^{-1} \circ \theta = \omega \circ \theta$$

where $\omega = \phi^{-1} \circ s \circ \psi^{-1}$. Then ω is a holomorphic function from $\mathbb{D}$ to $\mathbb{D}$, and $\omega(0) = 0$. Since ω is not one-to-one (because s is not one-to-one) and hence is not a rotation, we have $|\omega'(0)| < 1$ by the Schwarz Lemma. By the Chain Rule,

$$|g'(z_0)| = |\omega'(\theta(z_0))||\theta'(z_0)| = |\omega'(0)||\theta'(z_0)|$$

and, since $|\omega'(0)| < 1$ and $|\theta'(z_0)| > 0$, we have $|g'(z_0)| < |\theta'(z_0)| \leq d$.

XII.8.4.9. Step 4 is easy: If f is a conformal bijection from Ω to $\mathbb{D}$ (e.g. the one constructed in 2), then $f'(z_0) \neq 0$. Set $\alpha = \frac{f'(z_0)}{|f'(z_0)|}$ and $g(z) = \bar{\alpha} f(z)$. Then g is also a conformal bijection with $g(z_0) = 0$, and $g'(z_0) = |f'(z_0)| > 0$.

This completes the proof of the Riemann Mapping Theorem. 

XII.8.5. Extension to the Boundary

It will be convenient in this section to consider a conformal bijection f from $\mathbb{D}$ to a simply connected region Ω (the inverse of a conformal bijection from Ω to $\mathbb{D}$). Note that f is defined only on the open set $\mathbb{D}$. We now examine the question whether f can be extended to a function on $\bar{\mathbb{D}}$ which is at least continuous, or better. The existence of an extension is a somewhat subtle matter.

Elementary Considerations

XII.8.5.1. The first observation is that if one conformal bijection extends, so do all. If f and g are conformal bijections from $\mathbb{D}$ to Ω, then $f^{-1} \circ g$ is a conformal bijection from $\mathbb{D}$ to $\mathbb{D}$, hence a fractional linear transformation (XII.8.6.3.), which extends to a conformal bijection h from a neighborhood of $\bar{\mathbb{D}}$ to $\mathbb{C}$, and $h(\bar{\mathbb{D}}) = \bar{\mathbb{D}}$. If f extends to $\bar{f} : \bar{\mathbb{D}} \to \mathbb{C}$, then g extends to $\bar{g} = \bar{f} \circ (h|_{\bar{\mathbb{D}}})$. If $\bar{f}$ is a bijection and/or smooth, so is $\bar{g}$.

XII.8.5.2. If f extends to a continuous function $\bar{f}$ from $\bar{\mathbb{D}}$ to $\mathbb{C}$, then $\bar{f}(\bar{\mathbb{D}})$ is compact, hence bounded, and in particular Ω is bounded. So there is only a hope of such an extension if Ω is bounded, so we will assume this from now on (until ()). Also, if f extends to $\bar{f}$, then $\bar{f}$, and hence f also, is uniformly continuous (); conversely, if f is uniformly continuous, then it extends to $\bar{\mathbb{D}}$ since $\mathbb{C}$ is complete ().

XII.8.5.3. Since $\mathbb{D}$ is dense in $\bar{\mathbb{D}}$, if f extends to $\bar{f}$ on $\bar{\mathbb{D}}$ we must have $f(\mathbb{D}) = \Omega$ dense in $\bar{f}(\bar{\mathbb{D}})$, i.e. $\bar{f}(\bar{\mathbb{D}}) \subseteq \bar{\Omega}$. Since $\bar{f}(\bar{\mathbb{D}})$ is compact and contains Ω, we must have $\bar{f}(\bar{\mathbb{D}}) = \bar{\Omega}$. Since $\bar{\mathbb{D}}$ is the disjoint union of $\mathbb{D}$ and $\partial\mathbb{D} = \mathbb{T}$ and $f(\mathbb{D}) = \Omega$, we must have $\bar{f}(\mathbb{T}) \supseteq \bar{\Omega} \setminus \Omega = \partial\Omega$.

XII.8.5.4. PROPOSITION. If f extends continuously to $\bar{f}$ on $\bar{\mathbb{D}}$, then $\bar{f}(\mathbb{T}) = \partial\Omega$.

PROOF: For $0 < r < 1$ let $\mathbb{D}_r$ be the open disk of radius r around 0, and set $\Omega_r = f(\mathbb{D}_r)$. Then Ω_r is open. We have $f(\mathbb{D} \setminus \mathbb{D}_r) = \Omega \setminus \Omega_r$ since f is a homeomorphism, so

$$\bar{f}(\mathbb{T}) \subseteq \bar{f}(\bar{\mathbb{D}} \setminus \mathbb{D}_r) \subseteq \overline{\Omega \setminus \Omega_r} = \bar{\Omega} \setminus \Omega_r$$

for any r. Since $\cup_{r<1}\Omega_r = \Omega$, the result follows. 👍

XII.8.5.5. Thus if f extends to $\bar{f}$, we must have that $\partial\Omega$ is a continuous image of $\mathbb{T}$ (so could be called a "closed curve" although this can be misleading). This is a nontrivial restriction: $\partial\Omega$ can be very bad, e.g. the Warsaw Circle (XIII.36), a pseudocircle (XIII.19.5.34.), the Land of Wada, a set as in XIII.66 closed off on the right, or worse (but recall (XII.8.3.2.(x)) that it is always at least connected). In fact, since $\mathbb{T}$ is a continuous image of $[0, 1]$ and conversely, by the Hahn-Mazurkiewicz Theorem (XIII.14.5.18.) $\partial\Omega$ is a continuous image of $\mathbb{T}$ if and only if $\partial\Omega$ is locally connected. (Of course, there are more restrictions: for example, $\partial\Omega$ must have empty interior and "surround" a bounded simply connected open set. But $\partial\Omega$ can be rather pathological even if it is locally connected (XII.8.5.10.).)

The Locally Connected Case

The locally connected property of XII.8.5.5. suffices:

XII.8.5.6. THEOREM. Let Ω be a bounded simply connected open set in $\mathbb{C}$, and f a conformal bijection from $\mathbb{D}$ to Ω. Then f extends to a continuous $\bar{f} : \bar{\mathbb{D}} \to \bar{\Omega}$ if and only if $\partial\Omega$ is locally connected.

The proof of this theorem is beyond the scope of this book. See [Pom92] or, for a good alternate exposition, [Con95].

XII.8.5.7. EXAMPLE. Let Ω be $\mathbb{D}$ with the ray $[0, 1)$ on the real axis deleted (the *slit disk*). Then Ω is simply connected and $\partial\Omega$ is $\mathbb{T} \cup [0, 1)$, which is locally connected, so a conformal bijection from $\mathbb{D}$ to Ω extends to a continuous function $\bar{f}$ from $\bar{\mathbb{D}}$ onto $\bar{\Omega} = \bar{\mathbb{D}}$. But $\bar{f}$ is not a bijection since $\partial\Omega$ is not homeomorphic to $\mathbb{T}$, and the inverse function $f^{-1} : \Omega \to \mathbb{D}$ does not extend continuously to $\bar{\Omega}$ (this type of behavior is why we consider conformal bijections from $\mathbb{D}$ to Ω, not the reverse).

XII.8.5.8. The boundary $\partial\Omega$ can be much worse, even if locally connected. It is useful to make the following distinction. If Ω is a bounded simply connected open set in $\mathbb{C}$, let Λ be the interior of $\bar{\Omega}$. Then Λ is a regular open set, $\Omega \subseteq \Lambda$, and $\bar{\Lambda} = \bar{\Omega}$, so Ω is dense in Λ and $\partial\Lambda \subseteq \partial\Omega$.

XII.8.5.9. DEFINITION. $\partial\Lambda$ is called the *regular boundary* of Ω, denoted $\partial_r\Omega$. The complement $\partial\Omega \setminus \partial_r\Omega = \Lambda \cap \partial\Omega$ is called the *inner boundary* of Ω, denoted $\partial_i\Omega$.

XII.8.5.10. EXAMPLES. (i) If Ω is a regular open set, e.g. $\mathbb{D}$ or, more generally, the interior of a Jordan curve, or even the interior of the Warsaw Circle, then $\Lambda = \Omega$ so $\partial_r\Omega = \partial\Omega$ and $\partial_i\Omega = \emptyset$.

(ii) We can have $\partial_r\Omega = \partial\Omega$ even if Ω is pathological, e.g. if Ω is one of the Lakes of Wada.

(iii) Let Ω be the slit disk $\mathbb{D} \setminus [0,1)$. Then $\Lambda = \mathbb{D}$, $\partial_r\Omega$ is $\mathbb{T}$, and $\partial_i\Omega = [0,1)$.

(iv) $\partial_r\Omega$ can be nice even if $\partial\Omega$ is bad. Let Ω be the interior of (); then $\partial_r\Omega$ is the bounding square.

(v) Let Ω be the annulus $\{z : 1 < |z| < 2\}$ with the open segment $(1,2)$ removed. Then Ω is simply connected, and $\partial\Omega$ consists of the two boundary circles and the segment $(1,2)$. Λ is the entire annulus, and thus is not simply connected; $\partial_r\Omega$ consists of the two boundary circles, and $\partial_i\Omega$ the segment $(1,2)$.

(vi) Let C be the Cantor set () as a subset of $[0,1]$. Remove from $\mathbb{D}$ the points of C, and for each open interval I in $[0,1]$ removed in the construction of C, also remove the closed disk with I as diameter. See Figure (). Let Ω be the remainder. Then Ω is a simply connected regular open set, so $\partial_r\Omega = \partial\Omega$ consists of $\mathbb{T}$ along with C and the circles bounding the removed disks. Thus $\partial\Omega$ is locally connected but looks nothing like a dendrite (). This example can be extended by removing small copies of the removed set radially from a dense set of points in $\mathbb{T}$ and densely around all boundary circles of the removed disks, so $\partial\Omega$ is not locally a dendrite around any point.

(vii) Remove from $\mathbb{D}$ the segment $[0,1)$ and the closed disk of radius $\frac{1}{2(n+1)^2}$ around $\frac{1}{n}$ for every $n \geq 2$. Then Ω is simply connected, and $\partial\Omega$ is locally connected. Λ consists of $\mathbb{D}$ with 0 and a sequence of disjoint closed disks converging to 0 removed. $\partial_r\Omega$ consists of $\mathbb{T}$ along with 0 and the boundary circles of the removed disks, and thus is not locally connected at 0. $\partial_i\Omega$ consists of the open segments on the real axis between the removed disks.

These examples show that $\partial\Omega$, even if locally connected, need not look anything like a Jordan curve. See XII.8.6.7. for more properties.

The Jordan Curve Case

XII.8.5.11. Let Ω be a bounded simply connected open set in $\mathbb{C}$, and f a conformal bijection from $\mathbb{D}$ to Ω. If f extends to an injective (hence bijective) continuous function (hence a homeomorphism) from $\bar{\mathbb{D}}$ to $\bar{\Omega}$, then $\bar{f}$ gives a homeomorphism from $\mathbb{T}$ to $\partial\Omega$, i.e. $\partial\Omega$ is a Jordan curve () in $\mathbb{C}$. Conversely:

XII.8.5.12. THEOREM. Let Ω be a bounded simply connected open set in $\mathbb{C}$, and f a conformal bijection from $\mathbb{D}$ to Ω. Then f extends to a continuous bijection $\bar{f} : \bar{\mathbb{D}} \to \bar{\Omega}$ if and only if $\partial\Omega$ is a Jordan curve.

This can be deduced from XII.8.5.6.; for more direct proofs see e.g. [Rud87] or [GK06].

Uniqueness

We can refine XII.8.5.12. to show an existence and uniqueness result once three points on the boundary are specified:

XII.8.5.13. THEOREM. Let Ω be a bounded simply connected open set in $\mathbb{C}$ with $\partial\Omega$ a Jordan curve. Let $\{a,b,c\}$ be three distinct points in $\partial\Omega$ in counterclockwise order. If $\{a',b',c'\}$ are three distinct points in $\mathbb{T}$ in counterclockwise order, then there is a unique homeomorphism $\bar{f}$ from $\bar{\mathbb{D}}$ onto $\bar{\Omega}$ such that $f = \bar{f}|_\Omega$ is a conformal bijection from $\mathbb{D}$ onto Ω and $\bar{f}(a') = a$, $\bar{f}(b') = b$, $\bar{f}(c') = c$.

PROOF: There is a conformal bijection g from $\mathbb{D}$ to Ω, which extends to a homeomorphism $\bar{g} : \bar{\mathbb{D}} \to \bar{\Omega}$. Set $a'' = \bar{g}^{-1}(a)$, $b'' = \bar{g}^{-1}(b)$, $c'' = \bar{g}^{-1}(c)$. Then $\{a'', b'', c''\}$ are in counterclockwise order on $\mathbb{T}$. There is a (unique) Möbius transformation h sending $\mathbb{D}$ to $\mathbb{D}$ and sending $\{a', b', c'\}$ to $\{a'', b'', c''\}$ (). Then $\bar{f} = \bar{g} \circ h$ is the desired map. For uniqueness, if $\bar{f}_1$ and $\bar{f}_2$ are such functions, then $f_2^{-1} \circ f_1$ is a Möbius transformation from $\mathbb{D}$ to $\mathbb{D}$ which extends to a neighborhood and leaves a', b', c' fixed, hence is the identity. 👍

The Smooth Case

XII.8.5.14. THEOREM. Let Ω be a bounded simply connected open set in $\mathbb{C}$, and f a conformal bijection from $\mathbb{D}$ to Ω. Then f extends to a $\mathcal{C}^\infty$-diffeomorphism $\bar{f} : \bar{\mathbb{D}} \to \bar{\Omega}$ if and only if $\partial\Omega$ is a $\mathcal{C}^\infty$ Jordan curve.

One direction is obvious; for the other direction only smoothness of the extension is an issue. The proof of this direction is not easy [BK87]. (One has the feeling there should be a simpler proof.)

XII.8.5.15. It is much easier to show using the Schwarz Reflection Principle that if $\partial\Omega$ is an analytic Jordan curve, then the extension is analytic. See ().

XII.8.6. Exercises

XII.8.6.1. Let (Ω_n) be an increasing sequence of open subsets of $\mathbb{C}$, and $\Omega = \cup_n \Omega_n$. Show that if each Ω_n is simply connected, so is Ω. [Any closed path in Ω is contained in some Ω_n by compactness.]

XII.8.6.2. Let f be a bounded analytic function on $\mathbb{D}$, with $|f(z)| \leq M$ for all $z \in \mathbb{D}$. Let

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} z^k$$

be the Taylor expansion around 0.

(a) Let $0 < r < 1$. Integrate $|f(z)|^2$ around the circle of radius r around 0 to obtain

$$\sum_{k=0}^{\infty} \left[\left| \frac{f^{(k)}(0)}{k!} \right|^2 r^k \right] \leq M^2 \ .$$

[Justify term-by-term integration and show the cross terms integrate to 0.]

(b) Let $r \to 1$ to conclude that

$$\sum_{k=0}^{\infty} \left| \frac{f^{(k)}(0)}{k!} \right|^2 \leq M^2 \ .$$

(c) Conclude from (b) that for any k, $|f^{(k)}(0)| \leq k!M$, with equality if and only if $f(z) = cz^k$ for some c, $|c| = M$.

XII.8.6.3. Fractional Linear Transformations. Let $a, b, c, d \in \mathbb{C}$, $ad - bc \neq 0$. A function

$$T_{a,b,c,d}(z) = \frac{az + b}{cz + d}$$

is called a *fractional linear transformation* or *Möbius transformation.* If $c = 0$, $T_{a,b,c,d}$ is an affine function from $\mathbb{C}$ to $\mathbb{C}$.

(a) Show that we may assume $ad - bc = 1$ without loss of generality (it is sometimes not convenient to do this, e.g. in the second part of (i)).

(b) $T_{a,b,c,d}$ is a meromorphic function whose derivative is never zero, hence a conformal map. If $c \neq 0$, it has a simple pole at $-\frac{d}{c}$ (if $c = 0$ it is an entire function). If $c \neq 0$, then

$$\lim_{z \to \infty} T_{a,b,c,d}(z) = \frac{a}{c}$$

(c) A composition of two fractional linear transformations is a fractional linear transformation (if a possible removable singularity is filled in).

(d) If $ad - bc = 1$, then $T_{a,b,c,d} \circ T_{d,-b,-c,a}$ and $T_{d,-b,-c,a} \circ T_{a,b,c,d}$ become the identity map on $\mathbb{C}$ when a removable singularity is filled in (if $c \neq 0$). Thus $T_{a,b,c,d}$ is a conformal equivalence from $\mathbb{C} \setminus \left\{-\frac{d}{c}\right\}$ to $\mathbb{C} \setminus \left\{\frac{a}{c}\right\}$ (if $c \neq 0$; otherwise it is a conformal bijection from $\mathbb{C}$ to $\mathbb{C}$), and the set of fractional linear transformations (with removable singularities removed) forms a group under composition.

(e) The map

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \mapsto T_{a,b,c,d}$$

is a group-homomorphism from $SL_2(\mathbb{C})$ onto the group of fractional linear transformations with kernel $\{\pm I\}$, thus drops to an isomorphism from $PSL_2(\mathbb{C}) = SL_2(\mathbb{C})/\{\pm I\}$.

(f) A fractional linear transformation $T_{a,b,c,d}$ can be thought of as a conformal map from the Riemann sphere $\mathbb{S}$ () to $\mathbb{S}$, sending $-\frac{d}{c}$ to ∞ and ∞ to $\frac{a}{c}$, or sending ∞ to ∞ if $c = 0$. (When thought of this way one does not have to worry about removable singularities.)

(g) A fractional linear transformation sends lines and circles to lines and circles (circles to circles if a line is regarded as a circle through ∞).

(h) If $\{z_1, z_2, z_3\}$ and $\{w_1, w_2, w_3\}$ are two sets of distinct points in $\mathbb{S}$, then there is a unique fractional linear transformation T with $T(z_j) = w_j$ for $1 \leq j \leq 3$.

(i) $T_{a,b,c,d}$ $(ad - bc = 1)$ sends $\mathbb{R} \cup \{\infty\}$ to $\mathbb{R} \cup \{\infty\}$ if and only if $a, b, c, d \in \mathbb{R}$. Determine when it sends the upper half plane to the upper half plane. T sends $\mathbb{D}$ onto $\mathbb{D}$ if and only if it is of the form

$$T(z) = \alpha \frac{z - z_0}{1 - \bar{z}_0 z}$$

for some $z_0 \in \mathbb{D}$ and $|\alpha| = 1$.

(j) Every conformal bijection from $\mathbb{D}$ to $\mathbb{D}$ is a fractional linear transformation. The same is true between any Ω and Λ each conformally equivalent to $\mathbb{D}$ via a fractional linear transformation (e.g. a half plane).

Caution: In some texts, fractional linear transformations (or "linear fractional transformations") are called "linear transformations." But most are not linear in the linear algebra sense, or even affine. More commonly, fractional linear transformations are called "bilinear," which is also not good terminology since they are not bilinear in any sense similar to III.11.13.1..

XII.8.6.4. Use fractional linear transformations to prove the following generalization of the Schwarz Lemma (*Pick's Theorem*):

THEOREM. Let f be an analytic function from $\mathbb{D}$ to $\mathbb{D}$, and $z_0 \in \mathbb{D}$. Then either f is a conformal bijection (fractional linear transformation of the form in XII.8.6.3.(i)), or, for all $z \in \mathbb{D}$, $z \neq z_0$,

$$\left|\frac{f(z)-f(z_0)}{1-\overline{f(z_0)}f(z)}\right| < \left|\frac{z-z_0}{1-\bar{z}_0 z}\right|$$

and

$$|f'(z_0)| < \frac{1-|f(z_0)|^2}{1-|z_0|^2} < \frac{1}{1-|z_0|^2}\ .$$

More generally, if f is an analytic function on $\{z : |z| < R\}$ and $|f(z)| < S$ for all such z, and $|z_0| < R$, then either f is a conformal bijection (which must be a fractional linear transformation), or

$$\left|S\frac{f(z)-f(z_0)}{S^2-\overline{f(z_0)}f(z)}\right| < \left|R\frac{z-z_0}{R^2-\bar{z}_0 z}\right|$$

for all z with $|z| < R$, $z \neq z_0$, and

$$|f'(z_0)| < \frac{R}{S}\frac{S^2-|f(z_0)|^2}{R^2-|z_0|^2}\ .$$

XII.8.6.5. Conformally Rigid Sets. Let Ω be $\mathbb{D}$ with finitely many points removed.

(a) Show that any conformal equivalence f from Ω to Ω extends to a conformal equivalence from $\mathbb{D}$ to $\mathbb{D}$, hence is a fractional linear transformation. [All the singularities are removable.]

(b) If Ω has three points z_1, z_2, z_3 removed, then f must permute $\{z_1, z_2, z_3\}$. There is at most one such f for any such permutation. [Use XII.8.6.3..] If z_1, z_2, z_3 are collinear, or lie on a circle not contained in $\mathbb{D}$, the permutation can only be the identity or the one interchanging z_1 and z_3.

(c) In the last case of (b), give a necessary and sufficient condition that there is an f that reverses order. This condition is rarely satisfied, and if not, the only conformal equivalence of Ω with itself is the identity. In particular, the only conformal equivalence of $\mathbb{D} \setminus \{0, 1/2, 3/4)\}$ with itself is the identity. (If Ω is $\mathbb{D}$ with more than three points removed, it is even harder for Ω to have a self-equivalence other than the identity.)

(d) More generally, two regions consisting of $\mathbb{D}$ with finitely many points removed are rarely conformally equivalent (for one thing, they cannot be if different numbers of points have been removed – they are not even homeomorphic!), and a conformal equivalence, if it exists, is usually unique. Give a necessary and sufficient condition that two regions consisting of $\mathbb{D}$ with three points removed are conformally equivalent. Contrast with the case where only two points have been removed.

XII.8.6.6. Let Ω and Λ be bounded simply connected open sets in $\mathbb{C}$ (so there is a homeomorphism, even a conformal bijection, from Ω to Λ).

(a) If $\partial\Omega$ is homeomorphic to $\partial\Lambda$, are $\bar{\Omega}$ and $\bar{\Lambda}$ necessarily homeomorphic under a homeomorphism sending Ω to Λ? (Taking Ω and Λ to be two of the Lakes of Wada may give a counterexample depending on how the construction is done.)

(b) If there is a homeomorphism from $\bar{\Omega}$ to $\bar{\Lambda}$ sending Ω to Λ (and hence sending $\partial\Omega$ to $\partial\Lambda$), does a conformal bijection from Ω to Λ necessarily extend to $\bar{\Omega}$? If the answer is no in general, give necessary and sufficient conditions for an extension.

(c) What if Ω and Λ are connected but not simply connected? (If Ω and Λ are not connected, the answer to (a) is no in general: let Ω be one of the Lakes of Wada and Λ the union of two of the lakes. And if the answer to (a) is yes, there is not a conformal bijection from Ω to Λ in general (XII.8.6.5.(d) or annuli XII.8.2.4.). The question in (b) asks: if there is a homeomorphism from $\bar{\Omega}$ to $\bar{\Lambda}$ sending Ω to Λ *and* there is a conformal bijection from Ω to Λ, does it extend?)

XII.8.6.7. Let Ω be a bounded simply connected open set in $\mathbb{C}$, and $\partial_r\Omega$ and $\partial_i\Omega$ as in XII.8.5.9.. Then $\partial\Omega$ is the disjoint union of $\partial_r\Omega$ and $\partial_i\Omega$; $\partial_r\Omega$ is closed, but $\partial_i\Omega$, if nonempty, is not closed.

(a) Show that $\partial_i\Omega$ is locally connected. [It is relatively open in $\partial\Omega$.] Note that $\partial_r\Omega$ need not be locally connected (XII.8.5.10.(vii)).

(b) Show that $\partial_i\Omega$ can contain no Jordan curves. [Ω is connected.] Note that $\partial_r\Omega$ can contain Jordan curves, even of arbitrarily small diameter (XII.8.5.10.(vi)).

(c) Let X be a connected (hence path-connected) component of $\partial_i\Omega$. Show that X is not closed, and $\bar{X} \setminus X$ consists of either one or two points of $\partial_r\Omega$ and if it is two, they must be in different connected components of $\partial_r\Omega$. Conclude that $\bar{X}$ is a dendrite ().

XII.8.6.8. Use the Jordan Curve Theorem XIII.19.1.20., Theorem XII.8.3.2.[(x) $\Rightarrow$ (iv)], and Theorem XII.8.5.12. to prove the Jordan-Schönflies Theorem XIII.19.1.21.. [Some additional argument is needed to match up the homeomorphisms on the inside and outside of the curve.]

XII.8.6.9. If X is a connected topological space, a point $p \in X$ is a *cut point* of X if $X \setminus \{p\}$ is not connected.

(a) If Ω is a bounded simply connected open set in $\mathbb{C}$, show that $\partial\Omega$ is a Jordan curve if and only if it is locally connected and has no cut points.

(b) There are one-dimensional continua in the plane which are locally connected with no cut points which are not Jordan curves, for example the letter θ. There are even one-dimensional continua in the plane which are locally connected with no "local cut points" which are not Jordan curves, such as the Sierpiński Gasket (XIII.19.5.10.) and Sierpiński Carpet (XIII.19.5.11.).

XII.8.6.10. Adapt the arguments in this section to the Riemann sphere to handle boundary behavior if Ω is unbounded.

XII.9. Analytic Continuation and Riemann Surfaces

In this section we will assume all open sets we work on are connected. We will use the term *region* to mean a nonempty connected open subset of $\mathbb{C}$.

XII.9.1. Analytic Function Elements

The Rigidity Theorem implies that if Ω and $\tilde{\Omega}$ are regions in $\mathbb{C}$, whose intersection is nonempty, and f is holomorphic on Ω, then there is at most one function $\tilde{f}$ which is holomorphic on $\tilde{\Omega}$ and agrees with f on the intersection. There may or may not be such an $\tilde{f}$; but if there is, f and $\tilde{f}$ can be reasonably considered to be "pieces" of the same function which is holomorphic on some unspecified domain containing both Ω and $\tilde{\Omega}$.

Things can be more complicated, however. The intersection of Ω and $\tilde{\Omega}$ is not necessarily connected, and the functions f and $\tilde{f}$ can agree on part of the intersection but not all. An example is two branches of $\log z$ () corresponding to α and β with $|\alpha - \beta| < 2\pi$.

A related difficulty which actually provides a solution to the first difficulty is that if Ω_1, Ω_2, and Ω_3 are regions, and f_j is holomorphic on Ω_j, if $f_1 = f_2$ on $\Omega_1 \cap \Omega_2$ and $f_2 = f_3$ on $\Omega_2 \cap \Omega_3$, then f_1 and f_3 should be regarded as "pieces" of the same function even if Ω_1 and Ω_3 are disjoint, or if they overlap but f_1 and f_3 do not agree on $\Omega_1 \cap \Omega_3$, as can happen, e.g. with suitable branches of the logarithm. The solution is to regard f on Ω and $\tilde{f}$ on $\tilde{\Omega}$ as "pieces" of the same function if there is a finite chain of intermediate "pieces" f_j on Ω_j which agree on consecutive overlaps. Using this extension of equivalence, we may regard any two branches of $\log z$, even ones whose angles differ by more than 2π, to be "pieces" of the same big function. The domain of this big function is not a subset of $\mathbb{C}$, but a large abstract set made up of overlapping pieces which can be identified with regions in $\mathbb{C}$ (technically a complex 1-manifold ()). This domain is called the *Riemann surface* of the function.

We now discuss the details.

XII.9.1.1. Definition. An *analytic function element* is a pair (Ω, f), where Ω is a region (nonempty connected open set) in $\mathbb{C}$ and f is holomorphic on Ω.

Two analytic function elements (Ω, f) and $(\tilde{\Omega}, \tilde{f})$ are *equivalent* if there is a finite chain (Ω_j, f_j) $(1 \le j \le n)$ such that $(\Omega_1, f_1) = (\Omega, f)$, $(\Omega_n, f_n) = (\tilde{\Omega}, \tilde{f})$, and $\Omega_j \cap \Omega_{j+1} \neq \emptyset$ and $f_j = f_{j+1}$ on $\Omega_j \cap \Omega_{j+1}$ for $1 \le j < n$.

A *total analytic function* is an equivalence class of analytic function elements.

XII.9.1.2. It is obvious that equivalence of analytic function elements is indeed an equivalence relation. Every analytic function element determines a unique total analytic function, namely its equivalence class. Equivalent analytic function elements determine the same total analytic function.

XII.9.1.3. Proposition. (i) If (Ω, f) and $(\tilde{\Omega}, \tilde{f})$ are analytic function elements, $\Omega \cap \tilde{\Omega} \neq \emptyset$, and $f = \tilde{f}$ on $\Omega \cap \tilde{\Omega}$, then (Ω, f) and $(\tilde{\Omega}, \tilde{f})$ are equivalent.

(ii) If (Ω, f) is an analytic function element, $\tilde{\Omega}$ a subregion of Ω, and $\tilde{f} = f|_{\tilde{\Omega}}$, then the analytic function element $(\tilde{\Omega}, \tilde{f})$ is equivalent to (Ω, f).

(ii) If $\{(\Omega_k, f_k) : k \in K\}$ are analytic function elements, such that $f_j = f_k$ on $\Omega_j \cap \Omega_k$ for all j and k, and $\Omega = \cup_k \Omega_k$ is connected (e.g. if $\cap_k \Omega_k \neq \emptyset$), set $f = f_k$ on Ω_k for each k. Then f is a well-defined holomorphic function on Ω, and the analytic function element (Ω, f) is equivalent to each (Ω_k, f_k).

XII.9.2. The Riemann Surface of a Total Analytic Function

XII.9.2.1. We can put the domains of the elements of a total analytic function together to get a "surface" which is the natural domain of the function. Informally, we take the "union" of the domains of the elements, with points in overlapping elements identified *provided* the function agrees around these points. More precisely:

XII.9.2.2. DEFINITION. Let $\boldsymbol{f} = \{(\Omega_k, f_k) : k \in K\}$ be a total analytic function, i.e. the (Ω_k, f_k) are equivalent analytic function elements. Let D_f be the disjoint union () of the Ω_k (even if $\Omega_j = \Omega_k$ as subsets of $\mathbb{C}$, if $j \neq k$ they are disjoint subsets of D_f). Put an equivalence relation on D_f as follows. If $z_j \in \Omega_j$, $z_k \in \Omega_k$, set $z_j \sim z_k$ if $z_j = z_k \in \Omega_j \cap \Omega_k \subseteq \mathbb{C}$ and $f_j(z) = f_k(z)$ for all z in a neighborhood of $z_j = z_k$. Write $\hat{\Omega}_k = \{[z] : z \in \Omega_k\}$. Let S_f be the set of equivalence classes. S_f is the *Riemann surface* of the total analytic function $\boldsymbol{f}$.

The total analytic function $\boldsymbol{f}$ gives a well-defined function f from S_f to $\mathbb{C}$ by $f([z]) = f_k(z)$ for $z \in \Omega_k$. The function f is called the *(complete) analytic function* defined by $\boldsymbol{f}$. Each (Ω_k, f_k) is a *branch* of f or $\boldsymbol{f}$.

XII.9.2.3. There is a natural injective map ι_k from Ω_k to S_f for each k, and the restriction of f to the image $\hat{\Omega}_k$ of Ω_k is exactly f_k.

We topologize S_f by taking the $\hat{\Omega}_k$ as a base for the topology. Then each $\hat{\Omega}_k$ is an open subset of S_f, and the inclusion of Ω_k into S_f is a homeomorphism onto its image. The function $f : S_f \to \mathbb{C}$ is continuous.

XII.9.2.4. The natural identifications of the subsets $\hat{\Omega}_k$ of S_f with the corresponding subsets Ω_k of $\mathbb{C}$ give an atlas defining the structure of a complex analytic 1-manifold on S_f (). This atlas has the special property that the transition functions between charts are identity maps on open sets in $\mathbb{C}$. (This natural atlas is not a maximal atlas.)

The next proposition is an immediate corollary of ().

XII.9.2.5. PROPOSITION. Let $\boldsymbol{f}$ be a total analytic function, and (Ω_j, f_j) and (Ω_k, f_k) analytic function elements in $\boldsymbol{f}$. If $z_0 \in \Omega_j \cap \Omega_k$, then either $\iota_j(z) = \iota_k(z)$ for all z in a neighborhood of z_0 (if $\iota_j(z_0) = \iota_k(z_0)$), or $f_j(z) \neq f_k(z)$ for all z in a deleted neighborhood of z_0 and in particular $\iota_j(z) \neq \iota_k(z)$ for all z in a neighborhood of z_0 (if $\iota_j(z_0) \neq \iota_k(z_0)$).

XII.9.2.6. COROLLARY. Let $\boldsymbol{f}$ be a total analytic function with Riemann surface S_f. Then S_f is Hausdorff.

XII.9.2.7. PROPOSITION. Let $\boldsymbol{f}$ be a total analytic function with Riemann surface S_f. Then S_f is path-connected.

PROOF: Suppose (Ω_j, f_j) and $\Omega_k, f_k)$ are elements of $\boldsymbol{f}$, and $z_j \in \Omega_j$, $z_k \in \Omega_k$. Let $(\Omega_1, f_1), \ldots, (\Omega_n, f_n)$ be a chain with $(\Omega_1, f_1) = (\Omega_j, f_j)$, $(\Omega_n, f_n) = (\Omega_k, f_k)$, and $\Omega_m \cap \Omega_{m+1} \neq \emptyset$ and $f_m = f_{m+1}$ on $\Omega_m \cap \Omega_{m+1}$ for $1 \leq m < n$. For each m, $1 \leq m \leq n-1$, choose $z_m \in \Omega_m \cap \Omega_{m+1}$. We have $\iota_m(z_m) = \iota_{m+1}(z_m)$ for each

m. There is a path from z_j to z_1 in Ω_1, a path from z_m to z_{m+1} in Ω_{m+1} for $1 \leq m \leq n-2$, and a path from z_{n-1} to z_k in Ω_n. The corresponding paths in the $\hat{\Omega}_m$ match up at the endpoints and give a path in S_f from $\iota_j(z_j)$ to $\iota_k(z_k)$. 👍

XII.9.2.8. EXAMPLES. (i) The Riemann surface of a total analytic function can sometimes be identified with a subset of $\mathbb{C}$. The simplest case is when f is an entire function. Then $(\mathbb{C}, f)$ is an analytic function element, and the analytic function elements equivalent to it are precisely of the form $(\Omega, f|_\Omega)$, where Ω is a region in $\mathbb{C}$. Since the functions in any two elements of the equivalence class agree on the intersection of the regions, every region in S_f is identified with itself as a subset of $\mathbb{C}$. Thus $S_f \cong \mathbb{C}$.

(ii) Similarly, if f is holomorphic on $\mathbb{C}$ except for isolated non-removable singularities, the set P of (non-removable) singularities is a countable subset of $\mathbb{C}$ with no limit point. Then $(\mathbb{C} \setminus P, f)$ is an analytic function element, and as before every equivalent analytic function element is of the form $(\Omega, f|_\Omega)$, where Ω is a region in $\mathbb{C} \setminus P$, and $S_f \cong \mathbb{C} \setminus P$. (If f also has removable singularities, then S_f is the domain of f with the removable singularities removed.)

(iii) Here is one of the most important examples, of a total analytic function whose Riemann surface is not a subset of $\mathbb{C}$. If Ω is $\mathbb{C}$ with a branch cut (ray out from 0) removed (), then there are infinitely many branches of $\log z$ defined on Ω. For each such f, (Ω, f) is an analytic function element, and the copies of Ω corresponding to different branches are disjoint. If $\tilde{\Omega}$ is $\mathbb{C}$ with a different branch cut removed, then $\Omega \cap \tilde{\Omega}$ is not connected, and the pieces are identified with parts of different copies of Ω in S_f. The Riemann surface for $\log z$ can be visualized by taking a sequence of copies of Ω, and connecting each sheet along the branch cut with the sheet above, giving a helical surface (Figure XII.1, from http://upload.wikimedia.org/wikipedia/commons/4/41/Riemann_surface_log.jpg). Choosing a branch cut just allows the surface to be constructed or visualized in this way; the branch cut has no intrinsic significance as part of the Riemann surface, and any other branch cut would work as well and give the same surface.

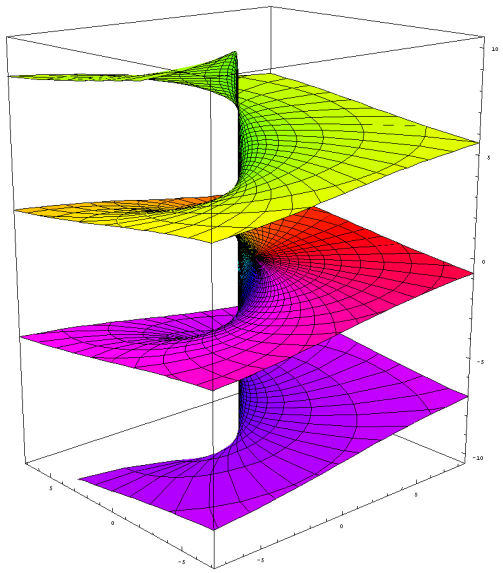


Figure XII.1: Riemann Surface for $\log z$ (Three levels shown)

(iv) Similarly, the Riemann surface for $\sqrt{z}$ can be constructed by again taking a branch cut out from 0. This time, there are only two branches for each branch cut. The Riemann surface is obtained by gluing each copy of the plane with branch cut removed to the other copy along the branch cut (this cannot be physically accomplished in $\mathbb{R}^3$ since the surfaces would have to pass through each other along the branch cut, but can be done in $\mathbb{R}^4$ or $\mathbb{C}^2$; see Figure XII.2, from http://en.wikipedia.org/wiki/File:Riemann_sqrt.jpg).

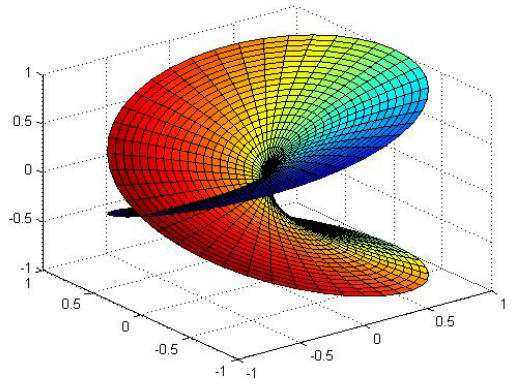


Figure XII.2: Riemann Surface for $\sqrt{z}$

The Riemann surface for $\sqrt[n]{z}$ is similar, with n sheets, each attached to the next one along the branch cut and the last attached to the first.

(v) If (Ω_j, f_j) and (Ω_k, f_k) are branches of a total analytic function $\boldsymbol{f}$, and $z \in \Omega_j \cap \Omega_k$, we may have $\iota_j(z) \neq \iota_k(z)$ even though $f_j(z) = f_k(z)$: the function with elements $(\Omega, g(z)[g(z) - 2\pi i])$, where Ω is a simply connected open set in $\mathbb{C} \setminus \{0\}$ and g is a branch of $\log z$ holomorphic on Ω, has two distinct branches over neighborhoods of 1 both taking the value 0 at 1. (These branches differ at all points of a deleted neighborhood of 1.)

XII.9.2.9. If $\boldsymbol{f}$ is a total analytic function with Riemann surface S_f, there is a projection map p_f from S_f to $\mathbb{C}$ which sends each $\hat{\Omega}$ on the surface to the corresponding subset Ω of $\mathbb{C}$. This projection is continuous and open, its range is an open subset of $\mathbb{C}$, and may be either one-to-one (e.g. in cases XII.9.2.8.(i)–(ii)) or many-to-one (e.g. in cases XII.9.2.8.(iii)–(iv)). If p_f is one-to-one, it is a homeomorphism onto its image, and S_f can be identified with an open set in $\mathbb{C}$; in this case, $\boldsymbol{f}$ is said to be *single-valued*; if p_f is not one-to-one, $\boldsymbol{f}$ is said to be *multivalued*.

In cases XII.9.2.8.(iii)–(v), the map p_f is a covering map, but this is not true in general, and the preimage of different points can even have different cardinality () (each point preimage is always finite or countable since it is relatively discrete in S_f ()). The range of p_f is always a region in $\mathbb{C}$, but it may be quite small ().

XII.9.2.10. Any analytic function element (Ω, f) defines a unique total analytic function $\boldsymbol{f}$ and hence a Riemann surface S_f, which is the "largest domain" f can be extended to as a holomorphic function. But if f is defined on a small open set Ω, e.g. on a finite disk by a power series with finite radius of convergence, it can

be difficult if not impossible to determine the full Riemann surface of f, or indeed whether $\boldsymbol{f}$ is single-valued or multivalued.

XII.9.3. Analytic Continuation

XII.9.3.1. We now change our point of view slightly, and discuss the problem from XII.9.2.10. in more detail. Suppose f is a holomorphic function on a region Ω (i.e. (Ω, f) is an analytic function element). How far outside Ω can f be extended as a holomorphic function, and to what degree is the extension unique?

XII.9.3.2. The first question to discuss is: if z is a boundary point of Ω, can a (necessarily unique) holomorphic function g be defined on an open disk centered at z which agrees with f on the part of the disk contained in Ω? If not, z is called a *singular point* of the boundary (for f). The set of singular points is a closed subset of the boundary of Ω. In some cases, it can be shown that there must be singular boundary points, e.g.:

XII.9.3.3. PROPOSITION. Let

$$\sum_{k=0}^{\infty} a_k (z - z_0)^k$$

be a power series with radius of convergence R, $0 < R < \infty$, and let Ω be the open disk of radius R centered at z_0. If f is the holomorphic function on Ω defined by the power series, then there is at least one singular point for f on the boundary of Ω.

PROOF: Suppose not. Then for any z on the boundary circle there is an open disk D_z of radius ϵ_z centered at z and a holomorphic function g_z on this disk agreeing with f on the part of the disk inside Ω. By the Rigidity Theorem, any two g_z must agree on the intersection of the D_z. Thus there is a well-defined holomorphic function g on $U = \Omega \cup \cup_z D_z$ which extends f. But U is an open neighborhood of $\bar{\Omega}$, which is compact, so U contains a disk of radius $R + \epsilon$ centered at z_0, for some $\epsilon > 0$ (). Then by XII.4.4.6. the Taylor series for f around z_0 has radius of convergence at least $R + \epsilon$, a contradiction. 👍

XII.9.3.4. EXAMPLES. (i) There may be only one singular boundary point. The power series

$$\sum_{k=0}^{\infty} z^k$$

converges to $f(z) = \frac{1}{1-z}$ on the open unit disk $\mathbb{D}$. This function can be extended holomorphically to every boundary point of $\mathbb{D}$ except $z = 1$. In fact, it extends to a function holomorphic on $\mathbb{C} \setminus \{1\}$.

(ii) At the other extreme, all boundary points can be singular points. The power series

$$\sum_{k=0}^{\infty} z^{2^k}$$

has radius of convergence 1, hence defines a holomorphic function on $\mathbb{D}$. But if θ is a dyadic rational multiple of 2π, then it is easily checked that

$$\lim_{r \to 1-} |f(re^{i\theta})| = +\infty$$

so $e^{i\theta}$ is a singular point of the boundary. Since such points are dense in the unit circle and the set of singular points is closed, every boundary point is a singular point.

(iii) Here is an interesting variation of (ii), a holomorphic function on $\mathbb{D}$ which can be extended continuously, even smoothly, to the boundary, but for which all boundary points are nonetheless singular. The power series

$$\sum_{k=1}^{\infty} 2^{-k^2} z^{2^k}$$

has radius of convergence 1 and converges uniformly on the closed unit disk $\bar{\mathbb{D}}$, as do the derived series of all orders. Thus the function

$$f(z) = 1 + 2z + \sum_{k=1}^{\infty} 2^{-k^2} z^{2^k}$$

is a smooth function (in the vector calculus sense) on $\bar{\mathbb{D}}$ which is holomorphic on $\mathbb{D}$. But nonetheless all boundary points are singular by the Hadamard gap theorem. (It turns out that f is one-to-one on $\bar{\mathbb{D}}$, the reason for including the $2z$ term. Including the constant term 1 is conventional.)

Examples (ii) and (iii) are *lacunary series*, power series where the gaps between the indices of consecutive nonzero coefficients become arbitrarily large. There is a discussion of these and related series in [Rem98], where the comment is made that "The gap theorem is in some sense a paradox: power series that, because of their gaps, converge especially fast in the interior of their disk of convergence have singularities almost everywhere on the boundary precisely on account of these gaps."

We now consider analytic continuation along a path.

XII.9.3.5. DEFINITION. Let Ω be a region in $\mathbb{C}$, and f a holomorphic function on Ω. Let γ a path in $\mathbb{C}$ from z_0 to z_1, with $z_0 \in \Omega$, parametrized by $\zeta : [a, b] \to \mathbb{C}$. Then f can be *analytically continued along* γ if, for every $t \in [a, b]$, there is an open disk D_t centered at $\zeta(t)$ and a function f_t holomorphic on D_t such that, for every $t_0 \in [a, b]$, $f_t = f_{t_0}$ on $D_t \cap D_{t_0}$ for all t sufficiently close to t_0, and such that $f_a = f|_{D_a}$.

The essential point is that the path γ begins in Ω but may pass or end outside Ω. In fact this is the only interesting situation: if γ lies inside Ω, then f can be analytically continued along γ simply by restriction.

If $\{(D_t, f_t) : t \in [a, b]\}$ is an analytic continuation of f along γ, then each (D_t, f_t) is an analytic function element.

XII.9.3.6. PROPOSITION. If $\{(D_t, f_t) : t \in [a, b]\}$ is an analytic continuation of f along γ, all analytic function elements (D_t, f_t) are equivalent.

PROOF: It is clear that for any t_0, the function elements (D_t, f_t) are equivalent to (D_{t_0}, f_{t_0}) for t sufficiently close to t_0. Thus the set of t in each equivalence class is open in $[a, b]$. Since the complement of an equivalence class is the union of the other equivalence classes, the set of t in each equivalence class is also closed in $[a, b]$. Thus there can be only one equivalence class since $[a, b]$ is connected. 👍

XII.9.3.7. Analytic continuation along a path γ can be alternately considered a finite process: if $\{(D_t, f_t)\}$ is an analytic continuation, there is a finite partition $\{a = t_0, t_1, \ldots, t_n = b\}$ of $[a, b]$ such that $\zeta(t) \in D_{t_k} \cup D_{t_{k+1}}$ for $t_k \leq t \leq t_{k+1}$, and $f_{t_k} = f_{t_{k+1}}$ on $D_{t_k} \cap D_{t_{k+1}}$ for $0 \leq k < n$. Conversely, if there is such a partition, disks D_k centered at ζ_{t_k} such that $\zeta(t) \in D_{t_k} \cup D_{t_{k+1}}$ for $t_k \leq t \leq t_{k+1}$, and functions f_k holomorphic on D_k agreeing on the consecutive overlaps, then (D_t, f_t) can be defined for each t by restricting f_k or f_{k+1} for $t_k \leq t \leq t_{k+1}$ to a disk D_t centered at $\zeta(t)$ contained in $D_k \cup D_{k+1}$.

XII.9.3.8. PROPOSITION. Let Ω be a region in $\mathbb{C}$, and f a holomorphic function on Ω. Let γ a path in $\mathbb{C}$ from z_0 to z_1, with $z_0 \in \Omega$, parametrized by $\zeta : [a, b] \to \mathbb{C}$. Then an analytic continuation of f along γ, if it exists, is unique: if $\{(D_t, f_t)\}$ and $\{\tilde{D}_t, g_t)\}$ are analytic continuations of f along γ, then $f_t = g_t$ on $D_t \cap \tilde{D}_t$ for all t.

PROOF: Let S be the set of $t \in [a, b]$ such that $f_t = g_t$ on $D_t \cap \tilde{D}_t$. Then $a \in S$ since both f_a and g_a must agree with f on a neighborhood of z_0, hence everywhere on $D_a \cap \tilde{D}_a$ by the Rigidity Theorem. S is obviously (relatively) open in $[a, b]$, again by the Rigidity Theorem. If t_0 is a limit point of S, then $f_{t_0} = g_{t_0}$ on a sequence of points in S converging to t_0, and thus $f_{t_0} = g_{t_0}$ on all of $D_{t_0} \cap \tilde{D}_{t_0}$ by the Rigidity Theorem, i.e. $t_0 \in S$. Thus S is also closed. So $S = [a, b]$ since $[a, b]$ is connected. 👍

XII.9.3.9. Thus, if f can be analytically continued along γ, the value of $f(\zeta(t))$ is uniquely determined for each t, and in particular a branch of f at z_1 is uniquely determined by the analytic continuation. Note, however, that the curve γ may cross itself, and at crossing or repeated points the value of f obtained can be different at different times. This is particularly common and important if γ is a closed curve. A good example is $f(z) = \log z$, $z_0 = z_1 = 1$, γ the unit circle parametrized once around counterclockwise. This example also shows that the parametrization (or at least its equivalence class) is essential: if γ is parametrized once around clockwise, or more than once around, different values of f at z_1 are obtained.

XII.9.3.10. PROPOSITION. Let Ω be a region in $\mathbb{C}$, and f a holomorphic function on Ω. If γ is a (parametrized) curve in $\mathbb{C}$ from $z_0 \in \Omega$ to $z_1 \in \mathbb{C}$, then f can be analytically continued along γ if and only if there is a (necessarily unique) curve $\hat{\gamma}$ in S_f from the point $\hat{z}_0$ corresponding to z_0 on the branch (Ω, f) to another point $\hat{z}_1$, such that $p(\hat{\gamma}) = \gamma$.

PROOF: If there is such a $\hat{\gamma}$, f can clearly be analytically continued along γ. Conversely, if f can be analytically continued along γ, there is a partition $\{a = t_0, t_1, \ldots, t_n = b\}$ of the interval $[a, b]$ such that the disks D_{t_k} cover γ and $f_{t_k} = f_{t_{k+1}}$ on $D_{t_k} \cap D_{t_{k+1}}$ for each k. The pieces of γ for $t_k \leq t \leq t_{k+1}$ can be successively lifted to $\hat{D}_{t_k} \subseteq S_f$ in a unique way via the local homeomorphisms so that the entire lifted path is continuous. 👍

XII.9.3.11. As a result, if f can be analytically continued along γ, there is a "largest" analytic continuation (D_t, f_t), where D_t is the open disk of convergence of the Taylor series for f_t around $\zeta(t)$. We can use this analytic continuation as the standard one.

XII.9.3.12. If f can be analytically continued along γ, and γ is not entirely contained inside Ω, then there is a smallest t, say $t = t_1$, for which $\zeta(t_1) \in \partial\Omega$. Then $\zeta(t_1)$ is not a singular point of $\partial\Omega$. There is then a smallest $t_2 > t_1$ such that $\zeta(t_2) \in \partial(\Omega \cup D_{t_1})$ (unless $\gamma \subseteq \Omega \cup D_{t_1}$), and this point is not a singular point. The process can be continued.

Conversely, if f cannot be analytically continued along γ, then there is a smallest $c \in [a, b]$ such that f cannot be analytically continued along the restriction γ_c of γ to $[a, c]$. We have $c > a$ since f can be analytically continued along γ_d for $d > a$ small enough that $\gamma_d \subseteq \Omega$. Then $\zeta(c)$ can be regarded as a singular point of f along γ. For example, if $z_0 \in \Omega$, $z_1 \in \partial\Omega$, and γ is a path from z_0 to z_1 with $\zeta(t) \in \Omega$ for $a \leq t < b$, then z_1 is a singular point of f along γ if and only if z_1 is a singular point of $\partial\Omega$ in the sense of XII.9.3.2.. (There is not always such a path for every $z_1 \in \partial\Omega$ if $\partial\Omega$ is not well behaved topologically.)

XII.9.3.13. If f is holomorphic on Ω, and γ is a path beginning in Ω ending in a point z_1 outside Ω along which f can be analytically continued, then there is a well-defined value for f at z_1 (and for all z in some neighborhood of z_1) given by the analytic continuation along γ. But if $\tilde{\gamma}$ is another path beginning in Ω and ending at z_1, the value of f at z_1 defined by analytic continuation along this path need not be the same. A simple example is the principal branch of $\log z$ defined on the right half plane Ω, and the upper and lower unit semicircles from 1 to -1. However, homotopic paths give the same result:

XII.9.3.14. Theorem. [Monodromy Theorem] Let f be holomorphic on a region Ω, $z_0 \in \Omega$, $z_1 \in \mathbb{C}$. Let γ_0 and γ_1 be paths in $\mathbb{C}$ from z_0 to z_1, which are homotopic by a homotopy $h : [0, 1] \times [a, b]$ (XII.3.5.1.). For each s, let γ_s be the path parametrized by ζ_s, where $\zeta_s(t) = h(s, t)$. Suppose f can be analytically continued along γ_s for every s. Then the values of f at z_1 defined by analytic continuation along γ_0 and γ_1 are the same.

XII.9.4. The Regular Riemann Surface of a Function

XII.9.4.1. If $\boldsymbol{f}$ is a total analytic function with Riemann surface S_f, and $f : S_f \to \mathbb{C}$ is the corresponding complete analytic function, then f has a well-defined derivative at each point of S_f obtained from a local function element representative, and in fact well-defined derivatives of all orders (hence a well-defined Taylor series representation) at each point.

Viewed another way, if (Ω_k, f_k) is a function element of $\boldsymbol{f}$, then (Ω, f') is also an analytic function element, and the corresponding total analytic function $\boldsymbol{f'}$ can be regarded as the derivative of the total analytic function $\boldsymbol{f}$.

Although there is a natural map from S_f to $S_{f'}$, which is surjective (), the Riemann surface $S_{f'}$ is not the same as the Riemann surface S_f in general; it is only a quotient. (Consider the case of $\log z$; cf. Exercise ().) The map from S_f to $S_{f'}$ is a covering map in general ().

XII.9.4.2. If $\boldsymbol{f}$ is a total analytic function with Riemann surface S_f, and $\boldsymbol{f}$ is not constant, then the set of points of S_f where f' is zero is a closed subset Z_f which is discrete in the relative topology (). Thus the complement $R_f = S_f \setminus Z_f$ is a dense open subset of S_f, called the *regular Riemann surface of* $\boldsymbol{f}$.

XII.9.4.3. If $\zeta \in R_f$, then there is a function element (Ω, g) of $\boldsymbol{f}$ such that $\hat{\Omega}$ is a neighborhood of ζ and f is one-to-one on $\hat{\Omega}$, i.e. g is one-to-one on Ω (). Then $(g(\Omega), g^{-1})$ is an analytic function element which is a "branch" of $\boldsymbol{f}^{-1}$. It is not hard to see that any two such function elements corresponding to small open sets

in R_f are equivalent, so the total analytic function defined by them can be called $\boldsymbol{f}^{-1}$. The correspondence $\Omega \mapsto g(\Omega)$ gives a conformal bijection between R_f and $R_{f^{-1}}$; thus f can be regarded as a conformal bijection from R_f onto $R_{f^{-1}}$.

XII.9.4.4. Another way of viewing this is: if R_f and $R_{f^{-1}}$ are identified in this way, and p and q are the natural projections from R_f and $R_{f^{-1}}$ to $\mathbb{C}$, then f is completely determined by the two projections p and q. Thus an analytic function can be thought of as consisting of a Riemann surface and two projections to $\mathbb{C}$.

XII.9.5. Branch Points

XII.9.5.1. The regular Riemann surface R_f naturally sits as a dense subset of S_f, and the (isolated) points in $Z_f = S_f \setminus R_f$ do not appear in the Riemann surface $S_{f^{-1}}$. Conversely, there may be points in $S_{f^{-1}}$ not corresponding to points of S_f. The points of Z_f and $Z_{f^{-1}}$ can all be added to $R_f \cong R_{f^{-1}}$ to make a larger Riemann surface, and both f and f^{-1} can be extended in a meaningful way to these points. Such added points are called *branch points*.

XII.9.5.2. Example. Here is a simple example. Let $f(z) = \sqrt{z}$. Then the Riemann surface $S_f = R_f$ is a two-sheeted surface which projects to $\mathbb{C} \setminus \{0\}$ (XII.9.2.8.(iv)). The inverse function $f^{-1}(z) = z^2$ has $S_{f^{-1}} = \mathbb{C}$ and $R_{f^{-1}} = \mathbb{C} \setminus \{0\}$. The point 0 can be added to R_f, and f can be defined to be 0 at this added point, so S_f can be extended by adding a branch point. There are in this case no branch points to be added to $R_{f^{-1}}$. The roles of f and f^{-1} can of course be interchanged.

XII.9.5.3. There is an intrinsic way to describe branch points. To motivate this, it is simplest to describe branch points of $\boldsymbol{f}^{-1}$ corresponding to points of Z_f. Suppose $\zeta_0 \in Z_f$. Set $z_0 = p(\zeta_0)$. Then there is a small disk Ω centered at z_0 such that the corresponding function element (Ω, g) of $\boldsymbol{f}$ has $g'(z_0) = 0$ and no other zeroes of g'. Then z_0 is a zero of g of finite order n, and g is exactly n-to-one on $\Omega \setminus \{z_0\}$. The open set U in $R_{f^{-1}}$ corresponding to $\Omega \setminus \{z_0\}$ evenly covers a deleted neighborhood V of $g(z_0)$ n times via q, so there are locally n corresponding branches of $\boldsymbol{f}^{-1}$ over V which fit together like the branches of $\sqrt[n]{z}$ around 0. (Note that there is no reason there cannot be other branches of $\boldsymbol{f}^{-1}$ over points of V corresponding to different branches of $\boldsymbol{f}$ near z_0 than the one corresponding to ζ_0.)

If C is a small circle in V centered at $g(z_0)$, w_1 a point of C, and γ_k is C parametrized k times around counterclockwise beginning and ending at w_1, and h is any of the n corresponding branches of g^{-1} in a small neighborhood of w_1, then h can be analytically continued along γ_k for any k. For $1 \leq k < n$ the terminal branch of g^{-1} around w_1 differs from the initial branch chosen, but for $k = n$ the terminal branch is the same as the initial branch. As the diameter of C approaches 0, all the values on the branches at w_1 approach z_0. So we can "unwind" U, which is homeomorphic to a punctured disk, and add a center point where g^{-1} takes the value z_0.

XII.9.5.4. In general, a branch point of $\boldsymbol{f}$ corresponds to an open set U in S_f homeomorphic to to a punctured disk, such that p gives an even covering () from U onto a punctured disk $D \setminus \{z_0\}$ centered at a point z_0, and such that

$$w_0 = \lim_{p(\zeta) \to z_0} f(\zeta)$$

exists in the sense that for every $\epsilon > 0$ there is a $\delta > 0$ such that $|f(\zeta) - w_0| < \epsilon$ whenever $\zeta \in U$ and $|p(\zeta) - z_0| < \delta$ (it actually suffices that f just be bounded on U because of ()). Then there is an $n \in \mathbb{N}$ such that $p|_U$ is exactly n-to-one, and the image of every simple closed curve in U not homotopic to 0 has winding number $\pm n$ around z_0. If $n = 1$, there will be a point ζ_0 in the closure of U in S_f such that $p(\zeta_0) = z_0$ and p maps $U \cup \{\zeta_0\}$ homeomorphically onto D; and $f(\zeta_0) = w_0$. If $n > 1$, there is no such point on S_f; but we can then add a point ζ_0 to S_f, and topologize $S_f \cup \{\zeta_0\}$ by taking sets of the form

$$\{\zeta_0\} \cup \left[U \cap p^{-1}\left(\{z : |z - z_0| < \delta\}\right)\right]$$

for $\delta > 0$ as a local base at ζ_0. Extend f to $S_f \cup \{\zeta_0\}$ by setting $f(\zeta_0) = w_0$.

XII.9.5.5. The extended surface $S_f \cup \{\zeta_0\}$ is a topological 2-manifold with a natural structure as a complex 1-manifold, defined around ζ_0 by taking a "local uniformizing variable." Consider the map $h(s) = z_0 + s^n$ from a disk $\tilde{D}$ centered at 0 onto D. Fix an $s_1 \in \tilde{D}$, $s_1 \neq 0$, and let $z_1 = h(s_1)$ be the corresponding point of D. Define a holomorphic function G on $\tilde{D}$ as follows. Fix a branch g of f on a neighborhood Ω of z_1 corresponding to a point ζ_1 of U lying above it. If $s \in \tilde{D} \setminus \{0\}$, let γ be a path from s_1 to s in $\tilde{D} \setminus \{0\}$; then $h(\gamma)$ is a path from z_1 to $h(s)$ within $D \setminus \{z_0\}$, and g can be analytically continued along $h(\gamma)$ by () and XII.9.3.10.. Let $G(s)$ be the value of the analytic continuation at the terminal point $h(s)$. Then G is well defined on $\tilde{D} \setminus \{0\}$ (cf. [Ahl78, p. 221-222]). We have

$$\lim_{s \to 0} G(s) = w_0$$

so G has a removable singularity at 0. There is then an obvious induced homeomorphism ϕ from $\tilde{D}$ to $U \cup \{\zeta_0\}$ sending 0 to ζ_0 and s_1 to ζ_1, for which $G = f \circ \phi$, and this homeomorphism can be used to transfer the analytic structure on $\tilde{D}$ to $U \cup \{z_0\}$. This analytic structure agrees with the one already defined on U as a subset of S_f. It is easily seen that this analytic structure does not depend on the choices made in the construction of G. If the branch point arises from a point of $S_{f^{-1}}$ as in XII.9.5.3., this procedure gives the same analytic structure as the usual analytic structure from $S_{f^{-1}}$.

XII.9.5.6. Example. Not all branch points of $\boldsymbol{f}$ arise from $\boldsymbol{f}^{-1}$. Consider $f(z) = z^{2/3}$. The Riemann surface S_f has three sheets over $\mathbb{C} \setminus \{0\}$, and $\boldsymbol{f}$ has a branch point of order 3 at 0. The inverse function is $f^{-1}(z) = z^{3/2}$, whose Riemann surface has two sheets over $\mathbb{C} \setminus \{0\}$ and a branch point of order 2 at 0.

XII.9.6. Points at Infinity and Meromorphic Functions

Essentially everything done so far in this section works identically if we replace the complex plane $\mathbb{C}$ with the Riemann sphere $\bar{\mathbb{C}}$ and the words *analytic* and *holomorphic* with *meromorphic*.

XII.9.6.1. Definition. A *meromorphic function element* is a pair (Ω, f), where Ω is a region in $\mathbb{C}$ and f is a meromorphic function on Ω. Two meromorphic function elements (Ω, f) and $(\tilde{\Omega}, \tilde{f})$ are *equivalent* if there is a finite chain (Ω_j, f_j) $(1 \leq j \leq n)$ such that $(\Omega_1, f_1) = (\Omega, f)$, $(\Omega_n, f_n) = (\tilde{\Omega}, \tilde{f})$, and $\Omega_j \cap \Omega_{j+1} \neq \emptyset$ and $f_j = f_{j+1}$ on $\Omega_j \cap \Omega_{j+1}$ for $1 \leq j < n$.

A meromorphic function on Ω can be regarded as an analytic function from Ω to $\bar{\mathbb{C}}$ in the sense of (). Any analytic function element is a meromorphic function element.

XII.9.6.2. A total meromorphic function can be thought of as an equivalence class of meromorphic function elements (we will later () give a slightly different definition, allowing meromorphic function elements including the point at infinity also).

If $\boldsymbol{f} = \{(\Omega_k, f_k)\}$ is an equivalence class of meromorphic function elements, we can form a Riemann surface S_f as in the holomorphic case, where $z_j \in \Omega_j$ is identified with $z_k \in \Omega_k$ if $z_j = z_k$ in $\mathbb{C}$ and $f_j = f_k$ on a deleted neighborhood of $z_j = z_k$. (If we regard f_j and f_k as maps to $\bar{\mathbb{C}}$, they then agree on a neighborhood of $z_j = z_k$). There is an induced function from S_f to $\bar{\mathbb{C}}$. Thus every meromorphic function element gives a well-defined Riemann surface in this way corresponding to its equivalence class.

XII.9.6.3. Note that each analytic function element (Ω, f) defines a Riemann surface S_f^h as in XII.9.2.2. when thought of as an analytic function element, and a Riemann surface S_f^m when thought of as a meromorphic function element. These are not the same in general. S_f^h is naturally a subset of S_f^m; S_f^h is exactly the preimage of $\mathbb{C}$ in S_f^m under the induced function from S_f^m to $\bar{\mathbb{C}}$. Thus S_f^m can be regarded as S_f^h with points added where the extended f has poles. For example, if $f(z) = \frac{1}{z}$, then $S_f^h \cong \mathbb{C} \setminus \{0\}$ and $S_f^m \cong \mathbb{C}$.

XII.9.6.4. We can incorporate the point at infinity more fully into this picture. A meromorphic function element (Ω, f), where Ω is a region in $\mathbb{C}$ with bounded complement and f is a meromorphic function on $\mathbb{C}$ which has a removable singularity or pole at ∞ in the sense of () (this implies that f is holomorphic in the complement of a bounded set), can be regarded as a region in $\bar{\mathbb{C}}$ containing ∞ and an analytic function from this set to $\bar{\mathbb{C}}$, and will be called a *meromorphic function element at infinity*.

For each meromorphic function element (Ω, f) at infinity, for which $0 \notin \Omega$, define a meromorphic function element $(\bar{\Omega}, g)$, where

$$\bar{\Omega} = \{0\} \cup \left\{ \frac{1}{z} : z \in \Omega \right\}$$

$$g(z) = f\left(\frac{1}{z}\right) \text{ for } z \in \Omega, \quad g(0) = \lim_{z \to \infty} f(z) \in \bar{\mathbb{C}} \ .$$

We incorporate $(\bar{\Omega}, g)$ into the construction of S_f by identifying $w \in \bar{\Omega} \setminus \{0\}$ with $\frac{1}{w} \in \Omega$. This effectively adds a "point at infinity" to S_f, at which f takes the value $g(0)$. There may be many such points at infinity added corresponding to different branches of $\boldsymbol{f}$.

There is then a natural projection map p from the extended S_f to $\bar{\mathbb{C}}$.

XII.9.6.5. Branch points can then be defined and added as in the holomorphic case. A branch point corresponds to an open set U in S_f for which p is an even covering from U onto a deleted open disk $D \setminus \{z_0\}$ in $\bar{\mathbb{C}}$, where $z_0 = \infty$ is allowed in which case $D \setminus \{\infty\}$ is the complement of a closed disk in $\mathbb{C}$ centered at 0, and

$$w_0 = \lim_{p(\zeta) \to z_0} f(\zeta)$$

exists in $\bar{\mathbb{C}}$, i.e. $w_0 = \infty$ is allowed.

When all points at infinity and branch points have been added, a Riemann surface $\bar{S}_f$ is obtained which is called the *total Riemann surface* of $\boldsymbol{f}$. If $\boldsymbol{f}$ is not constant, the total Riemann surfaces of $\boldsymbol{f}$ and $\boldsymbol{f}^{-1}$ are conformally equivalent ().

XII.9.7. Germs of Analytic Functions

XII.9.7.1. We now go to the opposite extreme from considering the Riemann surface of a total analytic function. If $z_0 \in \mathbb{C}$ is a fixed point of $\mathbb{C}$, and if $\boldsymbol{f}$ is a total analytic function containing an analytic function element (Ω, f) with $z_0 \in \Omega$, then the restriction of f to any neighborhood of z_0 completely determines $\boldsymbol{f}$. Thus every function holomorphic in a neighborhood of z_0 determines a unique total analytic function, and two holomorphic functions agreeing on a neighborhood of z_0 determine the same total analytic function. (But two different holomorphic functions on small neighborhoods of z_0 can give different branches of the same total analytic function.)

XII.9.7.2. Definition. Let $z_0 \in \mathbb{C}$. A *germ* of an analytic function at z_0 is an equivalence class of functions holomorphic in neighborhoods of z_0, where two such functions are equivalent if they agree on some neighborhood of z_0.

Thus the exact domain of a germ of an analytic function at z_0 is not specified or relevant.

XII.9.7.3. The algebraic operations of addition, multiplication, and scalar multiplication are well defined on germs. If f and g are holomorphic on neighborhoods U and V of z_0, then $f+g$ is defined and holomorphic on $U \cap V$. If $\tilde{f}$ is holomorphic near z_0 and agrees with f on $\tilde{U} \subseteq U$, and $\tilde{g}$ agrees with g on $\tilde{V} \subseteq V$, then $\tilde{f} + \tilde{g}$ agrees with $f + g$ on $\tilde{U} \cap \tilde{V}$, so addition is well defined. The argument for multiplication and scalar multiplication is almost identical.

XII.9.7.4. Thus the set $\mathcal{G}(z_0)$ of germs of analytic functions has a natural structure as a ring, actually an algebra over $\mathbb{C}$.

XII.9.7.5. The algebra $\mathcal{G}(z_0)$ can be described in a different way. If f is holomorphic in a neighborhood of z_0, then the Taylor series of f around z_0 determines f on at least a small disk centered at z_0. Conversely, any power series centered at z_0 with a positive radius of convergence defines a holomorphic function on a disk centered at z_0. Thus there is a one-one correspondence between germs of analytic functions at z_0 and power series centered at z_0 with a positive radius of convergence (we will call such a series a *convergent power series* centered at z_0). This correspondence preserves addition, multiplication, and scalar multiplication, i.e. is an algebra isomorphism.

Up to isomorphism, the algebra $\mathcal{G}(z_0)$ does not depend on z_0; the map $f(z) \mapsto f(z+z_0)$ is an isomorphism of $\mathcal{G}(z_0)$ onto $\mathcal{G}(0)$, which is isomorphic to the algebra of convergent power series centered at 0. So $\mathcal{G}(0)$ can be identified with a subalgebra of the algebra $\mathbb{C}[[X]]$ of formal power series over $\mathbb{C}$ ().

XII.9.7.6. Proposition. The algebra $\mathcal{G}(z_0)$ has no zero divisors: if $[f], [g]$ are nonzero elements of $\mathcal{G}(z_0)$, then $[f][g] \neq 0$. That is, $\mathcal{G}(z_0)$ is an integral domain ().

Proof: This can be seen in at least two ways. $\mathcal{G}(z_0)$ is isomorphic to a subring of the integral domain $\mathbb{C}[[X]]$. Alternatively, if $[f]$ and $[g]$ are germs of analytic functions f and g respectively, neither of which are identically zero in a neighborhood of z_0, then f and g have at most an isolated zero at z_0, i.e. there is a deleted neighborhood U of z_0 on which both f and g are never zero. Thus fg is nonzero on U, i.e. $[fg] = [f][g] \neq 0$ in $\mathcal{G}(z_0)$. 👍

Germs of Meromorphic Functions

We can extend the notion of germ to functions holomorphic on a neighborhood of z_0 except possibly for a pole at z_0.

XII.9.7.7. DEFINITION. Let $z_0 \in \mathbb{C}$. A *germ* of a meromorphic function at z_0 is an equivalence class of functions holomorphic in deleted neighborhoods of z_0 with a removable singularity or pole at z_0, where two such functions are equivalent if they agree on some deleted neighborhood of z_0.

XII.9.7.8. As before, the set $\mathcal{M}(z_0)$ of germs of meromorphic functions at z_0 is a complex algebra. There is an obvious injective homomorphism from $\mathcal{G}(z_0)$ to $\mathcal{M}(z_0)$, so we may identify $\mathcal{G}(z_0)$ with a subalgebra of $\mathcal{M}(z_0)$. We again have $\mathcal{M}(z_0) \cong \mathcal{M}(0)$.

There is a one-one correspondence (algebra isomorphism) between $\mathcal{M}(z_0)$ and the algebra of Laurent series () centered at z_0 which converge on a deleted neighborhood of z_0 (*convergent Laurent series*). Thus $\mathcal{M}(z_0)$ is isomorphic to a subalgebra of the algebra $\mathbb{C}((X))$ of formal Laurent series with complex coefficients ().

XII.9.7.9. PROPOSITION. The algebra $\mathcal{M}(z_0)$ is a field; in fact, it is the quotient field of the ring $\mathcal{G}(z_0)$.

PROOF: If f is holomorphic on a deleted neighborhood of z_0 and has a pole at z_0, then $\frac{1}{f}$ has a removable singularity at z_0, so agrees on a deleted neighborhood of z_0 with a function holomorphic in a neighborhood of z_0. Thus $[f]$ is invertible in $\mathcal{M}(z_0)$, and its reciprocal is in $\mathcal{G}(z_0)$. And if $[g] \in \mathcal{G}(z_0)$, $[g]$ not the zero element (i.e. g not identically zero near z_0), then either $g(z_0) \neq 0$, in which case $[g]$ is already invertible in $\mathcal{G}(z_0)$, or g has an isolated zero at z_0, in which case $\frac{1}{g}$ is holomorphic in a deleted neighborhood of z_0 and has a pole at z_0, i.e. $[g]$ is invertible in $\mathcal{M}(z_0)$. 👍

XII.9.7.10. Note that this proposition cannot be proved simply by using the isomorphism with a subring of the field $\mathbb{C}((X))$, since it is not obvious (although it turns out to be true) that the inverse Laurent series to a convergent Laurent series is also convergent.

XII.9.8. General Riemann Surfaces

XII.9.9. Exercises

XII.9.9.1. (a) (H. Looman) For $z \in \mathbb{C}$, define $f(z) = e^{-1/z^4}$ if $z \neq 0$, and set $f(0) = 0$. Show that the partial derivatives of the real and imaginary parts of f exist and satisfy the Cauchy-Riemann equations everywhere on $\mathbb{R}^2$, but f is not complex differentiable, or even continuous, at 0.

(b) (D. Menchoff) Set $f(z) = \frac{z^5}{|z|^4}$ for $z \neq 0$, and $f(0) = 0$. Show that f is continuous and has first-order partials everywhere on $\mathbb{C}$, and the first-order partials at 0 satisfy the Cauchy-Riemann equations, but f is not complex differentiable at 0.

XII.9.9.2. Let $\mathbb{D}$ be the open unit disk in $\mathbb{C}$.

(a) Show that the power series $\displaystyle\sum_{k=1}^{\infty} \frac{z^k}{k^2}$ converges uniformly on $\mathbb{D}$.

(b) Show that the series $\displaystyle\sum_{k=1}^{\infty} \frac{z^{k-1}}{k}$ of term-by-term derivatives converges u.c. but not uniformly on $\mathbb{D}$.

XII.9.9.3. (Another proof of the Fundamental Theorem of Algebra, the "standard" one) Let f be a polynomial with complex coefficients and no root in $\mathbb{C}$, and let $g = \frac{1}{f}$. Then g is an entire function.

(a) Use Lemma XII.1.4.1. to show that g is bounded.

(b) Apply Liouville's Theorem () to conclude that g, and hence f, is constant.

XII.9.9.4. (Yet another proof of the Fundamental Theorem of Algebra [Boa64]) Suppose f is a polynomial of degree $n \geq 1$ with no root in $\mathbb{C}$.

(a) Set $p(z) = f(z)\bar{f}(z)$, where $\bar{f}$ is the polynomial obtained from f by replacing each coefficient by its complex conjugate. Then p is a polynomial of degree $2n$ with no root in $\mathbb{C}$, taking strictly positive values on the real axis.

(b) Set $I = \displaystyle\int_{-\pi}^{\pi} \frac{1}{p(2\cos\theta)}\, d\theta$. Then $I > 0$.

(c) Show that $I = -i \displaystyle\int_{\gamma} \frac{1}{\zeta p(\zeta + \zeta^{-1})}\, d\zeta$, where γ is the unit circle.

(d) Show that $p(z + z^{-1}) = z^{-2n} q(z)$ for a polynomial q with no roots in $\mathbb{C}$.

(e) Conclude from Cauchy's Theorem that $I = -i \int_{\gamma} \frac{\zeta^{2n-1}}{q(\zeta)}\, d\zeta = 0$, contradicting (b).

XII.9.9.5. Prove the following complex analog of Rolle's Theorem for polynomials:

Theorem. Let $f(z)$ be a polynomial with complex coefficients. Then the roots of f' are contained in the (necessarily closed) convex hull K of the roots of f.

(a) Since K is an intersection of closed half planes, it suffices to show that if H is a closed half plane containing all the roots of f, then H also contains all the roots of f'. Show that it suffices to prove this if H is the closed upper half plane.

(b) We may assume f is monic without loss of generality. Use the Fundamental Theorem of Algebra to write

$$f(z) = (z - z_1)^{m_1}(z - z_2)^{m_2} \cdots (z - z_n)^{m_n}$$

where $z_1, \ldots, z_n \in \mathbb{C}$ and $m_1, \ldots, m_n \in \mathbb{N}$.

(c) Use either the Product Rule or logarithmic differentiation to show that

$$\frac{f'(z)}{f(z)} = \frac{m_1}{z - z_1} + \frac{m_2}{z - z_2} + \cdots + \frac{m_n}{z - z_n}$$

for $z \in \mathbb{C} \setminus \{z_1, \ldots, z_n\}$.

(d) Under the assumption that $Im(z_k) \geq 0$ for all k, show that if $Im(z) < 0$, then $Im\left(\frac{f'(z)}{f(z)}\right) > 0$ and hence $f'(z) \neq 0$.

(e) Explain why the theorem is a complex analog of Rolle's Theorem.

XII.9.9.6. The aim of this problem is to further justify the definition of the complex exponential function (XII.1.1.8.(iii)).

(a) Let $f(z) = e^z$. If g is holomorphic on a connected region Ω, and satisfies $g' = g$ on Ω, then $g = cf$ for some $c \in \mathbb{C}$. [Apply the Quotient Rule to $\frac{g}{f}$.] If $0 \in \Omega$ and $g(0) = 1$, then $g = f$.

(b) Similarly characterize the complex trigonometric and hyperbolic functions.

XII.9.9.7. Discuss the Riemann surface for the function $\log \log z$.

XII.9.9.8. Discuss the validity of the following formulas, where all letters denote complex numbers, and interpret them in terms of Riemann surfaces:

(a) $\log(zw) = \log z + \log w$.

(b) $\log(z^c) = c \log z$.

(c) $z^{c+d} = z^c z^d$.

(d) $(z^c)^d = z^{cd}$.

(e) $(zw)^c = z^c w^c$.

XII.9.9.9. Develop the argument in the last part of the proof of XII.3.5.4. to show that if γ_0 and γ_1 are piecewise-smooth contours from p to q which are homotopic in an open set Ω, then they are homotopic in Ω via a piecewise-smooth homotopy. [Choose partitions $\mathcal{P}$ and $\mathcal{Q}$ as in the proof; for $1 < j < m$ let $H(s_j, t)$ be piecewise-linear in t with $H(s_j, t_k) = h(s_j, t_k)$ for each k, and then extend H by making it piecewise-linear in s for fixed t.]

XII.9.9.10. (a) Let Ω be a connected open set in $\mathbb{C}$ containing an interval I on the real axis. Show that the complex exponential function is the only function which is holomorphic on Ω and which agrees with the usual real exponential function on I. [Use the Rigidity Theorem.] Do the same for the sine and cosine functions.

(b) Show that the power series

$$\sum_{k=0}^{\infty} \frac{z^k}{k!}$$

converges u.c. on all of $\mathbb{C}$ to the function e^z. Similarly, show that the series expansions

$$\cos z = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} z^{2k}$$

$$\sin z = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} z^{2k+1}$$

are valid for all $z \in \mathbb{C}$. Give similar series representations for $\cosh z$ and $\sinh z$.

Thus the usual power series representations of the real exponential, trigonometric, and hyperbolic functions are valid also for the complex ones. This problem provides additional justification for why the extensions we gave of the exponential function, etc., to $\mathbb{C}$ are reasonable (and indeed are the *only* reasonable ones.) In fact, in many references these functions are defined using the power series.

XII.9.9.11. [**?**] Here is an example of an improper real integral which can be computed by a more complicated contour integration. Consider the integral

$$\int_0^{\infty} \frac{x^{-k}}{1+x}\, dx$$

(where k is a real constant and x^{-k} is interpreted in the usual real sense, i.e. as a positive real number for $x > 0$).

(a) Show that this improper integral converges if and only if $0 < k < 1$.

We make this restriction in the rest of the problem. Define

$$f(z) = \frac{z^{-k}}{1+z}$$

where z^{-k} is defined using a branch cut along the positive real axis and taking the argument between 0 and 2π. Thus the argument of z^{-k} lies between $-2\pi k$ and 0.

We also consider the contour $\gamma = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$:
where γ_1 and γ_3 are line segments along the ray of points with argument ϵ and $2\pi - \epsilon$ respectively, and γ_2 and γ_4 are circle arcs of radius R and r respectively.

(b) Show that if $0 < r < 1$, $R > 1$, and $0 < \epsilon < \pi$ we have

$$\int_{\gamma} f = 2\pi i\, Res(f, -1) = 2\pi i e^{-\pi i k}\ .$$

(c) Show that

$$\int_{\gamma_2} f \to 0$$

as $R \to +\infty$ and

$$\int_{\gamma_4} f \to 0$$

as $r \to 0$, uniformly in ϵ.

(d) Show that, for fixed r and R, as $\epsilon \to 0+$ we have

$$\int_{\gamma_1} f \to \int_r^R \frac{x^{-k}}{1+x}\,dx$$

and that

$$\int_{\gamma_3} f \to -e^{-2\pi ik} \int_r^R \frac{x^{-k}}{1+x}\,dx$$

[The integrals do not approach the same limit because of the branch cut.]

(e) Conclude that

$$\int_0^\infty \frac{x^{-k}}{1+x}\,dx = \frac{2\pi i e^{-\pi ik}}{1-e^{-2\pi ik}}\ .$$

(f) Show that

$$1 - e^{-2\pi ik} = 2ie^{-\pi ik}\sin \pi k$$

so that

$$\int_0^\infty \frac{x^{-k}}{1+x}\,dx = \frac{\pi}{\sin \pi k}\ .$$

XII.9.9.12. Suppose f has a singularity at z_0, and that f is bounded in a deleted neighborhood of z_0. Mimic the proof of (iv) $\Rightarrow$ (i) in XII.5.2.7. to show that the singularity is removable, by showing directly that for fixed z near z_0,

$$\lim_{\delta \to 0+} \left[\int_{\gamma_\delta} \frac{f(\zeta)}{\zeta - z}\,d\zeta \right] = 0$$

by estimating the size of the integral.

XII.9.9.13. (a) Show that the implication (ii) $\Rightarrow$ (i) in XII.5.2.7. is equivalent to the following:

THEOREM. Let D be an open disk in $\mathbb{C}$ centered at z_0, and $f : D \to \mathbb{C}$. If f is continuous on D and holomorphic on $D \setminus \{z_0\}$, then f is holomorphic on D (i.e. differentiable at z_0).

(b) Prove this theorem by showing that the integral of f around any triangle in D is zero and applying Morera's Theorem.

XII.9.9.14. Using the Big Picard Theorem (XII.5.2.19.) and XII.5.3.5., prove the Little Picard Theorem (XII.3.3.12.). Do the case of a polynomial separately using the Fundamental Theorem of Algebra.

XII.9.9.15. For $z \in \mathbb{C}$ set $f(z) = e^{e^z}\,(= e^{(e^z)})$. Then f is an entire function.

What is wrong with the following argument? Since e^z is never 0, f never takes either the value 0 or the value 1. Thus f is a counterexample to the Little Picard Theorem (XII.3.3.12.).

XII.9.9.16. Let f be a continuous function from $\mathbb{C}$ to $\mathbb{C}$ such that the range of f contains the unit circle but not 0 (e.g. $f(z) = e^z$). Give a topological proof that f cannot be one-to-one.

XII.9.9.17. Convergent Power Series. This problem is a continuation of III.5.3.2., III.5.3.7.,and III.5.3.8.. Let F be a nondiscrete normed field. The prototype example is $\mathbb{C}$, but F could be $\mathbb{R}$ or $\mathbb{Q}_p$. We do not assume F is complete, so F could be any subfield of $\mathbb{C}$ or $\mathbb{Q}_p$ such as $\mathbb{Q}$ with either the usual absolute value or the p-norm. (There are many other possibilities too.) Denote the norm on F by $|\cdot|$.

(a) Let $\sum_{k=0}^{\infty} a_k X^k \in F[[X]]$. Show that the following conditions are equivalent:

(i) $\sup_k |a_k|^{1/k} < \infty$.

(ii) There is an $x \in F$, $x \neq 0$, such that $(|a_k x^k|)$ is bounded. [Since F is nondiscrete, there are nonzero x's in F of arbitrarily small norm.]

(iii) There is an $R > 0$ such that the partial sums of the series $\sum_{k=0}^{\infty} a_k x^k$ form a Cauchy sequence in F for all $x \in F$, $|x| < R$. [The norm on a normed field F satisfies $|xy| = |x||y|$ for all $x, y \in F$.]

A power series satisfying these conditions will be said to have a *positive radius of convergence*. The theoretical radius of convergence is

$$R = \frac{1}{\limsup |a_k|^{1/k}} .$$

If $F = \mathbb{C}$, these conditions are equivalent to having a positive radius of convergence in the usual sense. But since F is not required to be complete, a power series with positive radius of convergence will not necessarily converge (in F) for any nonzero elements of F. And the exact "radius of convergence" may not be well defined since the norm on a normed field does not necessarily take a dense set of values in $\mathbb{R}$.

(b) Let $\mathcal{G}_F[[X]]$ be the subset of $F[[X]]$ consisting of the power series with positive radius of convergence. (There does not seem to be standard notation for this set; we use this notation since $\mathcal{G}_{\mathbb{C}}[[X]]$ can be identified with the set $\mathcal{G}(0)$ of germs of analytic functions around 0.) $\mathcal{G}_F[[X]]$ can be thought of as the set of germs of "analytic" functions from F to F around 0, although this does not make literal sense in general. Show that $\mathcal{G}_F[[X]]$ is a subring (subalgebra) of $F[[X]]$. The polynomial ring $F[X]$ is a subring (subalgebra) of $\mathcal{G}_F[[X]]$. (The subalgebra $\mathcal{G}_F[[X]]$ of $F[[X]]$ depends on the norm used: if $F = \mathbb{Q}$, it is different for the usual absolute value and for the various p-norms.)

(c) Using the notation of III.5.3.7., if $f \in \mathcal{G}_F[[X]]$ and $f(0) \neq 0$, show that f is invertible in $\mathcal{G}_F[[X]]$, i.e. show that the inverse $\frac{1}{f}$ in $F[[X]]$ is in $\mathcal{G}_F[[X]]$. Show that $\mathcal{G}_F[[X]]$ has exactly one maximal ideal, the set of f with $f(0) = 0$, which is the principal ideal generated by $f(X) = X$.

(d) Define the set $\mathcal{G}_F((X))$ of Laurent series with positive radius of convergence. $\mathcal{G}_{\mathbb{C}}((X))$ can be identified with the set of germs of meromorphic functions around 0 (). Show that $\mathcal{G}_F((X))$ is a subfield of $F((X))$ and is the quotient field of $\mathcal{G}_F[[X]]$.

(e) Give an appropriate definition of positive radius of convergence for power series of several variables, and define $\mathcal{G}_F[[X_1, \ldots, X_n]]$ and $\mathcal{G}_F((X_1, \ldots, X_n))$. Formulate and prove analogs of the constructions of III.5.3.7.–III.5.3.8.: composition and the Implicit and Inverse Function Theorems, i.e. show in the results of III.5.3.7.–III.5.3.8., if the given functions have positive radius of convergence, so do the various functions asserted to exist in III.5.3.7.–III.5.3.8. (III.5.3.7.(a) does not work unrestrictedly).

If $F = \mathbb{C}$ (or $F = \mathbb{R}$), these theorems have analytic proofs. But there are direct proofs valid if F is any nondiscrete normed field. In proving the theorems, one may assume F is complete: if $\bar{F}$ is the completion, and the asserted functions are in $\mathcal{G}_{\bar{F}}[[X_1, \ldots, X_n]]$, they are actually in $F[[X_1, \ldots, F_n]] \cap \mathcal{G}_{\bar{F}}[[X_1, \ldots, X_n]] = \mathcal{G}_F[[X_1, \ldots, X_n]]$.

XII.10. Analytic Functions of Several Variables

We now describe the basic theory of complex analytic functions of several variables. Some aspects of the theory are virtually identical to the one-variable case, but there are also significant differences.

This is a large and active subject, and we only cover the most elementary aspects of it; even [BM48] gives a far more comprehensive treatment. The standard general reference for the theory of several complex variables is [GR09]. See [Gun90a]–[Gun90c] for a more modern and detailed study. See also [H90] and [AW98].

XII.10.1. Elementary Theory

XII.10.1.1. Definition. Let U be an open set in $\mathbb{C}^n$. A function $f : U \to \mathbb{C}$ is *analytic* if, for every $\boldsymbol{c} \in U$, f is represented in a neighborhood of $\boldsymbol{c}$ by a multivariable power series (VIII.8.1.4.) centered at $\boldsymbol{c}$. A function $\boldsymbol{f} = (f_1, \ldots, f_m) : U \to \mathbb{C}^m$ is analytic if each f_k is analytic.

XII.10.1.2. As in the one-variable case, this definition is a local one: a function f on an open set U is analytic on U if and only if it is analytic on a neighborhood of $\boldsymbol{c}$ for each $\boldsymbol{c} \in U$. In particular, if $U = \cup_j U_j$ is a union of open sets, then f is analytic on U if and only if it is analytic on each U_j.

An immediate corollary of VIII.8.1.13. is:

XII.10.1.3. Corollary. Let U be an open set in $\mathbb{C}^n$, and f analytic on U. Then f is $\mathcal{C}^\infty$ on U, and for any multi-index α, $\frac{\partial^\alpha f}{\partial z_\alpha}$ is analytic on U.

XII.10.1.4. An analytic function f is "jointly analytic" in its variables. If all but one variable is fixed, it is analytic in the remaining variable in the usual sense of (), since the multivariable power series for f around any point has a drastic revision obtained by collecting together all terms of like power in the remaining variable, which will converge in a neighborhood of the point. (More generally, if some of the variables are fixed, the function is analytic in the remaining variables.) Thus an analytic function is "separately analytic" in its variables. Conversely, one of the major theorems of the subject is that a separately analytic function is (jointly) analytic:

XII.10.1.5. Theorem. [Hartogs' Theorem] Let U be an open set in $\mathbb{C}^n$, and $f : U \to \mathbb{C}$. Suppose f is analytic (holomorphic) on U in each variable separately when the other variables are held fixed, i.e. all first-order partials of f (in the complex sense) exist everywhere on U. Then f is analytic on U.

We will prove the theorem only under the additional assumption that f is continuous (which is of course a consequence of the theorem); for the considerably more difficult proof without this assumption, see [BM48].

XII.10.1.6. THEOREM. [OSGOOD'S LEMMA] Let U be an open set in $\mathbb{C}^n$, and $f : U \to \mathbb{C}$. Suppose f is continuous on U, and analytic (holomorphic) on U in each variable separately when the other variables are held fixed, i.e. all first-order partials of f (in the complex sense) exist everywhere on U. Then f is analytic on U.

PROOF: The proof is a multidimensional version of the proof of XII.4.4.6.. Let $\boldsymbol{c} = (c_1, \ldots, c_n) \in U$, and let $R = \{(z_1, \ldots, z_n) : |z_k - c_k| \leq r_k\}$ be a closed polydisk around $\boldsymbol{c}$ contained in U. Fix $(z_1, \ldots, z_n)$ in the interior R° of R, i.e. $|z_k - c_k| < r_k$ for all k. We have, by repeated application of the Cauchy Integral Formula (),

$$f(z_1, \ldots, z_n) = \frac{1}{2\pi i} \int_{\gamma_n} \frac{f(z_1, \ldots, z_{n-1}, \zeta_n)}{\zeta_n - z_n}\, d\zeta_n$$

$$= \frac{1}{2\pi i} \int_{\gamma_{n-1}} \left[\frac{1}{2\pi i} \int_{\gamma_n} \frac{f(z_1, \ldots, z_{n-2}, \zeta_{n-1}, \zeta_n)}{\zeta_n - z_n}\, d\zeta_n \right] \frac{1}{\zeta_{n-1} - z_{n-1}}\, d\zeta_{n-1}$$

$$= \cdots = \frac{1}{(2\pi i)^n} \int_{\gamma_1} \int_{\gamma_2} \cdots \int_{\gamma_n} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - z_1) \cdots (\zeta_n - z_n)}\, d\zeta_n \cdots d\zeta_2\, d\zeta_1$$

where γ_k is the circle $\{z_k : |z_k - c_k| = r_k\}$ parametrized once around counterclockwise.

By parametrizing the integrals, $f(z_1, \ldots, z_n)$ becomes an iterated integral of a complex-valued continuous function over a rectangle in $\mathbb{R}^n$. By Fubini's Theorem (), this integral can be converted to a multiple integral (or computed as an iterated integral in any order). We will symbolically denote this multiple integral by

$$f(z_1, \ldots, z_n) = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - z_1) \cdots (\zeta_n - z_n)}\, d\zeta \ .$$

For each k set $b_k = \frac{|z_k - c_k|}{r_k} < 1$. We have, for $(\zeta_1, \ldots, \zeta_n) \in \partial R$,

$$\frac{1}{\zeta_k - z_k} = \frac{1}{\zeta_k - c_k} \frac{1}{1 - \frac{z_k - c_k}{\zeta_k - c_k}} = \frac{1}{\zeta_k - c_k} \sum_{m=0}^{\infty} \left[\frac{z_k - c_k}{\zeta_k - c_k} \right]^m$$

and the terms in the series are dominated in absolute value by the terms of $\sum_{m=0}^{\infty} b_k^m$, so the series converges absolutely and uniformly on ∂R. Iterating, the iterated sum

$$\sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} \cdots \sum_{m_n=0}^{\infty} \frac{(z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}$$

converges absolutely and uniformly on ∂R to $\frac{1}{(\zeta_1 - z_1) \cdots (\zeta_n - z_n)}$. Thus the unordered sum

$$\sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} \frac{(z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}$$

converges uniformly on ∂R to $\frac{1}{(\zeta_1 - z_1) \cdots (\zeta_n - z_n)}$. So we have

$$f(z_1, \ldots, z_n) = \frac{1}{(2\pi i)^n} \int_{\partial R} f(\zeta_1, \ldots, \zeta_n) \left[\sum_{(m_1, \ldots, m_n) \in \mathbb{N}_0^n} \frac{(z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \right] d\zeta$$

$$= \frac{1}{(2\pi i)^n} \int_{\partial R} \left[\sum_{(m_1,\ldots,m_n)\in\mathbb{N}_0^n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n} \frac{f(\zeta_1,\ldots,\zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \right] d\zeta \ .$$

Since f is bounded on R, by comparison the unordered sum

$$\sum_{(m_1,\ldots,m_n)\in\mathbb{N}_0^n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n} \frac{f(\zeta_1,\ldots,\zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}$$

converges uniformly on ∂R, and hence the summation can be moved past the integral (). Thus

$$f(z_1,\ldots,z_n) = \sum_{(m_1,\ldots,m_n)\in\mathbb{N}_0^n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n} \left[\frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1,\ldots,\zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \, d\zeta \right]$$

$$= \sum_{(m_1,\ldots,m_n)\in\mathbb{N}_0^n} a_{m_1,\ldots,m_n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}$$

where

$$a_{m_1,\ldots,m_n} = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1,\ldots,\zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}} \, d\zeta$$

(where the last integral can be interpreted as an iterated contour integral in any order). Thus, since $(z_1,\ldots,z_n)$ is arbitrary in R°, the multivariable power series

$$\sum_{(m_1,\ldots,m_n)\in\mathbb{N}_0^n} a_{m_1,\ldots,m_n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}$$

converges to f on R°, a neighborhood of $\boldsymbol{c}$. Since $\boldsymbol{c} \in U$ is arbitrary, f is analytic on U. 👍

XII.10.1.7. Corollary. Let U be an open set in $\mathbb{C}^n$, and $f : U \to \mathbb{C}$ a function. Then f is analytic on U if and only if f, when regarded as a function $f(x_1, y_1, \ldots, x_n, y_n)$ from $\mathbb{R}^{2n}$ to $\mathbb{R}^2$ with coordinate functions u and v, is differentiable in the vector calculus sense on U and the Cauchy-Riemann equations hold for each pair (x_k, y_k), i.e. for all k,

$$\frac{\partial u}{\partial x_k} = \frac{\partial v}{\partial y_k} \quad \text{and} \quad \frac{\partial u}{\partial y_k} = -\frac{\partial v}{\partial x_k} \ .$$

Proof: If f is analytic, it is differentiable in the vector calculus sense and separately analytic, so by XII.1.1.5. the Cauchy-Riemann equations are satisfied. Conversely, if it is differentiable, hence continuous, and the Cauchy-Riemann equations are satisfied, apply XII.1.1.5. to conclude that f is separately analytic. 👍

XII.10.1.8. Using the Looman-Menchoff Theorem (XII.1.1.6.), the assumption that f be differentiable in the vector calculus sense can be relaxed to just assuming f is continuous, and that the first-order partial derivatives of u and v (in the vector calculus sense) all exist everywhere on U and satisfy the Cauchy-Riemann equations. The assumptions can be further relaxed using Hartogs' Theorem, but they cannot be relaxed to simply assuming the partials exist everywhere and satisfy the Cauchy-Riemann equations (this does not imply analyticity even when $n = 1$; cf. XII.9.9.1.).

As a corollary of the proof, we obtain a multidimensional version of the Cauchy Integral Formula and of XII.4.4.6.:

XII.10.1.9. Corollary. Let U be an open set in $\mathbb{C}^n$, and $f : U \to \mathbb{C}$ an analytic function. For any $\boldsymbol{c} = (c_1, \ldots, c_n) \in U$, we have:

(i) If R is any closed polydisk in U containing $\boldsymbol{c}$ in its interior, we have

$$f(\boldsymbol{c}) = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1) \cdots (\zeta_n - c_n)}\, d\zeta$$

where the multiple integral can be regarded as an iterated contour integral in any order.

(ii) The multivariable power series

$$\sum_{(m_1,\ldots,m_n)\in\mathbb{N}_0^n} a_{m_1,\ldots,m_n} (z_1 - c_1)^{m_1} \cdots (z_n - c_n)^{m_n}$$

in standard form representing f in a neighborhood of $\boldsymbol{c}$ is unique and converges u.c. to f on any open polydisk centered at $\boldsymbol{c}$ contained in U; we have, for any $(m_1, \ldots, m_n) \in \mathbb{N}_0^n$,

$$\frac{\partial^{m_1+\cdots+m_n} f}{\partial z_1^{m_1} \cdots \partial z_n^{m_n}}(\boldsymbol{c}) = (m_1!) \cdots (m_n!) a_{m_1,\ldots,m_n} = \frac{(m_1!) \cdots (m_n!)}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}\, d\zeta$$

for any closed polydisk R contained in U and containing $\boldsymbol{c}$ in its interior. Thus

$$a_{m_1,\ldots,m_n} = \frac{1}{(2\pi i)^n} \int_{\partial R} \frac{f(\zeta_1, \ldots, \zeta_n)}{(\zeta_1 - c_1)^{m_1+1} \cdots (\zeta_n - c_n)^{m_n+1}}\, d\zeta$$

where the multiple integral can be regarded as an iterated contour integral in any order.

XII.10.2. Exercises

XII.10.2.1. Prove the complex version of Clairaut's Theorem: if f is analytic on $\Omega \subseteq \mathbb{C}^n$, then for any j and k we have

$$\frac{\partial^2 f}{\partial z_j \partial z_k} = \frac{\partial^2 f}{\partial z_k \partial z_j}$$

on Ω, and more generally that higher mixed partials are independent of the order of differentiation.

XII.11. Harmonic Functions

Harmonic functions form a bridge between complex analysis, partial differential equations, and mathematical physics. They are solutions to a PDE with wide applications in physics, and in the two-dimensional case are intimately connected with complex analytic functions.

XII.11.1. Definitions and Basic Properties

XII.11.1.1. Definition. Let Ω be a region (open set) in $\mathbb{R}^n$. A *harmonic function* on Ω is a $\mathcal{C}^2$ function $u : \Omega \to \mathbb{R}$ satisfying the *Laplace equation*

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2} = 0$$

or

$$u_{x_1 x_1} + u_{x_2 x_2} + \cdots + u_{x_n x_n} = 0$$

which can be written more compactly as

$$\Delta u = 0 \text{ or } \nabla^2 u = 0$$

where $\Delta = \nabla^2$ is the *Laplace operator*

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_n^2} \ .$$

XII.11.1.2. We will almost exclusively concern ourselves with the case $n = 2$, which has close connections with complex analytic functions; in fact, this case can almost be considered to be the same as the theory of analytic functions. In this case the Laplace operator can be written

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

and a function u is harmonic if $u_{xx} + u_{yy} = 0$.

XII.11.1.3. Examples. (i) Any linear or affine function (in particular, any constant function) is harmonic.

(ii) The function $u(x, y) = x^2 - y^2$ is harmonic on $\mathbb{R}^2$. More generally, if $\alpha_1, \ldots, \alpha_n$ satisfy $\sum_{k=1}^n \alpha_k = 0$, then $u(x_1, \ldots, x_n) = \alpha_1 x_1^2 + \cdots + \alpha_n x_n^2$ is harmonic on $\mathbb{R}^n$.

(iii) The function $u(x, y) = \log \sqrt{x^2 + y^2} = \frac{1}{2} \log(x^2 + y^2)$ is harmonic on $\mathbb{R}^2 \setminus \{(0, 0)\}$. This is a particularly important example (cf. XII.11.5.1.).

(iv) If Ω is a region in $\mathbb{C} \cong \mathbb{R}^2$, and f is a complex-analytic function on Ω, then the Cauchy-Riemann equations, plus the fact that f is $\mathcal{C}^\infty$, imply that the real and imaginary parts of f are harmonic. (This includes (ii), since u is the real part of $f(z) = z^2$, and also essentially includes (iii) since u is the real part of $f(z) = \log z$.)

XII.11.1.4. A function which is harmonic on Ω is, of course, also harmonic on any subregion. Conversely, a real-valued function on Ω which is harmonic on an open neighborhood of each point of Ω is harmonic on Ω (harmonicity is a *local property*).

XII.11.1.5. Proposition. Let Ω be a region in $\mathbb{R}^n$. Any linear combination of harmonic functions on Ω is harmonic (Δ is a *linear operator*).

XII.11.1.6. A simple formal calculation using Clairaut's Theorem seems to show that any partial derivative of a harmonic function is harmonic. But there is a technicality: a harmonic function is only assumed to have continuous second-order partials, so a partial derivative is not obviously $\mathcal{C}^2$. But harmonic functions are actually automatically $\mathcal{C}^\infty$, even real-analytic (see XII.11.2.4. for the two-dimensional case), so partial derivatives of any order of harmonic functions are harmonic.

XII.11.2. Harmonic and Analytic Functions

In this section, we will only consider harmonic functions on regions in $\mathbb{R}^2$, which we can (and will) identify with $\mathbb{C}$.

XII.11.2.1. XII.11.1.3.(iv) shows that there is a connection between harmonic and analytic functions. The connection is much more intimate: the converse also essentially holds (holds locally). This provides a powerful tool for analyzing harmonic functions.

XII.11.2.2. The converse is not unrestrictedly true: the function $u(x,y) = \log\sqrt{x^2+y^2}$ of XII.11.1.3.(iii) is harmonic on $\mathbb{C}\setminus\{0\}$, but is not the real part of an analytic function on all of $\mathbb{C}\setminus\{0\}$. Thus we have to work locally.

XII.11.2.3. Theorem. Let Ω be a region in $\mathbb{C}$. If u is harmonic on Ω, then in a neighborhood of each point of Ω there is a harmonic function v, unique up to adding a constant, such that $f(x+iy) = u(x,y) + iv(x,y)$ is analytic. (The function v is called the *harmonic conjugate* of u.) Thus u is locally the real part of an analytic function on Ω. If Ω is simply connected, there is a harmonic conjugate for u on all of Ω, i.e. u is the real part of an analytic function on Ω.

Proof: The Cauchy-Riemann equations imply that v, if it exists, is unique up to adding a constant. For existence, first suppose Ω is simply connected. Define $f(x+iy) = \frac{\partial u}{\partial x}(x,y) - i\frac{\partial u}{\partial y}(x,y)$ on Ω. Then the real and imaginary parts of f have continuous first-order partials, so f is differentiable in the vector calculus sense, and the partials satisfy the Cauchy-Riemann equations by harmonicity of u and Clairaut's Theorem. Thus f is holomorphic on Ω. By (), f has an antiderivative $F = U + iV$ on Ω. We have

$$F' = \frac{\partial U}{\partial x} - i\frac{\partial U}{\partial y} = f = \frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y}$$

by XII.1.1.5., so U has the same partials as u, and hence differs from u by a (real) constant. Correcting F by subtracting this constant, we have $U = u$, and we may take $v = V$.

Since every point of a general region Ω has a simply connected open neighborhood (e.g. a disk) contained in Ω, the first statement follows. 👍

Here are some crucial consequences:

XII.11.2.4. Corollary. Let Ω be a region in $\mathbb{R}^2$. Any harmonic function on Ω is real-analytic, and in particular $\mathcal{C}^\infty$.

XII.11.2.5. Corollary. Let Λ and Ω be regions in $\mathbb{C}$. If $u : \Omega \to \mathbb{R}$ is harmonic and $\phi : \Lambda \to \Omega$ is (complex-)analytic, then $u \circ \phi$ is harmonic on Λ.

Proof: Let $w_0 \in \Lambda$, and $z_0 = \phi(w_0)$. Let D be an open disk around z_0 contained in Ω, and $W = \phi^{-1}(D)$. Then W is an open neighborhood of w_0. There is an analytic function ψ on D with real part $u|_D$. So $\psi \circ \phi$ is analytic on W, and its real part is the restriction of $u \circ \phi$ to W. Thus $u \circ \phi$ is locally harmonic, hence harmonic. 👍

XII.11.3. The Mean Value Property and Local Extrema

We continue to only work in regions of $\mathbb{R}^2$. Using XII.11.2.3., we have the following important consequence of XII.3.3.3.:

XII.11.3.1. Theorem. Let u be harmonic in $\Omega \subseteq \mathbb{R}^2$, and let γ be a circle centered at $z_0 = (x_0, y_0) \in \Omega$ such that γ and its interior are contained in Ω. Then $u(x_0, y_0)$ is the average value of u around γ, i.e. if γ has radius r, then

$$u(x_0, y_0) = \frac{1}{2\pi} \int_0^{2\pi} u(z_0 + re^{i\theta})\, d\theta$$

and $u(x_0, y_0)$ is also the average value of u over the disk D bounded by γ, i.e.

$$u(x_0, y_0) = \frac{1}{\pi r^2} \iint_D u(x, y)\, dx\, dy \ .$$

Proof: The function u is the real part of an analytic function f on a slightly larger disk containing γ, and so the first formula follows from XII.3.3.3. by taking the real part. For the second part, integrating using polar coordinates and applying the first part to each circle inside γ, we have

$$\frac{1}{\pi r^2} \iint_D u(x, y)\, dx\, dy = \frac{1}{\pi r^2} \int_0^r \left[\int_0^{2\pi} u(z_0 + se^{i\theta})\, d\theta \right] s\, ds = \frac{1}{\pi r^2} \int_0^r [2\pi u(x_0, y_0)]\, s\, ds$$

$$= \frac{2\pi u(x_0, y_0)}{\pi r^2} \int_0^r s\, ds = \frac{2u(x_0, y_0)}{r^2} \frac{r^2}{2} = u(x_0, y_0) \ .$$

👍

XII.11.3.2. A continuous real-valued function u on $\Omega \subseteq \mathbb{R}^2$ satisfying the conclusion of the theorem for each circle γ is said to have the *mean value property*. (Note we do not assume a function with the mean value property is differentiable). There is a converse to XII.11.3.1.: a (continuous) function from Ω to $\mathbb{R}$ with the mean value property is necessarily harmonic (XII.11.4.8.) (and in particular $\mathcal{C}^\infty$).

Since the average value of a nonconstant real-valued continuous function over a closed bounded region is strictly less than the maximum in the region (), we have:

XII.11.3.3. Corollary. Let u have the mean value property on a region $\Omega \subseteq \mathbb{R}^2$. If u has a local maximum or minimum at a point $(x_0, y_0) \in \Omega$, then u is constant in a neighborhood of (x_0, y_0).

Proof: If u has a local maximum at (x_0, y_0), then there is a closed disk D centered at (x_0, y_0) with $u(x_0, y_0) = \max\{u(x, y) : (x, y) \in D\}$, and it follows that u is constant on D. The statement for local minima follows by considering $-u$, which also has the mean value property. 👍

Using the Rigidity Theorem, we get a vast extension (given XII.11.4.8.):

XII.11.3.4. Theorem. Let u be harmonic on a connected region $\Omega \subseteq \mathbb{R}^2$. If u has a local maximum or minimum at a point $(x_0, y_0) \in \Omega$, then u is constant on Ω.

Proof: Let $u(x_0, y_0) = \alpha$. Let

$$W = \{(x, y) \in \Omega : u \equiv \alpha \text{ in a neighborhood of } (x, y)\} .$$

Then W is open, and nonempty since $(x_0, y_0) \in W$. Let $(x_1, y_1) \in \partial W \cap \Omega$, and let D be an open disk centered at (x_1, y_1) contained in Ω. Then u is the real part of an analytic function f on D. Since the real part of f is constant on $D \cap W$, by the Cauchy-Riemann equations f is constant on $D \cap W$. Thus by the Rigidity Theorem f, and therefore u, is constant on D, and so $(x_1, y_1) \in W$. Therefore W is ralatively closed in Ω. Since Ω is connected, $W = \Omega$. 👍

This argument does not work directly if u just has the mean value property. But by a similar argument we get the following corollary of XII.11.3.4.:

XII.11.3.5. Corollary. Let Ω be a connected region in $\mathbb{C}$. Let u be continuous on $\bar{\Omega}$, with the mean value property (e.g. harmonic) on Ω, and not constant. If u has a maximum [resp. minimum] on $\bar{\Omega}$ (which will be automatic if Ω is bounded), then the maximum [resp. minimum] of u on $\bar{\Omega}$ can occur only on $\partial\Omega$.

Proof: Suppose α is the maximum value of u on $\bar{\Omega}$, and suppose $(x_0, y_0) \in \Omega$ with $u(x_0, y_0) = \alpha$. Set $W = \{(x, y) \in \Omega : u(x, y) = \alpha\}$. Then W is nonempty and relatively closed in Ω since u is continuous. If $(x_1, y_1) \in W$, then by XII.11.3.4. u is constant in a neighborhood of (x_1, y_1), so W is open. Since Ω is connected, $W = \Omega$. For the minimum, use $-u$. 👍

XII.11.3.6. Corollary. Let Ω be a region in $\mathbb{C}$. Let u and v be continuous on $\bar{\Omega}$, with the mean value property (e.g. harmonic) on Ω. If $u \equiv v$ on $\partial\Omega$, then $u \equiv v$ on all of $\bar{\Omega}$.

Proof: If Ω_0 is a connected component of Ω, then $u - v$ is identically 0 on $\partial\Omega_0$, hence is identically 0 on Ω_0. 👍

We also get a harmonic analog of the Rigidity Theorem. The proof is very similar to the proof of XII.11.3.4..

XII.11.3.7. Theorem. [Rigidity Theorem for Harmonic Functions] Let Ω be a connected region in $\mathbb{R}^2$, and let u and v be harmonic on Ω. If there is a sequence (x_n, y_n) in Ω with $(x_n, y_n) \to (x, y) \in \Omega$ and $u(x_n, y_n) = v(x_n, y_n)$ for all n, then $u \equiv v$ on all of Ω.

XII.11.4. The Dirichlet Problem

The Dirichlet Problem is a PDE boundary-value problem with many applications in physics:

XII.11.4.1. Problem. [Dirichlet Problem] Let Ω be a (usually bounded) region in $\mathbb{R}^n$, and let f be a real-valued continuous function on $\partial\Omega$. Does there exist a function $u : \bar{\Omega} \to \mathbb{R}$ which is continuous on $\bar{\Omega}$, harmonic on Ω, with $u|_{\partial\Omega} = f$? (In other words, can f be extended to $\bar{\Omega}$ to be harmonic on its interior?)

XII.11.4.2. XII.11.3.6. implies that if there is a solution, it is unique.

We will again restrict to the case $n = 2$.

XII.11.4.3. The Dirichlet Problem is not always solvable. Let Ω be the open unit disk with the origin removed. Then $\partial\Omega$ is the union of $\{0\}$ and the unit circle. If $f(0) = 1$ and f is identically 0 on the unit circle, the problem is not solvable [by uniqueness, a solution would have to be rotationally symmetric; apply XII.11.5.1.(c)].

In this example, Ω is not a regular open set, i.e. the interior of $\bar{\Omega}$ is larger than Ω. It is harder to give an example of a regular Ω for which the problem is not solvable ($\partial\Omega$ must be very complicated ()).

The Dirichlet Problem can be solved on $\mathbb{D}$, by explicit formula:

XII.11.4.4. Theorem. Let $f : [0, 2\pi] \to \mathbb{R}$ be a continuous function with $f(0) = f(2\pi)$. Then the function

$$u(re^{i\phi}) = \int_0^{2\pi} f(\theta) \frac{1 - r^2}{2\pi[1 - 2r\cos(\theta - \phi) + r^2]}\, d\theta$$

is harmonic on $\mathbb{D}$, and

$$\lim_{re^{i\phi} \to e^{i\theta}} u(re^{i\phi}) = f(\theta)$$

for all $\theta \in [0, 2\pi]$, i.e. u extends to a continuous function on $\bar{\mathbb{D}}$ agreeing with f on the unit circle. So u is the unique solution to the Dirichlet Problem on $\mathbb{D}$ for f.

See e.g. () for a proof. More generally, the Dirichlet Problem is solvable on any disk just by scaling.

XII.11.4.5. If Ω is a bounded simply connected region, and some (hence any) conformal bijection ϕ from Ω to $\mathbb{D}$ extends to a homeomorphism from $\bar{\Omega}$ to $\bar{\mathbb{D}}$, then the Dirichlet problem on Ω can be transferred to $\mathbb{D}$ via ϕ, so is solvable for any f. This condition is satisfied if $\partial\Omega$ is reasonably nice, e.g. a smooth curve or a polygon. But for this argument $\partial\Omega$ must be at least homeomorphic to a circle, and even this is not sufficient. See () for a precise description of the regions on which the problem is solvable.

XII.11.4.6. The Dirichlet Problem is also solvable if the boundary function is "not too discontinuous" ().

XII.11.4.7. Here is a simple but important example of a physical situation leading to the Dirichlet Problem. Let P be a thin metal plate whose boundary is kept at a known temperature which varies around the edge (but is constant in time). If the plate has reached steady-state temperature, the temperature at interior points is a harmonic function which is the solution to the Dirichlet Problem for the known boundary temperatures.

The Converse of the Mean Value Property

Using the fact that the Dirichlet Problem is solvable on a disk, we can obtain a converse to XII.11.3.1.:

XII.11.4.8. PROPOSITION. Let Ω be a region in $\mathbb{R}^2$, and let u be a (continuous) real-valued function on Ω with the mean value property. Then u is harmonic on Ω.

PROOF: Let $(x_0, y_0) \in \Omega$, and let D be an open disk centered at (x_0, y_0) with $\bar{D} \subseteq \Omega$. Let f be the restriction of u to ∂D. Then f is continuous, so there is a continuous function v on $\bar{D}$ which is harmonic on D and which agrees with f on ∂D. Then u and v are continuous functions on $\bar{D}$ with the mean value property, and $u \equiv v$ on ∂D, so $u \equiv v$ on $\bar{D}$ by XII.11.3.6.. In particular, u is harmonic on the open neighborhood D of (x_0, y_0). Since $(x_0, y_0) \in \Omega$ is arbitrary, u is harmonic on Ω. 👍

XII.11.4.9. Note that the analogous statement about complex-valued functions is false. If Ω is a region in $\mathbb{C}$ and f a complex-valued function on $\mathbb{C}$, then f has the mean value property if and only if its real and imaginary parts are harmonic, with no necessary relation between them, so f need not be analytic (indeed, the imaginary part may be zero, so f is just harmonic!)

XII.11.5. Exercises

XII.11.5.1. Let Ω be a region in $\mathbb{R}^2$ not containing the origin, and let u be a real-valued function $u(r, \theta)$ on Ω written in polar coordinates.
(a) Show that $\Delta u(r, \theta) = \frac{1}{r}(ru_r)_r + \frac{1}{r^2}u_{\theta\theta} = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}$.

(b) Show that if u is a function of r alone (i.e. its value is "rotationally symmetric"), and harmonic, then $u(r) = a \log r + b$, where a and b are (real) constants. This function with $a = \frac{1}{2\pi}$ and $b = 0$ is called the *fundamental solution* of Laplace's equation (in two variables). It can be used to give explicit formulas for solutions to the Dirichlet and Neumann problems. (In some references, this u is described as having a "pole" at 0. But this is potentially misleading since u is not the real part of an analytic function with a pole at 0, or any other kind of isolated singularity.)

(c) If u is a nonconstant rotationally symmetric harmonic function on the unit disk with the origin removed, then u cannot be extended to a function continuous at 0.

XII.11.5.2. This problem is a continuation of Problem XII.11.5.1.. Give an alternate description of the solution to the Dirichlet problem on the unit disk $\mathbb{D}$, as follows.

(a) Show directly that the functions $u(r,\theta) = r^k \cos k\theta$, $u(r,\theta) = r^k \sin k\theta$, and more generally $u(r,\theta) = r^k \cos(k\theta + \alpha)$, for $k \in \mathbb{Z}$ and $\alpha \in \mathbb{R}$, is harmonic on $\mathbb{C} \setminus \{0\}$ (all of $\mathbb{C}$ if $k \geq 0$). [A simple indirect proof: this function is the real part of $f(z) = cz^k$ for some $c \in \mathbb{C}$, $|c| = 1$.]

(b) Let f be a real-valued continuous function on the unit circle in $\mathbb{C}$, thought of as a function on $[0, 2\pi]$ with $f(2\pi) = f(0)$. Suppose the Fourier series

$$f(\theta) = \sum_{k=0}^{\infty} a_k \cos k\theta + \sum_{k=1}^{\infty} b_k \sin k\theta$$

converges in an appropriate sense (cf. ()). Show that the series

$$u(r,\theta) = \sum_{k=0}^{\infty} a_k r^k \cos k\theta + \sum_{k=1}^{\infty} b_k r^k \sin k\theta$$

gives the solution to the Dirichlet problem on $\mathbb{D}$ for f.

XII.11.5.3. Give alternate proofs of the results of this section without using complex analysis, and formulate and prove analogous theorems for harmonic functions on regions in $\mathbb{R}^n$. (See e.g. ().)

XII.12. Differential Algebra

Differential algebra is, roughly speaking, the algebraic theory of calculus. Versions appear in quite different contexts. A basic application concerns solutions of differential equations, in particular answering questions about whether antiderivatives or other solutions to differential equations have elementary closed formulas. A quite different context is the study of one-parameter automorphism groups of operator algebras, and in forms of amenability.

XII.12.1. Derivations

A derivation on a ring is an algebraic operation which abstracts the essential properties of the derivative, specifically the product rule.

XII.12.1.1. DEFINITION. Let R be a ring. A *derivation* on R is a function $\delta : R \to R$ with the properties

$$\delta(x+y) = \delta(x) + \delta(y)$$

$$\delta(xy) = \delta(x)y + x\delta(y)$$

for all $x, y \in R$. If R is an algebra over a field $\mathbb{F}$, we often (but not always) expand the first condition to require that δ be $\mathbb{F}$-linear, i.e. we add the requirement

$$\delta(\alpha x) = \alpha\delta(x)$$

for all $\alpha \in \mathbb{F}$, $x \in R$ (cf. XII.12.1.5.(iii)).

A *differential subring* of a differential ring (R, δ) is a subring S which is closed under δ; the restriction of δ to S is thus a derivation on S.

XII.12.1.2. Note that R need not be commutative; if R is not commutative, it is crucial to respect the order of the factors in the product rule terms (since addition is always commutative, the order of the terms does not matter).

Derivations can be defined more generally from a ring R to an R-bimodule M with the same properties; but we will primarily stick to the case of derivations from R to R, which are general enough for most purposes.

XII.12.1.3. EXAMPLES. (i) Let I be an interval in $\mathbb{R}$, and R be a ring of real-valued differentiable functions on I which is closed under differentiation, e.g. $\mathcal{C}^\infty(I)$ or the set of polynomials, and set $\delta(f) = f'$. Then δ is a derivation on R. This is the prototype example.

R must be closed under differentiation: a ring such as $\mathcal{C}^1(I)$ will not work, since the derivative of a $\mathcal{C}^1$ function is not in $\mathcal{C}^1$ in general. We could more generally define differentiation as a derivation from $\mathcal{C}^p(I)$ to the $C^p(I)$-bimodule $C^{p-1}(I)$ for any $p \geq 1$.

(ii) Similarly, if U is an open set in $\mathbb{R}^2$ and $R = \mathcal{C}^\infty(U)$, set $\delta_x(f) = \frac{\partial f}{\partial x}$ and $\delta_y(f) = \frac{\partial f}{\partial y}$. Then δ_x and δ_y are derivations on R.

(iii): Let $\mathbb{F}$ be a commutative ring, usually a field. On either the polynomial ring $\mathbb{F}[X]$ or the ring of formal power series $\mathbb{F}[[X]]$ (III.5.3.2.) define δ by

$$\delta\left(\sum_{k=0}^{\infty} a_k X^k\right) = \sum_{k=1}^{\infty} k a_k X^{k-1}$$

(the sum is finite in the $\mathbb{F}[X]$ case). Then δ is a derivation. This derivation on $\mathbb{F}[X]$ can be extended to the field $\mathbb{F}(X)$ of rational functions using the "quotient rule." An extension to the field of formal Laurent series $\mathbb{F}((X))$ (III.5.3.3.) can be made in the evident way. Similarly, on $\mathbb{F}[X,Y]$ or $\mathbb{F}[[X,Y]]$, derivations δ_X and δ_Y can be defined in the same way.

(iv): Let $\mathbb{F}$ be a unital commutative ring, usually a field, and let $X, X_1, X_2, \ldots$ be a sequence of indeterminates. Form the polynomial ring

$$R = \mathbb{F}[X, X_1, X_2, \ldots] \ .$$

We will regard X_n as the "n'th derivative" of X (we may think of X as X_0). Specifically, we define a derivation δ on R with $\delta(X_k) = X_{k+1}$ for all $k \geq 0$. Linearity and the product rule then dictates the following formula for δ:

$$\delta\left(\sum_{k,n} \alpha_{kn} X_k^n\right) = \sum_{k,n} n\alpha_{k,n} X_k^{n-1} X_{k+1}$$

(the sum is a finite sum). It is easily checked that δ is a derivation on R. The ring R with this derivation is called the *ring of differential polynomials* in X, over $\mathbb{F}$, denoted $\mathbb{F}\{X\}$. We may similarly form rings of differential polynomials in more than one variable, e.g. $\mathbb{F}\{X,Y\}$, with "partial derivatives" δ_X, δ_Y.

(v) Here is an important class of derivations of interest only on noncommutative rings. Let R be a ring, and $a \in R$. Set

$$\delta_a(x) = [a,x] = ax - xa \ .$$

Then δ_a is a derivation on R, called the *inner derivation* defined by a. It is the zero derivation if and only if a is in the center of R (e.g. if R is commutative).

More generally, if M is an R-bimodule and $a \in M$, then δ_a is a derivation from R to M. Amenability is characterized by certain types of derivations always being inner ().

(vi) We illustrate the connection between inner derivations and derivatives with the following example, which is a (very) special case of the theory to be developed in (). Let $\mathcal{H}$ be a (complex) Hilbert space, and let H be a self-adjoint operator in $\mathcal{B}(\mathcal{H})$. For each $t \in \mathbb{R}$, let $U_t = e^{itH}$. Then U_t is a unitary operator on $\mathcal{H}$, $U_{s+t} = U_s U_t$ for all s,t, and $t \mapsto U_t$ is norm-continuous. Consider the inner derivation $\delta = \delta_H$ on $\mathcal{B}(\mathcal{H})$, i.e. $\delta(X) = HX - XH$. Then δ is a derivation on $\mathcal{B}(\mathcal{H})$, and is also a bounded operator on $\mathcal{B}(\mathcal{H})$, i.e. $\delta \in \mathcal{B}(\mathcal{B}(\mathcal{H}))$. Thus we may form the operator

$$\alpha_t = e^{t\delta} \in \mathcal{B}(\mathcal{B}(\mathcal{H}))$$

for any $t \in \mathbb{R}$. Then α_t is a *-automorphism of $\mathcal{B}(\mathcal{H})$ for each t, $\alpha_{s+t} = \alpha_s \circ \alpha_t$ for all s,t, and $t \mapsto \alpha_t$ is norm-continuous ($\{\alpha_t\}$ is a *norm-continuous one-parameter group of automorphisms* of $\mathcal{B}(\mathcal{H})$). We have

$$\alpha_t(X) = U_t X U_t^*$$

$$H = \lim_{t\to 0} \frac{U_t - I}{t}$$

$$\delta(X) = \lim_{t\to 0} \frac{\alpha_t(X) - X}{t}$$

for all $X \in \mathcal{B}(\mathcal{H})$. Every norm-continuous one-parameter automorphism group of $\mathcal{B}(\mathcal{H})$ is of this form (), i.e. every bounded *-derivation on $\mathcal{B}(\mathcal{H})$ is inner.

The formal relation between derivations and one-parameter automorphism groups is quite simple (if convergence questions are ignored): if (α_t) is a one-parameter automorphism group and we define

$$\delta(X) = \lim_{t\to 0} \frac{\alpha_t(X) - X}{t}$$

(assuming the limit exists for all X), then for any X and Y we have

$$\delta(XY) = \lim_{t\to 0} \frac{\alpha_t(XY) - I}{t} = \lim_{t\to 0} \frac{\alpha_t(X)\alpha_t(Y) - X\alpha_t(Y) + X\alpha_t(Y) - I}{t}$$

$$= \lim_{t\to 0} \frac{\alpha_t(X) - I}{t} Y + X \lim_{t\to 0} \frac{\alpha_t(Y) - I}{t} = \delta(X)Y + X\delta(Y)$$

so δ is a derivation. Conversely, if δ is a derivation, a formal power-series calculation (assuming the appropriate series converge) shows that e^δ is an automorphism. Thus derivations are "derivatives" of one-parameter automorphism groups.

One-parameter groups of automorphisms of $\mathcal{B}(\mathcal{H})$ which are only point-norm continuous (strongly continuous) are more important in applications, e.g. in the time evolution of a dynamical system, and have a similar description except that the H is unbounded and only densely defined in general, as is the derivation δ.

Constants

XII.12.1.4. Let δ be a derivation on a ring R. An element $x \in R$ is a *constant* (for δ) if $\delta(x) = 0$. Write R_0 for the set of constants for δ.

XII.12.1.5. Proposition. Let δ be a derivation on a ring R, and R_0 the set of constants for δ. Then

(i) R_0 is a subring of R.

(ii) If R is a unital ring, then $1 \in R_0$, i.e. R_0 is a unital subring of R.

(iii) If K is a subring of R_0 which is contained in the center of R, and R is regarded as an algebra over K, then δ is K-linear, i.e.

$$\delta(\alpha x) = \alpha\delta(x)$$

for all $\alpha \in K$, $x \in R$.

Proof: Parts (i) and (iii) are obvious. For (ii), we have

$$\delta(1) = \delta(1 \cdot 1) = \delta(1)1 + 1\delta(1) = 2\delta(1)$$

so $\delta(1) = 0$. 👍

XII.12.1.6. If δ is a derivation on a ring R, and $\alpha \in R_0$ is in the center of R, then $\alpha\delta$ is also a derivation on R, where $(\alpha\delta)(x) = \delta(\alpha x)$:

$$(\alpha\delta)(xy) = \delta(\alpha xy) = \delta(\alpha)\delta(xy) + \alpha\delta(xy) = \alpha\delta(x)y + \alpha x\delta(y) = \alpha\delta(x)y + x\alpha\delta(y)\ .$$

XII.12.1.7. EXAMPLES. (i) In XII.12.1.3.(i), the constants are exactly the constant functions.

(ii) In XII.12.1.3.(ii), if U is connected, the constants for δ_x are just the functions whose value depends only on y. But if U is not connected, the constants for the derivation are a little more complicated (functions of only y on each component of U).

(iii)–(iv) In XII.12.1.3.(iii)–(iv), the constants are just the elements of $\mathbb{F}$.

(v) In XII.12.1.3.(v), the constants for the inner derivation δ_a are exactly the elements of R which commute with a.

(vi) The ring of constants is not necessarily commutative. In the extreme case, if δ is the zero derivation on any ring R, then $R_0 = R$.

XII.12.2. Differential Fields

In the subject of differential algebra, derivations on fields are usually considered. In this subsection, all rings will be commutative.

XII.12.2.1. DEFINITION. A *differential field* is a pair $(\mathbb{F}, \delta)$, where $\mathbb{F}$ is a field of characteristic zero (III.4.3.2.) and δ a derivation on $\mathbb{F}$. (In some references the characteristic zero requirement is omitted.)

The restriction to fields is rather natural, due to the following result:

XII.12.2.2. THEOREM. Let R be a unital commutative ring, and D a subset of R closed under multiplication, $1 \in D$, which contains neither 0 nor any zero divisors. If δ is a derivation on R, then δ extends uniquely to a derivation $\bar{\delta}$ on the ring $D^{-1}R$ of quotients (III.4.3.4.). The extension is given by the "quotient rule"

$$\bar{\delta}\left(\frac{x}{y}\right) = \frac{y\delta(x) - x\delta(y)}{y^2}$$

for all $x \in R$, $y \in D$. In particular, if R is an integral domain and δ is a derivation on R, then δ extends uniquely to the quotient field of R. If $x \in R$ and $y \in D$ are constants for δ, then $\frac{x}{y}$ is a constant for $\bar{\delta}$.

PROOF: Recall that elements of $D^{-1}R$ can be symbolically written as $\frac{x}{y}$ with $x \in R$, $y \in D$, with $\frac{x}{y} = \frac{z}{w}$ if and only if $xw = yz$. (We should really write $\left[\frac{x}{y}\right]$ or $[(x, y)]$ instead of $\frac{x}{y}$.)

To motivate the definition of $\bar{\delta}$ and prove uniqueness, suppose $\tilde{\delta}$ is a derivation on $D^{-1}R$ extending δ. For each $y \in D$, we have, by XII.12.1.5.(ii),

$$0 = \tilde{\delta}(1) = \tilde{\delta}\left(y\frac{1}{y}\right) = y\tilde{\delta}\left(\frac{1}{y}\right) + \frac{1}{y}\tilde{\delta}(y) = y\tilde{\delta}\left(\frac{1}{y}\right) + \frac{1}{y}\delta(y)$$

$$\tilde{\delta}\left(\frac{1}{y}\right) = -\frac{1}{y^2}\delta(y)$$

and if $x \in R$, we have

$$\tilde{\delta}\left(\frac{x}{y}\right) = \tilde{\delta}\left(x\frac{1}{y}\right) = \frac{1}{y}\tilde{\delta}(x) + x\tilde{\delta}\left(\frac{1}{y}\right) = \frac{1}{y}\delta(x) - \frac{x}{y^2}\delta(y) = \frac{y\delta(x) - x\delta(y)}{y^2} \ .$$

We now show that $\bar{\delta}$ is well defined. We first suppose $x \in \mathbb{R}$ and $y, z \in D$. Then

$$\bar{\delta}\left(\frac{xz}{yz}\right) = \frac{yz\delta(xz) - xz\delta(yz)}{y^2z^2} = \frac{yz(x\delta(z) + z\delta(x)) - xz(y\delta(z) + z\delta(y))}{y^2z^2}$$

$$= \frac{yz^2\delta(x) - xz^2\delta(y)}{y^2z^2} = \frac{y\delta(x) - x\delta(y)}{y^2} = \bar{\delta}\left(\frac{x}{y}\right) .$$

Now if $\frac{x}{y} = \frac{z}{w}$, i.e. $xw = yz$, we have

$$\bar{\delta}\left(\frac{x}{y}\right) = \bar{\delta}\left(\frac{xyw^2}{y^2w^2}\right) = \bar{\delta}\left(\frac{y^2zw}{y^2w^2}\right) = \bar{\delta}\left(\frac{z}{w}\right) .$$

The embedding of R into $D^{-1}R$ sends x to $\frac{x}{1}$, and

$$\bar{\delta}\left(\frac{x}{1}\right) = \frac{1 \cdot \delta(x) - x\delta(1)}{1^2} = \delta(x)$$

since $\delta(1) = 0$. Thus $\bar{\delta}$ extends δ.

To show that $\bar{\delta}$ is a derivation, let $x, z \in R$ and $y, w \in D$. Then

$$\bar{\delta}\left(\frac{x}{y} \cdot \frac{z}{w}\right) = \bar{\delta}\left(\frac{xz}{yw}\right) = \frac{yw\delta(xz) - xz\delta(yw)}{y^2w^2} = \frac{yw(x\delta(z) + z\delta(x)) - xz(y\delta(w) + w\delta(y))}{y^2w^2}$$

$$= \frac{ywz\delta(x) - xzw\delta(y) + xyw\delta(z) - xyz\delta(w)}{y^2w^2} = \frac{yz\delta(x) - xz\delta(y)}{y^2w} + \frac{xw\delta(z) - xz\delta(w)}{yw^2} = \frac{z}{w}\bar{\delta}\left(\frac{x}{y}\right) + \frac{x}{y}\bar{\delta}\left(\frac{z}{w}\right)$$

so $\bar{\delta}$ is a derivation.

The last statement is obvious from the formula for $\bar{\delta}$. 👍

A corollary of the first part of the proof is:

XII.12.2.3. Corollary. Let δ be a derivation on a unital ring R. If $y \in R$ is invertible, then

$$\delta(y^{-1}) = -y^{-1}\delta(y)y^{-1} .$$

In particular, if y is a constant for δ, so is y^{-1}. So if $(\mathbb{F}, \delta)$ is a differential field, the ring of constants for δ is a subfield of $\mathbb{F}$.

Here are the most important examples of differential fields for our purposes:

XII.12.2.4. EXAMPLES. (i) Let $\mathbb{F}$ be a field of characteristic zero (e.g. $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$). The polynomial ring $\mathbb{F}[X]$ in one variable over $\mathbb{F}$ becomes a differential ring under the derivation

$$\delta\left(\sum_{k=0}^{n}\alpha_k X^k\right)=\sum_{k=1}^{n}k\alpha_k X^{k-1}$$

and the quotient field $\mathbb{F}(X)$ of rational functions in one variable over $\mathbb{F}$ becomes a differential field with the unique extension from XII.12.2.2. given by the quotient rule.

There is a minor but technically important point about this example and the succeeding ones which is glossed over in many references on differential fields. The elements of $\mathbb{F}(X)$ are called "rational functions," but they are not exactly functions from $\mathbb{F}$ (or a fixed subset) to $\mathbb{F}$ since functions which agree except at finitely many points are identified. Thus, for example, $f(X)=X+1$, $g(X)=\frac{X^2-1}{X-1}$, and $h(X)=\frac{X^2-X-2}{X-2}$ all represent the same element of $\mathbb{F}(X)$ even though they represent different functions since their domains are different.

It is customary in differential algebra to play a little fast and loose with domains of "functions" since it turns out that the domains and other functional properties are not really relevant to the algebra, but one should be careful in the process.

(ii) Let Ω be a connected open set in $\mathbb{C}$. Let $\mathbb{M}(\Omega)$ be the set of meromorphic functions on Ω, complex-valued functions which are differentiable on Ω except for isolated poles (or removable singularities). We again identify two functions if they agree when all removable singularities have been filled in, i.e. if the functions agree on a nonempty open subset of Ω, or more generally on a subset of Ω with a limit point in Ω (XII.5.1.4.). The reciprocal of a meromorphic function on Ω which is not identically zero is again a meromorphic function on Ω, so $\mathbb{M}(\Omega)$ is a field. The usual complex differentiation is a derivation on $\mathbb{M}(\Omega)$ making it into a differential field. $\mathbb{C}(X)$ may be regarded as a differential subfield of $\mathbb{M}(\Omega)$ for any Ω. If $\Omega_1\subseteq\Omega_2$, then restriction gives a natural embedding of $\mathbb{M}(\Omega_2)$ into $\mathbb{M}(\Omega_1)$, so $\mathbb{M}(\Omega_2)$ can be regarded as a subfield of $\mathbb{M}(\Omega_1)$. (It is a proper subfield if $\Omega_2\neq\Omega_1$: for example, if $z_0\in\Omega_2\setminus\Omega_1$, then $f(z)=e^{1/(z-z_0)}$ is in $\mathbb{M}(\Omega_1)$ but not $\mathbb{M}(\Omega_2)$.) The field of constants in these differential fields is $\mathbb{C}$ (i.e. the usual constant functions).

If Ω is not connected, then $\mathbb{M}(\Omega)$ has zero divisors and is not a field: a function which is identically zero on one component of Ω but not on all of Ω is a zero divisor.

Differential Function Fields

These examples lead to a convenient definition of the principal objects of study in elementary differential algebra:

XII.12.2.5. DEFINITION. A *differential function field* or *dff* is a differential subfield of $\mathbb{M}(\Omega)$ containing $\mathbb{C}(X)$, for some nonempty connected open subset Ω of $\mathbb{C}$.

As mentioned earlier, the exact set Ω will turn out to be unimportant, and we will feel free to reduce it whenever necessary. The next convenient result illustrates the technique:

XII.12.2.6. PROPOSITION. Let $\mathbb{F}$ be a differential function field, and $\mathbb{K}$ a finite extension of $\mathbb{F}$. Then $\mathbb{K}$ is isomorphic to a differential function field, and the derivation on $\mathbb{K}$ is uniquely determined by the derivation on $\mathbb{F}$.

PROOF: Identify $\mathbb{F}$ with a subfield of $\mathbb{M}(\Omega)$ for some $\Omega \subseteq \mathbb{C}$. Suppose $\mathbb{K} = \mathbb{F}(\phi)$ for some $\phi \in \mathbb{K}$ ($\mathbb{K}$ can be constructed from $\mathbb{F}$ by finitely many such steps; in fact, it can be done in one step by the Primitive Element Theorem of algebra). Then ϕ is a root of an irreducible polynomial $p(Y)$ with coefficients in $\mathbb{F}$. Regard p as a function $p(z, w)$ from $\Omega \times \mathbb{C}$ to $\mathbb{C}$ which is analytic except where z is a pole of a coefficient function of p. By the Analytic Implicit Function Theorem (VIII.5.10.4.), there is an open set $\Omega' \subseteq \Omega$ and an analytic function g on Ω' such that $p(z, g(z)) = 0$ for all $z \in \Omega'$. The embedding of $\mathbb{K}$ into $\mathbb{M}(\Omega')$ which is restriction on $\mathbb{F}$ and which sends ϕ to g represents $\mathbb{K}$ as a differential function field. 👍

Algebraic Extensions of Derivations

The uniqueness of the derivation follows from a more general version of this result:

XII.12.2.7. PROPOSITION. Let $(\mathbb{F}, \delta)$ be a differential field, and $\mathbb{K}$ an algebraic extension of $\mathbb{F}$. Then δ extends uniquely to a derivation on $\mathbb{K}$.

PROOF: [Ros72] Define two derivations δ_0, δ_1 on $\mathbb{F}[X]$ by

$$\delta_0\left(\sum_{k=0}^{n} \alpha_k X^k\right) = \sum_{k=0}^{n} \delta(\alpha_k) X^k$$

$$\delta_1\left(\sum_{k=0}^{n} \alpha_k X^k\right) = \sum_{k=1}^{n} k\alpha_k X^{k-1}$$

(it is routine to check that δ_0 and δ_1 are derivations). Let $\tilde{\delta}$ be a derivation on $\mathbb{K}$ extending δ. If $p(X) = \sum_{k=0}^{n} \alpha_k X^k$, then for any $z \in \mathbb{K}$,

$$0 = \tilde{\delta}(p(z)) = \sum_{k=0}^{n} \tilde{\delta}(\alpha_k z^k) = \sum_{k=0}^{n} \tilde{\delta}(\alpha_k) z^k + \sum_{k=1}^{n} \alpha_k z^{k-1} \tilde{\delta}(z) = (\delta_0(p))(z) + (\delta_1(p))(z)\tilde{\delta}(z) \ .$$

Let f be the irreducible polynomial of z over $\mathbb{F}$. Taking $p = f$, we have

$$0 = \tilde{\delta}(0) = \tilde{\delta}(f(z)) = (\delta_0(f))(z) + (\delta_1(f))(z)\tilde{\delta}(z)$$

$$(\delta_1(f))(z)\tilde{\delta}(z) = -(\delta_0(f))(z) \ .$$

But $\delta_1(f)$ is the "derivative" of f in the usual algebraic sense. Since $\mathbb{F}$ has characteristic zero, z is not a root of $\delta_1(f)$ (), i.e. $(\delta_1(f))(z) \neq 0$, and hence

$$\tilde{\delta}(z) = -\frac{(\delta_0(f))(z)}{(\delta_1(f))(z)} \ .$$

Thus an extension of δ to $\mathbb{K}$ (if one exists) is unique, and for any $z \in \mathbb{K}$, we have that $\tilde{\delta}(z) \in \mathbb{F}(z)$.

To show existence, we may assume $\mathbb{K} = \mathbb{F}(z)$ for some z. We have $\mathbb{K} = \mathbb{F}(z) \cong \mathbb{F}[X]/(f(X))$, where $f(X)$ is the irreducible polynomial of z over $\mathbb{F}$. We need only show that the derivation

$$\bar{\delta} = \delta_0 - \frac{(\delta_0(f))(z)}{(\delta_1(f))(z)}\delta_1$$

of $\mathbb{F}[X]$ drops to a derivation of $\mathbb{F}[X]/(f(X))$ (XII.12.4.5.), since this drop will extend δ. Thus we only need to show that $\bar{\delta}$ maps the ideal $(f(X))$ into itself, i.e. that $\delta(f)$ is a multiple of f, or equivalently that z is a root of $\bar{\delta}(f)$. But

$$(\bar{\delta}(f))(z) = (\delta_0(f))(z) - \frac{(\delta_0(f))(z)}{(\delta_1(f))(z)}(\delta_1(f))(z) = 0 \ .$$



Exponentials and Logarithms

XII.12.2.8. Definition. Let $(\mathbb{F}, \delta)$ be a differential field, and $x, y \in \mathbb{F}$, $y \neq 0$. Then y is an *exponential* of x, and x is a *logarithm* of y, if $\delta(y) = y\delta(x)$, or equivalently $\delta(x) = \frac{\delta(y)}{y}$.

XII.12.2.9. Proposition. (i) If y_1 is an exponential of x, then y_2 is an exponential of x if and only if $y_2 = cy_1$ for some nonzero constant c.
(ii) If x_1 is a logarithm for y, then x_2 is a logarithm for y if and only if $x_2 = x_1 + c$ for some constant c.

Proof: (ii) is trivial since $\delta(x_2) = \delta(x_1)$ if and only if $\delta(x_1 - x_2) = 0$.
(i): The following statements are equivalent for $y_2 \neq 0$:

$$\frac{\delta(y_2)}{y_2} = \frac{\delta(y_1)}{y_1} (= \delta(x))$$

$$y_2\delta(y_1) = y_1\delta(y_2)$$

$$\frac{y_2\delta(y_1) - y_1\delta(y_2)}{y_2^2} = \delta\left(\frac{y_1}{y_2}\right) = 0 \ .$$



XII.12.2.10. Example. Let $f \in \mathbb{M}(\Omega)$ for some connected open set $\Omega \subseteq \mathbb{C}$. An exponential for f is a function g of the form ce^f for some nonzero $c \in \mathbb{C}$. The function g is meromorphic (holomorphic) on any subregion $\Omega' \subseteq \Omega$ containing no poles of f, i.e. a subset on which f is holomorphic.

If f is not identically zero, a logarithm for f is a function h of the form $c + \log f$ for some $c \in \mathbb{C}$. The function h can be chosen meromorphic (holomorphic) on any simply connected $\Omega' \subseteq \Omega$ not containing any zeroes or poles of f.

Thus, if $\mathbb{F}$ is any differential function field, and $f \in \mathbb{F}$, there is a dff $\mathbb{K}$ containing $\mathbb{F}$ containing an exponential for f; if f is nonzero there is also such a $\mathbb{K}$ containing a logarithm for f. The smallest such dff is well defined up to isomorphism, and is called the dff obtained from $\mathbb{F}$ by adding an exponential [logarithm] of $\mathbb{F}$; it is isomorphic to an evident quotient of the universal differential algebra $\mathbb{F}\{X\}$ (by the differential ideal generated by $X_1 - fX$ in the exponential case and $fX_1 - f'$ in the logarithm case).

XII.12.2.11. Proposition. [Logarithmic Differentiation Formula] Let $(\mathbb{F}, \delta)$ be a differential field, $a_1, \ldots, a_n$ nonzero elements of $\mathbb{F}$, and $r_1, \ldots, r_n \in \mathbb{N}$. Set $a = a_1^{r_1} a_2^{r_2} \cdots a_n^{r_n}$. Then

$$\frac{\delta(a)}{a} = r_1 \frac{\delta(a_1)}{a_1} + r_2 \frac{\delta(a_2)}{a_2} + \cdots + r_n \frac{\delta(a_n)}{a_n} \ .$$

Proof: This is an easy proof by induction on $r_1 + \cdots + r_n$, using the product rule. 👍

XII.12.3. Integration in Finite Terms

The name of this section has been the title of at least two important publications on the subject ([Rit48], [Ros72]). The general problem is: if δ is a derivation on a ring R, what is the range of δ, and in particular is δ surjective? If $x \in R$, can δ be extended to a well-behaved larger ring with x in the range of the extended derivation?

Rather than treating the general problem, we will have the much more modest goal of examining the question for differential function fields. In other words, we will determine conditions for when there is, or is not, a differentiable function g with an elementary formula whose derivative is a given function f which has an elementary formula, or equivalently whether

$$\int f(x)\, dx$$

has a formula in terms of standard functions.

We must first carefully determine what it means for a function to have an "elementary formula."

Elementary Formulas

XII.12.3.1. We will informally say a function has an elementary formula if it can be written in terms of algebraic operations and exponential, logarithm, trigonometric, and inverse trigonometric functions, which are defined on some common interval. (We could include hyperbolic and inverse hyperbolic functions too, but these functions can be written in terms of exponential and logarithm functions so they come for free.)

XII.12.3.2. It is convenient and useful to pass to complex functions. Each of the "building block" functions (algebraic, exponential, logarithm, trigonometric, inverse trigonometric) can be regarded as a holomorphic function on large regions of the complex plane, and hence any real-valued function defined on a real interval which has an elementary formula in the above sense will be holomorphic on a large region in $\mathbb{C}$. In addition, when we expand to complex functions, we get the pleasant fact that trigonometric and inverse trigonometric functions can also be written in terms of exponential and logarithm functions () and thus do not have to be separately considered. We can thus reduce the problem to consideration of only functions which can be written with formulas involving only algebraic operations and exponential and logarithm functions.

XII.12.3.3. Definition. Let $(\mathbb{F}, \delta)$ be a differential field. A differential field $\mathbb{K}$ containing $\mathbb{F}$ is an *elementary extension* of $\mathbb{F}$ if there is a finite chain

$$\mathbb{F} = \mathbb{F}_1 \subseteq \mathbb{F}_2 \subseteq \cdots \subseteq \mathbb{F}_n = \mathbb{K}$$

of field extensions where, for each k, $\mathbb{F}_{k+1}$ is either a finite extension of $\mathbb{F}_k$ or is obtained by adding to $\mathbb{F}_k$ an exponential or logarithm of an element of $\mathbb{F}_k$.

An *elementary function field* is an elementary extension of the field $\mathbb{C}(X)$ of rational functions over $\mathbb{C}$. A function f which is meromorphic on a region in $\mathbb{C}$ is an *elementary function* if it lies in some elementary function field.

XII.12.3.4. By XII.12.2.6.and XII.12.2.10., an elementary function field is a differential function field. Any elementary extension of an elementary extension of a differential field $\mathbb{F}$ is an elementary extension of $\mathbb{F}$; in particular, an elementary extension of an elementary function field is an elementary function field.

XII.12.3.5. Example. A function with a formula like

$$f(z) = \frac{\sqrt[3]{e^{\cos(z^2+1)} - \log(\sin^{-1} z))}}{z^5 - iz + 1}$$

is an elementary function. The reader can find a finite chain of exponentials, logarithms, and roots (beginning with $e^{i(z^2+1)}$) which when successively added to the field of rational functions include this function.

The problem of integration in finite terms thus becomes the following specific problem:

XII.12.3.6. Question. Let f be an elementary function. Is there an elementary function g with $g' = f$?

XII.12.3.7. It would be clean to take the field of all elementary functions, with its natural derivative (derivation), and ask what the range of this derivation is. However, it is not clear that this version of the question is well defined. (There is certainly no region in $\mathbb{C}$ on which all elementary functions are meromorphic.) Thus we have to work locally in elementary function fields.

A more general version of the question is thus:

XII.12.3.8. Question. Let $(\mathbb{F}, \delta)$ be a differential field, and $x \in \mathbb{F}$. Under what conditions is there an elementary extension $(\mathbb{K}, \tilde{\delta})$ of $\mathbb{F}$ and an element $y \in \mathbb{K}$ with $x = \tilde{\delta}(y)$?

XII.12.3.9. In 1835, Liouville gave a necessary and sufficient condition (XII.12.3.11.) for elements of elementary function fields, which was abstracted by Ostrowski [**?**] in 1946 (cf. [Ros72]):

XII.12.3.10. THEOREM. Let $(\mathbb{F}, \delta)$ be a differential field, and $x \in \mathbb{F}$. Then there is an elementary extension $(\mathbb{K}, \tilde{\delta})$ of $(\mathbb{F}, \delta)$ with the same field of constants as $(\mathbb{F}, \delta)$, and an element $y \in \mathbb{K}$ with $x = \tilde{\delta}(y)$, if and only if there are constants $c_1, \ldots, c_n$ and elements $u_1, \ldots, u_n, v$ of $\mathbb{F}$ such that

$$x = \sum_{k=1}^{n} c_k \frac{\delta(u_k)}{u_k} + \delta(v) \ .$$

XII.12.3.11. COROLLARY. Let $(\mathbb{F}, \delta)$ be an elementary function field, and $f \in \mathbb{F}$. Then there is an elementary function g with $f = g'$ if and only if there are constants $c_1, \ldots, c_n$ and functions $h_1, \ldots, h_n, \phi$ in $\mathbb{F}$ such that

$$f = \sum_{k=1}^{n} c_k \frac{h_k'}{h_k} + \phi' \ .$$

XII.12.3.12. To obtain the corollary from the theorem, it suffices to note that all differential function fields have the same field of constants, namely the constant functions. One direction of the theorem is very easy: if x is of the given form, one only needs to add logarithms $z_1, \ldots, z_n$ of $u_1, \ldots, u_n$ to $\mathbb{F}$, and then

$$y = c_1 z_1 + \cdots + c_n z_n + v$$

is an antiderivative of x. Thus the conclusion of the theorem can be strengthened, since $\mathbb{K}$ is obtained by just adding logarithms.

The other direction is a more involved argument requiring some elementary field theory, and is omitted. See e.g. [Ros72]. The requirement that $\mathbb{K}$ have the same constants as $\mathbb{F}$ is necessary: if $\mathbb{F} = \mathbb{R}(X)$, and $f(x) = \frac{1}{x^2+1}$, then f is not of the given form in $\mathbb{F}$ even though it has an elementary antiderivative in an extension of $\mathbb{C}(X)$ (which is a finite, hence elementary, extension of $\mathbb{F}$) by a logarithm. (It does have such an expression in $\mathbb{C}(X)$, to which the theorem applies.)

Applications

We now discuss some simple functions which have no elementary antiderivatives.

XII.12.3.13. We first examine integrals of the form

$$\int \frac{1}{\sqrt{p(x)}}\, dx$$

where p is a polynomial with complex coefficients of degree ≥ 3 and with no multiple roots. Integrals of this type include *elliptic integrals*, which arise in computing the arc length of an ellipse. Of course, what we want is an elementary function g whose derivative is $f(z) = \frac{1}{\sqrt{p(z)}}$ (say, on some subregion of $\mathbb{C}$; it could be the preimage under $1/p$ of any simply connected region not containing 0).

XII.12.3.14. We now examine integrals of the form

$$\int f(x)e^{g(x)}\,dx$$

where f and g are rational functions. LIOUVILLE showed:

XII.12.3.15. THEOREM. Let f and g be in $\mathbb{C}(X)$, with $f \neq 0$ and g not constant. Then there is an elementary function h such that $h'(z) = f(z)e^{g(z)}$ if and only if there is a rational function r such that

$$r'(X) + g'(X)r(X) = f(X)$$

in $\mathbb{C}(X)$. If there is such an r, then $h(z) = r(z)e^{g(z)}$ is an elementary antiderivative of fe^g.

XII.12.3.16. Note that the condition is considerably stronger than simply asserting that the differential equation

$$y' + g'(t)y = f(t)$$

has a solution, which will always be the case from existence theorems for differential equations (at least if f and g have real coefficients); the condition requires that it have a *rational* solution. A corollary (equivalent version) of the theorem is that if the differential equation does not have a rational solution, it does not have an elementary solution.

XII.12.3.17. For example, consider the integral

$$\int \frac{e^x}{x}\,dx\ .$$

Applying the theorem with $f(X) = \frac{1}{X}$ and $g(X) = X$, this integral has an elementary formula if and only if the differential equation

$$r'(X) + r(X) = \frac{1}{X}$$

has a rational solution. Suppose $r(X) = \frac{p(X)}{q(X)}$ is a solution with $p(X)$, $q(X)$ polynomials which are relatively prime. We have

$$\frac{q(X)p'(X) - p(X)q'(X)}{[q(X)]^2} + \frac{p(X)}{q(X)} = \frac{q(X)p'(X) - p(X)q'(X) + p(X)q(X)}{[q(X)]^2} = \frac{1}{X}$$

in $\mathbb{C}(X)$, and hence X divides $[q(X)]^2$ and therefore $q(X)$. Thus

$$q(X) = X^n\phi(X)$$

for some $n \geq 1$ and some polynomial $\phi(X)$ not divisible by X. Note that X does not divide $p(X)$. We have

$$\frac{X^n\phi(X)p'(X) - p(X)[nX^{n-1}\phi(X) + X^n\phi'(X)]}{X^{2n}[\phi(X)]^2} + \frac{p(X)}{X^n\phi(X)}$$

$$= \frac{X^n[\phi(X)p'(X) - \phi'(X)p(X) + \phi(X)p(X)] - nX^{n-1}\phi(X)p(X)}{X^{2n}[\phi(X)]^2}$$
$$= \frac{X[\phi(X)p'(X) - \phi'(X)p(X) + \phi(X)p(X)] - n\phi(X)p(X)}{X^{n+1}[\phi(X)]^2} = \frac{1}{X}$$

in $\mathbb{C}(X)$. The denominator is divisible by X^{n+1}, so the numerator must be divisible by X^n, and in particular by X. Thus $n\phi(X)p(X)$ must be divisible by X. But X does not divide $p(X)$, so it must divide $\phi(X)$, a contradiction. Thus there is no rational solution, and no elementary formula for the antiderivative.

The substitution $x = \log u$ converts this integral into

$$\int \frac{1}{\log u}\, du$$

which therefore also does not have an elementary formula. This integral arises in the Prime Number Theorem ().

XII.12.3.18. Now we turn to the best-known example of this type:

$$\int e^{-x^2}\, dx$$

which is important in probability (). The arguments for

$$\int e^{x^2}\, dx$$

are essentially identical.

Applying the theorem with $f(X) = 1$ and $g(X) = -X^2$, we need to show that there is no rational function $r(X)$ satisfying

$$r'(X) - 2Xr(X) = 1\ .$$

Clearly there is no polynomial satisfying this equation, since if $r(X)$ is a polynomial of degree d, the left side is a polynomial of degree $d+1$. Thus we may write $r(X) = \frac{p(X)}{q(X)}$ in lowest terms with q not constant. Then q has a root $z_0 \in \mathbb{C}$, i.e. r has a pole of some finite order $n \geq 1$ at z_0, so r can be written

$$r(X) = \frac{h(X)}{(X - z_0)^n}$$

with $h(X) \in \mathbb{C}(X)$ defined and nonzero at z_0, and h and h' are continuous at z_0. Set

$$\phi(X) = r'(X) - 2Xr(X) = \frac{(X - z_0)^n h'(X) - n(X - z_0)^{n-1}h(X)}{(X - z_0)^{2n}} - \frac{2Xh(X)}{(X - z_0)^n}$$
$$= \frac{(X - z_0)^n[h'(X) - 2Xh(X)] - n(X - z_0)^{n-1}h(X)}{(X - z_0)^{2n}} = \frac{(X - z_0)[h'(X) - 2Xh(X)] - nh(X)}{(X - z_0)^{n+1}}$$

and thus, if ϕ is regarded as a meromorphic function on $\mathbb{C}$, we have

$$\lim_{z \to z_0} (z - z_0)^{n+1}\phi(z) = \lim_{z \to z_0} [(z - z_0)[h'(z) - 2zh(z)] - nh(z)] = nh(z_0) \neq 0$$

which contradicts that $\phi(X) = 1$ in $\mathbb{C}(X)$.

Thus there is no rational solution to the differential equation, and no elementary formula for the antiderivative.

XII.12.4. Exercises

XII.12.4.1. Let δ be the derivation of $\mathbb{F}[X]$ defined as in XII.12.1.3.(iii). Show that the automorphism e^δ (XII.12.1.3.(vi)) is defined, and that $[e^\delta(f)](X) = f(X+1)$. Show why this does not work on $\mathbb{F}[[X]]$. But it does work on $\mathbb{F}\{X\}$ (XII.12.1.3.(iv)); give an explicit formula.

XII.12.4.2. Let (R,δ) and $(\tilde{R},\tilde{\delta})$ be differential rings. A ring-homomorphism $\phi : R \to \tilde{R}$ is a *differential homomorphism*, or *diffomorphism*, if ϕ intertwines δ and $\tilde{\delta}$, i.e. $\tilde{\delta}(\phi(x)) = \phi(\delta(x))$ for all $x \in R$. [The term "diffomorphism" is to "diffeomorphism" exactly as "homomorphism" is to "homeomorphism".]

Show that if (R,δ) is a commutative differential ring and $z \in R$, there is a unique diffomorphism from the differential polynomial ring $R_0\{X\}$ (XII.12.1.3.(iv)) to R sending X to z. Thus $R_0\{X\}$ is the "universal" differential algebra with one generator over the ring R_0 of constants.

XII.12.4.3. Generalize the construction of the differential polynomial ring as follows. Let $(\mathbb{F},\delta)$ be a commutative differential ring, and let $X, X_1, X_2, \dots$ be a sequence of indeterminates. Form the polynomial ring

$$R = \mathbb{F}[X, X_1, X_2, \dots] \ .$$

Define a map $\tilde{\delta} : R \to R$ by

$$\delta\left(\sum_{k,n} \alpha_{kn} X_k^n\right) = \sum_{k,n} \delta(\alpha_{k,n}) X_k^n + \sum_{k,n} n\alpha_{k,n} X_k^{n-1} X_{k+1} \ .$$

(i) Show that $\tilde{\delta}$ is a derivation on R which extends the derivation δ on the subring $\mathbb{F}$ of "constant polynomials" (which is not the ring of constants for $\tilde{\delta}$ in general!) Denote the ring R by $\mathbb{F}\{X\}$. Note that if δ is the zero derivation on $\mathbb{F}$, then this definition agrees with the previous definition of $\mathbb{F}\{X\}$.

(ii) Show that if $(\mathbb{K},\bar{\delta})$ is another commutative differential ring and $\phi : \mathbb{F} \to \mathbb{K}$ is a diffomorphism, and $z \in \mathbb{K}$, then there is a unique extension $\tilde{\phi}$ of ϕ to a diffomorphism from $\mathbb{F}\{X\}$ to $\mathbb{K}$ sending X to z. Thus $\mathbb{F}\{X\}$ is the "universal" differential algebra in one variable over the differential ring $(\mathbb{F},\delta)$.

XII.12.4.4. Let $\mathbb{F}$ be a unital commutative ring, and $\mathbb{F}\{X,Y\}$ the universal differential algebra in two variables over $\mathbb{F}$ (XII.12.1.3.(iv)). Give a careful description of $\mathbb{F}\{X,Y\}$, δ_X, and δ_Y, and prove the algebraic version of Clairaut's Theorem: δ_X and δ_Y commute, i.e. $\delta_X \circ \delta_Y = \delta_Y \circ \delta_X$. In fact, show that $\mathbb{F}\{X,Y\}$, $[\mathbb{F}\{X\}]\{Y\}$, and $[\mathbb{F}\{Y\}]\{X\}$ are naturally isomorphic, where the last two are defined as in XII.12.4.3.. Show that the automorphisms e^{δ_X} and e^{δ_Y} of $\mathbb{F}\{X,Y\}$ exist, give formulas, and show they commute.

XII.12.4.5. Let (R,δ) be a differential ring, and J a closed two-sided ideal of R (). If J is invariant under δ, i.e. if $\delta(J) \subseteq J$, show that δ drops to a derivation $\bar{\delta}$ on the quotient ring R/J satisfying $\bar{\delta}(\pi(x)) = \pi(\delta(x))$ for all $x \in R$, where $\pi : R \to R/J$ is the quotient map (i.e. π is a diffomorphism). An ideal of R invariant under δ is called a *differential ideal*.

XII.12.4.6. Let $(\mathbb{K},\delta)$ be a differential field, and $\mathbb{F}$ a subfield of $\mathbb{K}$ closed under δ. If $y \in \mathbb{K} \setminus \mathbb{F}$ is an exponential or logarithm of an element of $\mathbb{F}$, is y necessarily transcendental over $\mathbb{F}$? What if $\mathbb{F}$ contains all constants of $\mathbb{K}$?

XII.12.4.7. Define a one-parameter automorphism group of $\mathcal{C}_c^\infty(\mathbb{R})$ () by

$$[\alpha_t(f)](x) = f(x+t) \ .$$

Show that α_t is an automorphism for each t, that $\alpha_{t+s} = \alpha_t \circ \alpha_s$ for all t, s. Discuss continuity of this one-parameter automorphism group and how the "derivative" in the sense of XII.12.1.3.(vi) is the derivation given by ordinary differentiation.

XII.12.4.8. (a) Find an elementary closed formula for

$$\int \sqrt{\tan x}\ dx$$

(a computer algebra system is very helpful).

(b) (Liouville, 1835) Show that

$$\int \sqrt{\sin x}\ dx$$

has no elementary closed formula.

Chapter XIII

Topology

Topology is a part of mathematics which provides a general setting for the notions of limits and continuity.

Topology is a huge subject, and goes far beyond the parts which form an essential tool for analysis and other parts of mathematics. Our treatment of topology will mostly just cover the basic tools of the subject, with an eye to applications in analysis.

In his classic topology book [Kel75], J. Kelley described the contents as "what every young analyst should know." While his judgment was of course subjective, it can be taken as a good guide. The topology covered in this book largely represents my own judgment of What Every Analyst Should Know, but I admit some idiosyncracies due to my own personal interests.

The idea behind topology is to describe what it means for points of a set to be "close together." There are at least four standard and commonly used equivalent ways to define, specify, or think of a topology on a set X (usually called a *space*):

(1) Specifying what the neighborhoods of each point of X are, i.e. specifying which subsets of X are open.

(2) Specifying the closure of each subset of X.

(3) Specifying which sequences (or generalized sequences) in X converge, and what their limits are.

(4) Specifying which functions into or out of X are continuous.

A good way of doing (1) is to specify a metric on X, and hence open balls in X, but there are other ways to do (1) also.

Approach (1) is normally taken as the definition of a topology. We will in due course describe all four approaches and how to go from each to the others.

Topology as used by topologists (or as presented in most elementary topology classes) has a rather different emphasis than topology as used by analysts. Topologists by and large consider spaces which have only one obvious or natural topology (except sometimes for exotic topologies artificially constructed primarily to provide counterexamples; there are other exceptions too). When a topologist considers convergence of a sequence or continuity of a function, there is a natural interpretation, and it is not necessary to go back and worry about "with respect to what topology?" The topology tends to be taken for granted.

An analyst must quickly overcome this mindset. Spaces arising in analysis, notably function spaces, often have more than one *natural* topology, sometimes a great many, with varying properties. Convergence in one sense does not generally imply convergence in another sense, and some properties (e.g. continuity of functions) are preserved under certain kinds of limits but not others. Thus, for example, an analyst should never say of functions that $f_n \to f$ without carefully specifying what is meant, because such a statement is totally ambiguous otherwise. An analyst needs to be acutely aware of which topology is being considered in a given situation, and that there is usually a multitude of reasonable candidates.

Topology is a subject which is full of definitions; in fact, the density of definitions in topology may be the highest in all the parts of mathematics. This situation arises because of the great generality of the subject and the enormous variety of topological spaces with varied properties. Most definitions in topology are of little direct interest to analysts. We will for the most part stick to those definitions relevant to analysis to try to avoid getting too bogged down in terminology.



Figure XIII.1: Rumored original source of topology in Serbia

XIII.1. Topological Spaces

The most important tool used in the study of limits and continuity in Euclidean spaces and metric spaces is the notion of "open set." The idea behind the definition of a topological space is to abstract the notion of open set directly, without working through intermediate structures like metrics. Thus a topological space will just be an abstract set X together with a specified collection of subsets which we arbitrarily call "open" sets in X; the term "open" does not have any intrinsic or geometric meaning in this context. We only require that the "open" sets have some simple union and intersection properties which hold for open sets in Euclidean space, and which are necessary to make a reasonable theory of limits and continuity.

XIII.1.1. Definitions

XIII.1.1.1. DEFINITION. A *topology* on a set X is a subset $\mathcal{T}$ of $\mathcal{P}(X)$ with the following properties:

(i) $\emptyset, X \in \mathcal{T}$.

(ii) $\mathcal{T}$ is closed under finite intersections.

(iii) $\mathcal{T}$ is closed under arbitrary unions.

The sets in $\mathcal{T}$ are called the *open sets* (or $\mathcal{T}$-*open sets*) in X.

A *topological space* is a pair $(X, \mathcal{T})$, where X is a set and $\mathcal{T}$ is a topology on X.

There is a technical question whether a topological space is required to be nonempty (the pair $(\emptyset, \mathcal{P}(\emptyset))$ satisfies the definition). We will generally assume topological spaces are nonempty, but explicitly state this when it is important.

XIII.1.1.2. A set can have many different topologies. For example, the function spaces studied in analysis typically have several different natural topologies, and it is necessary to carefully distinguish between them. In some settings, there is one standard topology which is obvious from the context, so by abuse of language we often make statements like "let X be a topological space." However, it should be carefully noted and always kept in mind that the specification of a topology is an essential part of the definition of a topological space.

Topological spaces are the natural abstract setting for the notions of limits and continuity. We typically refer to elements of a topological space as "points." In this connection, it is convenient to think of an open set containing a point as a "neighborhood" of the point. Slightly more generally:

XIII.1.1.3. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, and $x \in X$. A *neighborhood* (or $\mathcal{T}$-*neighborhood*) of x is a set containing an open set which contains x.

Note that by this definition, a neighborhood of a point in X is not necessarily an open set. It is sometimes convenient to work with neighborhoods which are not open. But the most important neighborhoods of points are open neighborhoods; every neighborhood of a point x contains an open neighborhood of x, which is simply an open set containing x. We will try to always specify "open neighborhood" when open sets are required, even though this sometimes becomes a little pedantic.

Topologies and Metrics.

XIII.1.1.4. EXAMPLES.

(i) If X is any set, then $\mathcal{P}(X)$ is a topology on X, called the *discrete topology*.

(ii) If X is any set, then $\{\emptyset, X\}$ is a topology on X, called the *indiscrete topology*.

(iii) The most important examples of topologies are those defined by metrics. If (X, ρ) is a metric space, as usual define a subset of X to be *open* (or *ρ-open*) if it contains an open ball around each of its points. The set $\mathcal{T}_\rho$ of ρ-open sets is a topology on X, called the *topology defined by* (or *underlying topology of*) the metric space (X, ρ). Equivalent metrics define the same topology.

The discrete topology on a set X comes from a metric (the *discrete metric*) ρ, where $\rho(x, y) = 1$ if $x \neq y$.

The usual notions of convergence of sequences and closure of a set in a metric space, or of continuity of functions between metric spaces, depend only on the underlying topology and can all be phrased in terms of the topology. See () and Problem () for continuity, and () for the other notions.

XIII.1.1.5. There are, however, many examples of topological spaces which do not come from metric spaces (e.g. the indiscrete topology on a set with more than one element), some of which are important in applications. Even a metrizable topological space (one whose topology comes from a metric) need not have any *natural* metric defining the topology, and it is often unnatural to artificially pick one of the many possible metrics defining a topology which is given in another way. Thus it is useful to work directly with topological spaces rather than metric spaces in many situations.

XIII.1.1.6. Although topologies, even interesting topologies, do not come from metrics in general, most interesting topologies share some of the properties of metric spaces: there are "enough" open sets that points have "small" neighborhoods, and in particular distinct points have disjoint neighborhoods. There are varying degrees of similarity to metrizable spaces which are useful, described by the separation axioms developed in ().

XIII.1.2. Bases and Subbases

XIII.1.2.1. If $\{\mathcal{T}_\alpha\}$ is a family of topologies on a set X, then $\cap_\alpha \mathcal{T}_\alpha$ is also a topology on X (Exercise XIII.1.6.2.). In particular, if $\mathcal{S}$ is any subset of $\mathcal{P}(X)$, there is a smallest topology containing $\mathcal{S}$, called the *topology generated by* $\mathcal{S}$ (there is always at least one topology containing $\mathcal{S}$, the discrete topology).

Bases

It is usually convenient to use a smaller collection of "nice" open sets to describe a topology. For example, in a topology defined by a metric, the open balls generate the topology as in XIII.1.2.1.. We abstract this situation in the notion of base:

XIII.1.2.2. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, and $p \in X$. A subset $\mathcal{B}$ of $\mathcal{T}$ is a *local base* for $\mathcal{T}$ at p if, whenever $U \in \mathcal{T}$ and $p \in U$, there is a $V \in \mathcal{B}$ with $p \in V$ and $V \subseteq U$.

A subset $\mathcal{B}$ of $\mathcal{T}$ is a *base* for $\mathcal{T}$ if it is a local base at every point, i.e. whenever $U \in \mathcal{T}$ and $x \in U$, there is a $V \in \mathcal{B}$ with $x \in V$ and $V \subseteq U$. In other words, if $U \neq \emptyset$, $U = \cup_{\substack{V \in \mathcal{B} \\ V \subseteq U}} V$, i.e. every nonempty open set in $\mathcal{T}$ is a union of sets in $\mathcal{B}$.

In particular, if $\mathcal{B}$ is a base for $\mathcal{T}$, then $\mathcal{T}$ is the topology generated by $\mathcal{B}$.

When a base or local base $\mathcal{B}$ is specified, the sets in $\mathcal{B}$ are usually called *basic open sets*. Some authors use the term *basis* instead of *base*, in analogy with the use of "basis" for a vector space; but this is potentially misleading, since a base for a topology does not necessarily have any type of minimality condition like a basis for a vector space: for example, any collection of open sets containing a base for $\mathcal{T}$ is also a base for $\mathcal{T}$.

XIII.1.2.3. EXAMPLES. (i) If (X, ρ) is a metric space, and $p \in X$, the open balls centered at p form a local base for the topology at p. The set of all open balls in X forms a base for $\mathcal{T}_\rho$.

(ii) If X has the discrete topology, and $p \in X$, then $\{\{p\}\}$ is a local base for the topology at p. The collection of all singleton subsets of X is a base for the discrete topology.

(iii) If $\mathcal{T}$ is a topology on X, then the collection of all sets in $\mathcal{T}$, or the collection of all nonempty sets in $\mathcal{T}$, is a (not very interesting) base for $\mathcal{T}$.

XIII.1.2.4. There is nothing unique about a base for a given topology. For example, if ρ_1 and ρ_2 are equivalent metrics on a set X, the open *balls* from ρ_1 and ρ_2 may be quite different, but the open *sets* they generate will be the same. See () for an example.

It is useful to know when a collection $\mathcal{B}$ of open subsets of X forms a base for the topology it generates.

XIII.1.2.5. PROPOSITION. Let X be a set and $\mathcal{B} \subseteq \mathcal{P}(X)$, and let $\mathcal{T}$ be the topology generated by $\mathcal{B}$. Then $\mathcal{B}$ is a base for $\mathcal{T}$ if and only if

(i) $\cup_{V \in \mathcal{B}} V = X$ and

(ii) If $V_1, \ldots, V_n \in \mathcal{B}$ and $x \in \cap_{j=1}^n V_j$, there is a $V \in \mathcal{B}$ with $x \in V$ and $V \subseteq \cap_{j=1}^n V_j$.

XIII.1.2.6. Note that a base for a topology need not be closed under finite intersections. For example, an intersection of open balls with different centers in a metric space is generally not an open ball. (For XIII.1.2.5.(ii), it obviously *suffices* that $\mathcal{B}$ be closed under finite intersections, as is frequently the case.)

First and Second Countability

Most interesting topological spaces have uncountably many open sets. But frequently there is a countable base for the topology, or at least a countable local base at each point, and if so the topology is more easily describable and tends to behave more nicely.

XIII.1.2.7. DEFINITION. Let $(X, \mathcal{T})$ be a topological space.

(i) $(X, \mathcal{T})$ satisfies the *first axiom of countability*, or is *first countable*, if there is a countable local base for $\mathcal{T}$ at each point.

(ii) $(X, \mathcal{T})$ satisfies the *second axiom of countability*, or is *second countable*, if there is a countable base for $\mathcal{T}$.

The "first" and "second" are used for historical reasons, but the terminology persists because it is both appropriate and useful. A second countable space is first countable.

XIII.1.2.8. EXAMPLES.

(i) The discrete topology on a set X is always first countable; it is second countable if and only if X is countable.

(ii) The indiscrete topology on a set X is always second countable.

(iii) The usual topology on $\mathbb{R}^n$ is second countable: balls with rational radii around points with rational coordinates form a base for the topology. More generally, the topology of a metric space is always first countable (for any point p, balls of rational radius centered at p form a local base at p), and is second countable if and only if the metric space is separable (XIII.2.3.15.).

(iv) Let X be a set. The *finite complement topology* on X is the topology consisting of $\emptyset$ and all subsets of X with finite complement. (It is easily checked that this is a topology; cf. Exercise XIII.1.6.4..) If X is countable, then this topology is second countable: in fact, there are only countably many open sets in all. But if X is uncountable, then the finite complement topology is not first countable, and in fact no point has a countable neighborhood base. For if $p \in X$ and $\{U_n : n \in \mathbb{N}\}$ is a countable collection of open neighborhoods of p, set $E = \cap_{n=1}^{\infty} U_n$. Then E has countable complement and is therefore uncountable, and in particular contains a point $q \neq p$. Then $X \setminus \{q\}$ is an open neighborhood of p not containing any U_n, so the U_n cannot form a local base at p. (This argument requires the Countable AC.)

Topological spaces which are not first countable are rather unusual, but they do arise naturally in analysis; see (). First countability is important since it is essentially just in first countable topological spaces that the topology can be entirely described by convergence of sequences (although see XIII.3.5.2. and XXII.2.1.21.); in spaces which are not first countable it is usually necessary to work with generalizations of sequences called *nets*.

XIII.1.2.9. **Caution:** For a topological space to be first countable, it is not sufficient that every singleton subset be a countable intersection of open sets (which may be assumed to be decreasing since a finite intersection of open sets is open), i.e. that every singleton subset is a G_δ: there are countable (Hausdorff, even normal) topological spaces which are not first countable (Exercise XIII.11.14.10.). (The condition that singletons are G_δ's is not necessary either: the indiscrete topology on any set is even second countable. In a first countable T_1 space (), every singleton subset is a G_δ (XIII.7.3.3.). In fact, see XIII.7.10.2..)

XIII.1.2.10. PROPOSITION. Let $(X, \mathcal{T})$ be a second countable topological space, and $\mathcal{B}$ be a base for $\mathcal{T}$. Then $\mathcal{B}$ contains a countable base for $\mathcal{T}$. In particular, every $U \in \mathcal{T}$ is a *countable* union of sets in $\mathcal{B}$.

PROOF: Let $\mathcal{D} = \{V_n : n \in \mathbb{N}\}$ be a countable base for $\mathcal{T}$. Let S be the set of $(n, m) \in \mathbb{N}^2$ such that there is a $W_{n,m} \in \mathcal{B}$ with $V_m \subseteq W_{n,m} \subseteq V_n$. Let $U \in \mathcal{T}$. For each $x \in U$, there is an n such that $x \in V_n$ and $V_n \subseteq U$ since $\mathcal{D}$ is a base for $\mathcal{T}$. Since $\mathcal{B}$ is a base, there is a $W \in \mathcal{B}$ with $x \in W$ and $W \subseteq V_n$; then, since $\mathcal{D}$ is a base, there is an m with $x \in V_m$ and $V_m \subseteq W$. Thus $(n, m) \in S$ and $x \in W_{n,m} \subseteq U$. So $\{W_{n,m} : (n, m) \in S\}$ is a countable subset of $\mathcal{B}$ which is a base for $\mathcal{T}$. (This argument uses the Countable AC.) 👍

XIII.1.2.11. We can more generally define the *weight* $w(X)$ of a topological space X to be the minimum cardinality of a base for the topology of X (this is only well defined in general if the AC is assumed). Thus X is second countable if and only if $w(X) \leq \aleph_0$. Weight is an example of a *cardinal invariant* for topological spaces. The simplest cardinal invariant for X is $card(X)$; another cardinal invariant is *density* $d(X)$ (XIII.2.3.16.). There are many other cardinal invariants, such as:

(i) *character* $\chi(X)$, minimum cardinal κ such that every point of X has a local base of cardinality $\leq \kappa$, e.g. X is first countable if and only if $\chi(X) \leq \aleph_0$.

(ii) *Lindelöf number* $L(X)$, the smallest cardinal κ such that every open cover of X has a subcover (XIII.1.3.3.) of cardinality $\leq \kappa$, e.g. X is Lindelöf (XIII.1.3.4.) if and only if $L(X) \leq \aleph_0$.

(iii) *cellularity* $c(X)$, minimum cardinal κ such that every family of pairwise disjoint nonempty open subsets of X has cardinality $\leq \kappa$. X has the *Countable Chain Condition* if $c(X) \leq \aleph_0$.

The invariants $w(X)$, $d(X)$, $L(X)$, and $c(X)$ coincide for metrizable spaces, but differ in general. A thorough discussion of cardinal invariants can be found in [KV84], and there is also extensive discussion in [Eng89].

Subbases

It is sometimes convenient to work with a less restrictive set of generators:

XIII.1.2.12. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, and $\mathcal{S} \subseteq \mathcal{T}$. Then $\mathcal{S}$ is a *subbase* for $\mathcal{T}$ if the collection of finite intersections of sets in $\mathcal{S}$ forms a base for $\mathcal{T}$.

XIII.1.2.13. EXAMPLE. In $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, let $\mathcal{S}$ be the collection of all sets of the form $I \times \mathbb{R}$ or $\mathbb{R} \times I$, where I is an open interval in $\mathbb{R}$. Then $\mathcal{S}$ is a subbase for the usual topology on $\mathbb{R}^2$.

More generally, the topology generated by a family of pseudometrics is naturally defined by a subbase. See ().

XIII.1.2.14. It is sometimes convenient to extend the notion of base to allow sets which are not open. We say that a collection $\mathcal{B}$ of not necessarily open subsets of X is a base for the topology $\mathcal{T}$ if, whenever $U \in \mathcal{T}$ and $x \in U$, there is an $E \in \mathcal{B}$ which is a neighborhood of x for which $E \subseteq U$. For example, the set of closed balls of positive radius in a metric space is a base for the topology in this extended sense. We can similarly extend the notion of local base and subbase. Thus, for example, we can in some cases (e.g. in $\mathbb{R}^n$) take a base for a topology consisting of compact sets ().

If $\mathcal{B}$ is a base for the topology $\mathcal{T}$ in this sense, then the interiors (XIII.2.3.17.) of the sets in $\mathcal{B}$ form a base for the topology consisting of open sets. Thus if X has a countable base in this sense, it is second countable. The same applies to local bases; in particular, if every point of X has a countable local base in this sense, then X is first countable.

XIII.1.3. Covers and Subcovers

It is often convenient to cover a topological space with "small" subsets, particularly small open sets.

XIII.1.3.1. Definition. Let X be a set, and $\mathcal{U} = \{U_i : i \in I\}$ a collection of subsets of X. If $Y \subseteq X$, then $\mathcal{U}$ *covers* Y, or is a *cover of* Y, if $Y \subseteq \cup_{i\in I} U_i$. $\mathcal{U}$ covers X if $\cup_{i\in I} U_i = X$.
If X is a topological space, each U_i is open in X, and $\mathcal{U}$ covers X, then $\mathcal{U}$ is an *open cover* of X.

XIII.1.3.2. Examples. (i) Any base or subbase for the topology of X is an open cover of X.
(ii) If (X, ρ) is a metric space and $\epsilon > 0$ is fixed, then the open balls of radius ϵ form an open cover of X.

XIII.1.3.3. Definition. Let X be a set, $Y \subseteq X$, and $\mathcal{U}$ a cover of Y. A *subcover* of $\mathcal{U}$ is a subcollection of $\mathcal{U}$ which also covers Y.

Any subcover of an open cover is also an open cover. Although the terminology is imprecise, one must specify the set Y to discuss subcovers; normally one uses "subcover" only for covers of the whole set X.

XIII.1.3.4. Definition. Let X be a topological space.
(i) X is a *Lindelöf space* if every open cover of X has a countable subcover.
(ii) X is *compact* if every open cover of X has a finite subcover.

Compact spaces are discussed in detail in (). Obviously every compact space is Lindelöf; the converse is far from true.

XIII.1.3.5. Proposition. Every second countable topological space is Lindelöf.

Proof: Let $\{V_n : n \in \mathbb{N}\}$ be a countable base for the topology of X. Let $\mathcal{U}$ be an open cover of X. Fix a $U_0 \in \mathcal{U}$. For each $n \in \mathbb{N}$, if V_n is contained in some $U \in \mathcal{U}$, choose $U_n \in \mathcal{U}$ with $V_n \subseteq U_n$; if $V_n \subseteq U$ for no $U \in \mathcal{U}$, set $U_n = U_0$. We claim $\{U_n : n \in \mathbb{N}\}$ is a subcover of $\mathcal{U}$ (it is obviously countable). If $p \in X$, then there is a $U \in \mathcal{U}$ with $p \in U$ since $\mathcal{U}$ is a cover of X. Then there is an n such that $p \in V_n$ and $V_n \subseteq U$ since $\{V_n\}$ is a base for the topology. Then $p \in U_n$. Thus $\cup_{n\in\mathbb{N}} U_n = X$. 👍

Note that this proof requires the Countable AC.

XIII.1.3.6. In particular, $\mathbb{R}^n$ is Lindelöf for any n by XIII.1.2.8.(iii).

We can sharpen this result:

XIII.1.3.7. Definition. A topological space X is *locally second countable* if every $x \in X$ has an open neighborhood which is second countable in the relative topology.

Since every subspace of a second countable space is second countable, we may eliminate "open" from the definition without changing its meaning.

XIII.1.3.8. Proposition. Let X be a topological space. Then X is second countable if and only if it is Lindelöf and locally second countable.

Proof: If X is second countable, then it is obviously locally second countable, and Lindelöf by XIII.1.3.5.. Conversely, for each $x \in X$ let U_x be an open neighborhood of x which is second countable. Let $\{U_n : n \in \mathbb{N}\}$ be a countable subcover of $\{U_x : x \in X\}$. For each n, let $\mathcal{V}_n = \{V_{n1}, V_{n2}, \dots\}$ be a countable base for the relative topology on U_n. Then $\cup_n \mathcal{V}_n$ is a countable base for the topology on X. (Note that each V_{nk} is open in U_n, and U_n is open in X, so V_{nk} is open in X.) 👍

This proof uses the Countable AC.

XIII.1.4. Relative Topology on a Subspace

If $(X, \mathcal{T})$ is a topological space and Y is a subset of X, then there is one, and only one, reasonable way to put a topology on Y which is compatible with the topology on X:

XIII.1.4.1. Definition. Let $(X, \mathcal{T})$ and $Y \subseteq X$. The *relative topology*, or *subspace topology*, on Y from $\mathcal{T}$ is

$$\mathcal{T}_Y = \{U \cap Y : U \in \mathcal{T}\}$$

i.e. the open sets in Y are the intersections with Y of the open sets in X. Y with this topology is called a *subspace* of X.

XIII.1.4.2. Proposition. $\mathcal{T}_Y$ is a topology on Y.

Here is the proof as it appears in many standard references:
Proof: We have $\emptyset = \emptyset \cap Y$ and $Y = X \cap Y$, so $\emptyset, Y \in \mathcal{T}_Y$. If $V_1, \dots, V_n \in \mathcal{T}_Y$, then, for each k, $V_k = U_k \cap Y$ for some $U_k \in \mathcal{T}$; so

$$V_1 \cap \cdots \cap V_n = (U_1 \cap Y) \cap \cdots \cap (U_n \cap Y) = (U_1 \cap \cdots \cap U_n) \cap Y \in \mathcal{T}_Y$$

since $U_1 \cap \cdots \cap U_n \in \mathcal{T}$. If $\{V_i : i \in I\}$ is a collection of sets in $\mathcal{T}_Y$, then, for each i, $V_i = U_i \cap Y$ for some $U_i \in \mathcal{T}$; then

$$\bigcup_{i \in I} V_i = \bigcup_{i \in I} (U_i \cap Y) = \left(\bigcup_{i \in I} U_i \right) \cap Y \in \mathcal{T}_Y$$

since $\cup_{i \in I} U_i \in \mathcal{T}$. 👍

XIII.1.4.3. Note that the AC is used in this proof to choose the U_i. (Most standard references seem to overlook this point! See XIII.1.6.5..) But the result does not depend on the AC, because there is actually a systematic way to choose the U_i: for each i let U_i be the union of all $W \in \mathcal{T}$ such that $V_i = W \cap Y$ (I am indebted to S. Jabuka for this argument). There is also an alternate approach which does not use the AC (XIII.1.4.7.).

XIII.1.4.4. If $A \subseteq Y$, the phrase "A is open" is ambiguous since also $A \subseteq X$: does it mean that A is open in Y (i.e. $A \in \mathcal{T}_Y$) or that A is open in X (i.e. that $A \in \mathcal{T}$)? To avoid ambiguity, we use the term *relatively open* for the first: $A \subseteq Y$ is relatively open in Y if $A \in \mathcal{T}_Y$, i.e. $A = U \cap Y$ for some $U \in \mathcal{T}$. If Y is open in X, i.e. $Y \in \mathcal{T}$, there is no ambiguity: a subset of Y is relatively open if and only if it is open in X (the converse holds also).

XIII.1.4.5. Since $(X \setminus A) \cap Y = Y \setminus (A \cap Y)$ for any $A \subseteq X$, the closed sets in $(Y, \mathcal{T}_Y)$ are precisely the sets of the form $F \cap Y$ for F closed in X. These sets are called the *relatively closed* subsets of Y. If Y is closed in X, then a subset of Y is relatively closed if and only if it is closed in X.

If $A \subseteq Y$, the phrase "closure of A" is also ambiguous: we could mean the closure $\tilde{A}$ of A in $(Y, \mathcal{T}_Y)$ or the closure $\bar{A}$ in $(X, \mathcal{T})$. But the situation is well behaved:

XIII.1.4.6. Proposition. We have $\tilde{A} = \bar{A} \cap Y$ for any $A \subseteq Y$.

Proof: Let $y \in \tilde{A}$. Then for any $V \in \mathcal{T}_Y$ with $y \in V$ there is a $z \in A \cap V$. If $U \in \mathcal{T}$ with $y \in U$, set $V = U \cap Y$. If $z \in A \cap V$, then $z \in A \cap U$, so $y \in \bar{A}$. Conversely, if $y \in \bar{A} \cap Y$ and $V \in \mathcal{T}_Y$ with $y \in V$, then there is a $U \in \mathcal{T}$ with $V = U \cap Y$. Since $y \in \bar{A}$, there is a $z \in A \cap U$. Since $A \subseteq Y$, $z \in Y \cap U = V$. Thus $y \in \tilde{A}$.

The set $\tilde{A}$ is called the *relative closure* of A in Y.

In light of this result, we can give an alternate construction of the relative topology:

XIII.1.4.7. Proposition. Let $(X, \mathcal{T})$ be a topological space, and $Y \subseteq X$. For $A \subseteq Y$, define $\tilde{A} = \bar{A} \cap Y$ ($\bar{A}$ denotes the closure of A in X). Then $A \mapsto \tilde{A}$ is a Kuratowski closure operator (XIII.2.4.1.) on Y, and the closed sets in the corresponding topology are precisely the sets of the form $F \cap Y$ for F closed in X, i.e. is the relative topology on Y.

The proof is very simple and is left to the reader. It does not depend on the AC, and hence gives an alternate proof of XIII.1.4.2. not using the AC.

The relative topology also has other properties which make it the only reasonable choice:

XIII.1.4.8. Proposition. Let $(X, \mathcal{T})$ be a topological space, $Y \subseteq X$, and $\mathcal{T}_Y$ the relative topology on Y. Then

(i) If (y_i) is a net in Y and $y \in Y$, then $y_i \to y$ in $(Y, \mathcal{T}_Y)$ if and only if $y_i \to y$ in $(X, \mathcal{T})$.

(ii) If Z is another topological space and $f : X \to Z$ is continuous, then $f|_Y : Y \to Z$ is also continuous.

(iii) The topology $\mathcal{T}_Y$ is the weakest topology on Y making the inclusion map into X continuous.

(iv) If $(Z, \mathcal{S})$ is a topological space and $f : Z \to Y$ is a function, then f is continuous as a function from $(Z, \mathcal{S})$ to $(Y, \mathcal{T}_Y)$ if and only if it is continuous as a function from $(Z, \mathcal{S})$ to $(X, \mathcal{T})$.

XIII.1.4.9. Note that (i) does *not* say that if a net in Y converges in X, then it converges in Y. This will only be true if Y is closed in X.

XIII.1.4.10. Proposition. Let $(X, \mathcal{T})$ be a topological space, and $Z \subseteq Y \subseteq X$. Then the relative topology on Z as a subspace of $(Y, \mathcal{T}_Y)$ is the same as the relative topology as a subspace of $(X, \mathcal{T})$, i.e. $(\mathcal{T}_Y)_Z = \mathcal{T}_Z$.

Bases

XIII.1.4.11. If $(X, \mathcal{T})$ is a topological space and $(Y, \mathcal{T}_Y)$ is a subspace, and if $\mathcal{B}$ is a base for $\mathcal{T}$, then it is easily checked that $\mathcal{B}_Y = \{B \cap Y : B \in \mathcal{B}\}$ is a base for $\mathcal{T}_Y$. Similarly, if $y \in Y$ and $\mathcal{B}$ is a local base for $\mathcal{T}$ at y, then $\mathcal{B}_Y$ is a local base for $\mathcal{T}_Y$ at y. We thus obtain:

XIII.1.4.12. Proposition. Let $(X, \mathcal{T})$ be a topological space and $(Y, \mathcal{T}_Y)$ a subspace. Then

(i) If $(X, \mathcal{T})$ is first countable, so is $(Y, \mathcal{T}_Y)$.

(ii) If $(X, \mathcal{T})$ is second countable, so is $(Y, \mathcal{T}_Y)$.

XIII.1.4.13. It is not true that if $(X, \mathcal{T})$ is separable, then $(Y, \mathcal{T}_Y)$ is necessarily separable. There are many counterexamples, e.g. XIII.7.10.14.(h), XIII.7.10.18.(d), XIII.11.14.4..

XIII.1.5. Separated Union

There are many operations making new topological spaces from old ones. One of the simplest is separated union.

XIII.1.5.1. Definition. Let $(X, \mathcal{T})$ be a topological space, and $\{X_i : i \in I\}$ a collection of pairwise disjoint nonempty subsets of X whose union is X (i.e. a partition of X). Then X is the *separated union* of the X_i if a subset U of X is open (i.e. $U \in \mathcal{T}$) if and only if $U \cap X_i$ is relatively open in X_i for every $i \in I$.

XIII.1.5.2. If $U \in \mathcal{T}$, then by definition $U \cap X_i$ is relatively open in X_i for each i. So the important condition is the converse, which is not automatic [consider a partition of X into singleton sets; then X is the separated union if and only if its topology is the discrete topology.] Some authors call a separated union a *disjoint union*, but the term "separated union" seems clearer and more descriptive.

A subset U of a topological space X is called *clopen* if both U and $X \setminus U$ are open in X. ("Clopen" means "both open and closed"; cf. ().)

XIII.1.5.3. PROPOSITION. Let $\{X_i : i \in I\}$ be a partition of a topological space X. Then X is the separated union of the X_i if and only if each X_i is a clopen subset of X.

PROOF: Suppose X is the separated union of the X_i. For each fixed $j \in I$, $X_j \cap X_i$ is either X_j (if $i = j$) or $\emptyset$ (if $i \neq j$). Thus $X_j \cap X_i$ is relatively open in X_i for each i; hence X_j is open in X. So for any $J \subseteq I$, $\cup_{j\in J} X_j$ is open in X. In particular, for each j, $X \setminus X_j = \cup_{i\neq j} X_i$ is open in X, hence X_j is clopen in X. More generally, if $J \subseteq I$, $X \setminus \cup_{j\in J} X_j = \cup_{i\in I\setminus J} X_i$ is open in X; so $\cup_{j\in J} X_j$ is clopen in X.

Conversely, suppose each X_i is clopen (and in particular open) in X. Let U be a subset of X such that $U \cap X_i$ is relatively open in X_i for each i. Then, since X_i is open in X, $U \cap X_i$ is open in X for each i (), and hence $U = \cup_{i\in I}(U \cap X_i)$ is open in X. 👍

XIII.1.5.4. We can reverse the process. Suppose we are given a family $\{(X_i, \mathcal{T}_i) : i \in I\}$ of topological spaces. Form the set-theoretic disjoint union $X = \sqcup_{i\in I} X_i$ (). Let $\mathcal{T}$ be the collection of all subsets U of X such that $U \cap X_i \in \mathcal{T}_i$ for all i. It is easily checked that $\mathcal{T}$ is a topology on X and $(X, \mathcal{T})$ is the separated union of the $(X_i, \mathcal{T}_i)$.

XIII.1.6. Exercises

XIII.1.6.1. Give a logical critique of the following statement which can be found in many books:

> "A metric space is a topological space."

As part of your answer, you might consider, if ρ and σ are equivalent metrics on a set X, whether the spaces (X, ρ) and (X, σ) are the "same" or "different."

XIII.1.6.2. Show that if $\{\mathcal{T}_\alpha\}$ is a family of topologies on a set X, then $\cap_\alpha \mathcal{T}_\alpha$ is also a topology on X.

XIII.1.6.3. Prove Proposition XIII.1.2.5..

XIII.1.6.4. (a) Show that the finite complement topology (XIII.1.2.8.(iv)) on a set X is a topology. Show that it is the discrete topology if and only if X is finite.

(b) Let X be a set. The *countable complement topology* on X consists of $\emptyset$ and all subsets of X with countable complement. Show that this is a topology on X, which is the discrete topology if and only if X is countable. If X is uncountable, show that the countable complement topology is not first countable.

XIII.1.6.5. Explain why the following argument (which appears or is implicit in many books and does not use the AC) is not adequate to prove that the set $\mathcal{T}_Y$ defined in XIII.1.4.1. is closed under arbitrary unions and is thus a topology:

"Let $\{U_i \cap Y : i \in I, U_i \in \mathcal{T}\}$ be an indexed collection of relatively open sets (sets in $\mathcal{T}_Y$). Then

$$\bigcup_{i\in I}(U_i \cap Y) = \left(\bigcup_{i\in I} U_i\right) \cap Y \in \mathcal{T}_Y$$

since $\cup_{i\in I} U_i \in \mathcal{T}$."

XIII.1.6.6. Show that XIII.1.3.5. can be rephrased as follows:

PROPOSITION. Let X be a second countable topological space, and $\mathcal{U} = \{U_i : i \in I\}$ a collection of open sets in X with union U. Then there is a countable subcollection of $\mathcal{U}$ whose union is U.

XIII.1.6.7. Let $(X, \mathcal{T})$ be a topological space. A *pseudo-base* for $\mathcal{T}$ is a collection $\{U_i : i \in I\}$ of nonempty open sets such that every nonempty open set contains at least one U_i.

(i) Show that a base for a topology (consisting of nonempty sets) is a pseudo-base. In particular, a second countable topological space has a countable pseudo-base. (But a pseudo-base is not necessarily a base: the union of the sets in a pseudo-base is not necessarily even all of X in general, although it is dense.)

(ii) Show that every topological space with a countable pseudo-base is separable.

(iii) Show that a topological space which is separable and first countable has a countable pseudo-base (more generally, only a countable dense set of points which each have a countable local base is needed). Since there are even compact Hausdorff spaces which are separable and first countable but not second countable (XIII.11.14.4.), a space with a countable pseudo-base is not necessarily second countable. There are also separable compact Hausdorff spaces which do not have a countable pseudo-base ().

(iv) If $(X, \mathcal{T})$ is a topological space, and X has a dense open subspace Y which has a countable pseudo-base in the relative topology, then X has a countable pseudo-base. Thus if X is locally compact and has a countable pseudo-base, then any compactification of X (e.g. the Stone-Čech compactification βX) has a countable pseudo-base. In particular, $\beta\mathbb{N}$ and $\beta\mathbb{R}$ have countable pseudo-bases but are not even first countable (this also follows from the general version of (iii).)

(The Countable AC is needed for (ii) and (iii).)

XIII.2. Closed Sets and Limit Points

XIII.2.1. Closed Sets

Complements of open sets in a topological space are called closed sets:

XIII.2.1.1. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. A subset A of X is *closed* in the topology $\mathcal{T}$, or $\mathcal{T}$-*closed*, if $A^c = X \setminus A \in \mathcal{T}$.

There is an important alternate description of closed sets (XIII.2.3.6.).

XIII.2.1.2. EXAMPLES. (i) A subset of $\mathbb{R}$ is closed in the usual topology if and only if it is closed in the sense of ().

(ii) If X has the discrete topology, then every subset of X is closed.

(iii) If X has the indiscrete topology, then the only closed subsets of X are $\emptyset$ and X.

XIII.2.1.3. It is important to note that being closed is *not* the opposite of being open: on the one hand, a subset of X can be both open and closed (cf. XIII.2.1.2.(ii)), and on the other hand a subset of X need not be either open or closed – in fact typically "most" subsets of a topological space are neither open nor closed.

From de Morgan's laws we immediately obtain the following properties of closed sets:

XIII.2.1.4. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space, and $\mathcal{E}$ the collection of $\mathcal{T}$-closed sets. Then

(i) $\emptyset, X \in \mathcal{E}$.

(ii) $\mathcal{E}$ is closed under finite unions.

(iii) $\mathcal{E}$ is closed under arbitrary intersections.

XIII.2.1.5. Conversely, if X is a set and $\mathcal{E}$ is a collection of subsets of X satisfying XIII.2.1.4.(i)–(iii), then the complements of the sets in $\mathcal{E}$ form a topology $\mathcal{T}$ for which $\mathcal{E}$ is precisely the collection of closed sets. Thus a topology can be equally well defined by specifying the closed sets instead of the open sets.

XIII.2.1.6. There is an asymmetry between unions and intersections for both open sets and closed sets. An arbitrary union of closed sets is not closed in general: for example, in $\mathbb{R}$ with its usual topology, points (singleton subsets) are closed, and every subset of $\mathbb{R}$ is a union of singleton sets. Correspondingly, an arbitrary intersection of open sets is not open in general. In fact, this asymmetry is the crucial underlying feature of topology, and it is what gives the subject its power and usefulness.

XIII.2.2. Limit Points of a Set

One of the most fundamental notions in topology is limit point. Roughly speaking, a point x in a topological space X is a limit point of a set A if there are points of A "arbitrarily close" to x. More specifically:

XIII.2.2.1. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, $A \subseteq X$, and $x \in X$. Then x is a *limit point* of A (in the topology $\mathcal{T}$) if every open neighborhood of x contains at least one point of A.

Note that if $x \in A$, then x is trivially a limit point of A since every neighborhood of x contains at least one point of A, namely x itself. But the importance of the notion of limit point is that limit points of A are not necessarily in A. For example, 0 and 1 are limit points of the open interval $(0, 1)$ in $\mathbb{R}$ with its usual topology. Whether a point x is a limit point of a subset A can depend crucially on the topology used (if $x \notin A$); if there is a possibility of confusion we will use the terminology $\mathcal{T}$-limit point.

There is a closely related but distinct notion:

XIII.2.2.2. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, $A \subseteq X$, and $x \in X$. Then x is an *accumulation point* of A (in the topology $\mathcal{T}$) if every deleted open neighborhood of x contains at least one point of A, i.e. if every open neighborhood of x contains at least one point of A other than x.

Accumulation points are called *cluster points* by some authors, but for clarity we will only use the term "cluster point" in reference to sequences or nets (where the meaning is somewhat different, cf. XIII.3.1.7.), not subspaces.

XIII.2.2.3. Unlike for limit points, a point in A is not necessarily an accumulation point of A; a point of A which is not an accumulation point of A is called an *isolated point* of A. A point of X which is not in A is a limit point of A if and only if it is an accumulation point of A. There is some nonuniformity of terminology in the literature: some authors take "limit point" to mean what we call "accumulation point." It will sometimes be convenient to have separate terms for the two notions, and for the purposes of this section limit points are more important. See XIII.2.5. for a discussion of accumulation points and isolated points.

XIII.2.3. Closure of a Set

XIII.2.3.1. If $(X, \mathcal{T})$ is a topological space and $A \subseteq X$, then there is a smallest closed subset of X containing A, namely the intersection of all closed subsets of X containing A (there is at least one such set, X; recall that any intersection of closed sets is closed).

XIII.2.3.2. DEFINITION. Let $(X, \mathcal{T})$ be a topological space and $A \subseteq X$. The smallest $\mathcal{T}$-closed subset of X containing A is called the *closure* of A in X (or $\mathcal{T}$-closure of A).

The standard notation for the closure of A is $\bar{A}$; the notation $Cl(A)$ is also common. The closure of A depends on X and $\mathcal{T}$; when this needs to be made explicit we will use the notation $Cl_X(A)$, $Cl_{\mathcal{T}}(A)$, or $Cl_{(X,\mathcal{T})}(A)$.

Here are some elementary properties of closure:

XIII.2.3.3. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space and $A, B \subseteq X$. Then

(i) $A \subseteq \bar{A}$.

(ii) A is closed if and only if $A = \bar{A}$.

(iii) $\bar{\bar{A}} = \bar{A}$.

(iv) If $A \subseteq B$, then $\bar{A} \subseteq \bar{B}$.

(v) $\overline{A \cup B} = \bar{A} \cup \bar{B}$.

(vi) $\bar{\emptyset} = \emptyset$ and $\bar{X} = X$.

PROOF: (i) and (ii) are trivial from the definition, and (iii) and (vi) follow immediately from (ii). (iv) is also immediate from the definition [$\bar{B}$ is a closed set containing A]. For (v), note that $\bar{A} \cup \bar{B}$ is a union of two closed sets, hence closed, and contains $A \cup B$, so $\overline{A \cup B} \subseteq \bar{A} \cup \bar{B}$. On the other hand, $A \subseteq A \cup B$, so $\bar{A} \subseteq \overline{A \cup B}$ by (iv), and similarly $\bar{B} \subseteq \overline{A \cup B}$. 👍

XIII.2.3.4. If $A, B \subseteq X$, it is not generally true that $\overline{A \cap B} = \bar{A} \cap \bar{B}$. For example, if $A = (0, 1)$ and $B = (1, 2)$ in $\mathbb{R}$ (with its usual topology), then $\bar{A} \cap \bar{B} = \{1\}$. But $A \cap B = \emptyset$, so $\overline{A \cap B} = \emptyset$.

However, we do always have $\overline{A \cap B} \subseteq \bar{A} \cap \bar{B}$, since the set on the right is a closed set containing $A \cap B$. More generally, for any collection $\{A_i : i \in I\}$ of subsets of X, we have $\overline{\cap_{i \in I} A_i} \subseteq \cap_{i \in I} \bar{A}_i$ by the same argument. A similar argument shows that $\cup_{i \in I} \bar{A}_i \subseteq \overline{\cup_{i \in I} A_i}$, but we do not have equality in general if I is infinite.

The definition of closure is precise but often hard to apply directly. There is a more generally useful characterization of closure:

XIII.2.3.5. THEOREM. Let $(X, \mathcal{T})$ be a topological space and $A \subseteq X$. Then the closure of A is exactly the set of all limit points (XIII.2.2.1.) of A.

PROOF: Let U be the complement of $\bar{A}$; then U is open. Suppose x is a limit point of A. If $x \notin \bar{A}$, then U is an open neighborhood of x containing no points of A, contradicting that x is a limit point of A. Thus $x \in \bar{A}$.

Now suppose x is not a limit point of A. Then there is an open neighborhood V of x containing no point of A. So V^c is a closed set with $A \subseteq V^c$, and hence $\bar{A} \subseteq V^c$, $x \notin \bar{A}$. 👍

XIII.2.3.6. COROLLARY. Let $(X, \mathcal{T})$ be a topological space. A subset of X is closed if and only if it contains all its limit points.

Dense Sets and Separability

XIII.2.3.7. Definition. Let $(X, \mathcal{T})$ be a topological space, and $D \subseteq X$. Then D is *dense* in X if $\bar{D} = X$.

XIII.2.3.8. Proposition. Let X be a topological space, and $D \subseteq X$. Then D is dense in X if and only if every nonempty open subset of X contains a point of D.

An intersection of two dense subsets is not necessarily dense, or even nonempty: both $\mathbb{Q}$ and $\mathbb{J}$ are dense in $\mathbb{R}$. But if at least one of the sets is open, the intersection is dense:

XIII.2.3.9. Proposition. Let X be a topological space, and U and V dense subsets of X. If U is open, then $U \cap V$ is dense in X.

Proof: Let W be a nonempty open set in X. Then $U \cap W$ is an open set in X, which is nonempty since U is dense. Then $W \cap (U \cap V) = (U \cap W) \cap V$ is nonempty since V is dense. 👍

XIII.2.3.10. Corollary. Let X be a topological space. The intersection of any finite number of dense open subsets of X is a dense open subset of X.

In some, but not all, topological spaces, an intersection of a *countable* number of dense open sets is always dense (but not necessarily open). See XIII.12.2.1..

Some topological spaces have "small" dense subsets, which can cause better behavior:

XIII.2.3.11. Definition. A topological space is *separable* if it has a countable dense subset.

Note that separability in this sense has nothing to do with the separation axioms of XIII.7.. The coincidence of terminology is regrettable, but well established. ✋

XIII.2.3.12. Examples. (i) If X has the indiscrete topology, then any nonempty subset of X is dense. Thus X is separable.

(ii) If X has the discrete topology, no proper subset of X is dense in X. Thus X is separable if and only if it is countable. In fact, a countable set with any topology is separable.

(iii) If X is an infinite set with the finite complement topology, any infinite subset of X is dense. Thus X is separable if and only if it contains a countably infinite subset (this will always be true if the Countable AC is assumed).

(iv) If $\mathbb{R}$ has the usual topology, $\mathbb{Q}$ is a countable dense subset, so $\mathbb{R}$ is separable. More generally, the points of $\mathbb{R}^n$ with rational coordinates form a countable dense set (for the usual topology), so $\mathbb{R}^n$ is separable.

XIII.2.3.13. Separability seems to be related to first or (especially) second countability. There are some connections, as the next two results show. But a topological space can be separable without being first countable (an uncountable set with the finite complement topology; cf. also XIII.6.3.2.–XIII.6.3.4.), first countable without being separable (an uncountable set with the discrete topology), and both first countable and separable without being second countable (XIII.11.14.4.).

XIII.2.3.14. PROPOSITION. A second countable topological space is separable.

PROOF: If $\{U_n : n \in \mathbb{N}\}$ is a countable base for the topology consisting of nonempty sets, choose $x_n \in U_n$ for each n. Then $\{x_n : n \in \mathbb{N}\}$ is a countable dense set in X. 👍

This proof requires the Countable AC, and the result cannot be proved in ZF.

XIII.2.3.15. THEOREM. A metrizable topological space is second countable if and only if it is separable.

PROOF: Let (X, ρ) be a metric space. If X is second countable, it is separable by XIII.2.3.14.. Conversely, suppose X has a countable dense set $D = \{x_n : n \in \mathbb{N}\}$. Let $\mathcal{B}$ be the collection of all open balls of the form $B_r(x_n)$, where $x_n \in D$ and $r \in \mathbb{Q}$. Then $\mathcal{B}$ is a countable collection of open sets in X. We claim $\mathcal{B}$ is a base for the topology. To show this, we need only show that if $x \in X$ and $\epsilon > 0$, there is an n and r such that $x \in B_r(x_n)$ and $B_r(x_n) \subseteq B_\epsilon(x)$. Choose n such that $\rho(x, x_n) < \frac{\epsilon}{3}$ and $r \in \mathbb{Q}$ such that $\frac{\epsilon}{3} < r < \frac{2\epsilon}{3}$. Then $x \in B_r(x_n)$, and $B_r(x_n) \subseteq B_\epsilon(x)$ by the triangle inequality. 👍

The implication proved in this proof does not require any form of Choice, since only finitely many choices must be made.

XIII.2.3.16. More generally, we can define the *density* $d(X)$ of a topological space X to be the minimum cardinality of a dense subset of X (this is only well defined in general if the AC is assumed). X is separable if and only if $d(X) \leq \aleph_0$. A slight variation of the proof of XIII.2.3.14. shows that $d(X) \leq w(X)$ (XIII.1.2.11.) for any X. We can have $d(X) < w(X)$ (a separable space is not necessarily second countable), but if X is regular we have $w(X) \leq 2^{d(X)}$ (XIII.7.10.7.).

Interior of a Set

There is a dual notion to closure:

XIII.2.3.17. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, and $A \subseteq X$. The *interior* of A (in X, in the topology $\mathcal{T}$) is the largest open ($\mathcal{T}$-open) subset of X contained in A.

There is a largest such open set, the union of all open sets contained in A; there is always one such set, the empty set. The interior of A can very well be empty. The standard notation for the interior of A is A^o; the notation $Int(A)$ is also used.

XIII.2.3.18. Note that the interior of a subset A of X is a relative notion, not intrinsic to A (if A is given the relative topology, the interior of A with respect to A is all of A!) The interior of A is also dependent on the topology $\mathcal{T}$, although the notation A^o does not reflect this; the notation $Int(A)$ is thus more flexible, since it can be embellished as $Int_X(A)$, $Int_{\mathcal{T}}(A)$, or $Int_{(X,\mathcal{T})}(A)$ when necessary.

XIII.2.3.19. Example. Let $X = \mathbb{R}$ with its usual topology, $A = [0, 1]$. Then $A^o = (0, 1)$. If, however, $X = \mathbb{R}^2$ and A is the interval $[0, 1]$ on the x-axis, then $A^o = \emptyset$.

The relation between interior and closure is:

XIII.2.3.20. Proposition. Let X be a topological space, and $A \subseteq X$. Then $(A^o)^c = (A^c)^-$, i.e. $A^o = [(A^c)^-]^c$.

From this, de Morgan's laws, and XIII.2.3.3. we obtain the basic properties of interior:

XIII.2.3.21. Proposition. Let $(X, \mathcal{T})$ be a topological space and $A, B \subseteq X$. Then

(i) $A^o \subseteq A$.

(ii) A is open if and only if $A^o = A$.

(iii) $(A^o)^o = A^o$.

(iv) If $A \subseteq B$, then $A^o \subseteq B^o$.

(v) $(A \cap B)^o = A^o \cap B^o$.

(vi) $\emptyset^o = \emptyset$ and $X^o = X$.

XIII.2.3.22. We always have $A^o \cup B^o \subseteq (A \cup B)^o$, but equality does not hold in general. A dramatic example is to take $X = \mathbb{R}$, $A = \mathbb{Q}$, and $B = \mathbb{J}$; then $A^o = B^o = \emptyset$ but $(A \cup B)^o = X$.

Boundary of a Subset

XIII.2.4. Specification of a Topology via Closure

Instead of defining a topology by specifying the open sets, we can define it by specifying the closure of each set (this will determine the closed sets). This procedure can be axiomatized by a list of properties of the closure operation. This approach to topology is generally attributed to Kuratowski, although he himself in [Kur66] credited the procedure to F. Riesz.

XIII.2.4.1. DEFINITION. Let X be a set. A *Kuratowski closure operation* on X is an assignment to each subset A of X a subset $\bar{A}$ of X, called the *closure* of A, with the following properties:

(i) $\bar{\emptyset} = \emptyset$.

(ii) $A \subseteq \bar{A}$ for all $A \subseteq X$.

(iii) $\bar{\bar{A}} = \bar{A}$ for all $A \subseteq X$.

(iv) $\overline{A \cup B} = \bar{A} \cup \bar{B}$ for all $A, B \subseteq X$.

XIII.2.4.2. EXAMPLES. (i) Let $(X, \mathcal{T})$ be a topological space. For $A \subseteq X$, let $\bar{A}$ be the closure of A in the topological sense (XIII.2.3.2.). Then $A \mapsto \bar{A}$ is a Kuratowski closure operation (XIII.2.3.3.). We will show the converse: if $A \mapsto \bar{A}$ is a Kuratowski closure operation on X, then there is a unique topology on X such that the closure operation is closure with respect to the topology (XIII.2.4.4.).

(ii) If X is a nonempty set, set $\bar{A} = X$ for every $A \subseteq X$. Then this operation satisfies XIII.2.4.1.(ii), (iii), and (iv), but not (i).

(iii) If X is a nonempty set, set $\bar{A} = \emptyset$ for all $A \subseteq X$. Then this operation satisfies XIII.2.4.1.(i), (iii), and (iv), but not (ii).

(iv) Let $X = \{1, 2, 3\}$. Set $\bar{\emptyset} = \emptyset$, $\overline{\{1\}} = \{1, 2\}$, and $\bar{A} = X$ for every other A. Then this operation satisfies XIII.2.4.1.(i), (ii), and (iv), but not (iii).

(v) Let $X = \{1, 2, 3\}$. Set $\bar{\emptyset} = \emptyset$, $\overline{\{1, 2\}} = \{1, 2\}$, and $\bar{A} = X$ for every other A. Then this operation satisfies XIII.2.4.1.(i), (ii), and (iii), but not (iv).

(ii)–(v) show that the axioms in XIII.2.4.1. are independent, and hence all are necessary to specify a topology.

The most powerful of the axioms is (iv):

XIII.2.4.3. PROPOSITION. Let $A \mapsto \bar{A}$ be an operation on X satisfying (iv). Then

(i) If $A \subseteq B \subseteq X$, then $\bar{A} \subseteq \bar{B}$.

(ii) If $\{A_i : i \in I\}$ is a collection of subsets of X, then $\cup_{i \in I} \bar{A}_i \subseteq \overline{\cup_{i \in I} A_i}$.

(iii) If $\{A_i : i \in I\}$ is a collection of subsets of X, then $\overline{\cap_{i \in I} A_i} \subseteq \cap_{i \in I} \bar{A}_i$.

PROOF: (i): If $A \subseteq B$, then

$$\bar{B} = \overline{A \cup B} = \bar{A} \cup \bar{B}$$

which implies that $\bar{A} \subseteq \bar{B}$.

(ii): For each $j \in I$, $A_j \subseteq \cup_{i \in I} A_i$, so $\bar{A}_j \subseteq \overline{\cup_{i \in I} A_i}$ by (i).

(iii): For each $j \in I$, $\cap_{i \in I} A_i \subseteq A_j$, so $\overline{\cup_{i \in I} A_i} \subseteq \bar{A}_j$ by (i).



XIII.2.4.4. THEOREM. Let $A \mapsto \bar{A}$ be a Kuratowski closure operation on a set X. Then there is a unique topology $\mathcal{T}$ on X such that the closure operation is precisely closure in the topology $\mathcal{T}$.

PROOF: If there is such a topology, then the closed sets for the topology must be precisely the sets A for which $A = \bar{A}$ by axiom (iii), so there is at most one such topology. So we must show that the collection of such sets defines a topology (i.e. the complements of such sets form a topology on X). Call such sets *closed*. By XIII.2.1.5., we must show that $\emptyset$ and X are closed, and that the collection of closed sets is closed under finite unions and arbitrary intersections.
We have $\bar{\emptyset} = \emptyset$ by axiom (i), and $X \subseteq \bar{X}$ by axiom (ii), hence $X = \bar{X}$.
Suppose $A = \bar{A}$ and $B = \bar{B}$. Then $\overline{A \cup B} = \bar{A} \cup \bar{B} = A \cup B$ by axiom (iv), so $A \cup B$ is closed. By iteration (or induction), the collection of closed sets is closed under finite unions.
Suppose $A_i = \bar{A}_i$ for each $i \in I$. Then $\overline{\cap_{i\in I} A_i} \subseteq \cap_{i\in I} \bar{A}_i = \cap_{i\in I} A_i$ by XIII.2.4.3.(iii), and the opposite inclusion follows from axiom (ii). Thus the collection of closed sets is closed under arbitrary intersections. 👍

XIII.2.4.5. Suppose $A \mapsto \bar{A}$ is an operation on a set X satisfying axiom (iv) and also

(v) $\overline{\{p\}} = \{p\}$ for every $p \in X$.

Then the operation satisfies axiom (ii) by XIII.2.4.3.(ii) since any subset of X is a union of singleton subsets. If X has two distinct points p and q, then $\bar{\emptyset} \subseteq \overline{\{p\}} = \{p\}$ by XIII.2.4.3.(i), and similarly $\bar{\emptyset} \subseteq \{q\}$, so the operation also satisfies axiom (i). Thus if it also satisfies axiom (iii), it is a Kuratowski closure operation. A Kuratowski closure operation satisfies axiom (v) if and only if the corresponding topology is T_1 ().

XIII.2.5. Isolated Points and Derived Sets

We now look at isolated points and accumulation points (XIII.2.2.2.) in more detail. In order to avoid some cardinal-theoretic difficulties, we will assume the AC.

XIII.2.5.1. A point p in a topological space X is an *isolated point* of X if it is not in the closure of $X \setminus \{p\}$, i.e. if $\{p\}$ is an open set in X. (In some references, an isolated point is also required to be closed, which is automatic in a T_1 space; but we will not make this restriction.) A topological space is *dense-in-itself* if it contains no isolated points. A topological space X is *scattered* if every nonempty subset of X contains an isolated point in the relative topology, i.e. if no (nonempty) subset of X is dense-in-itself.

Any subset of a scattered space is scattered. A subset of a space which is dense-in-itself is not necessarily dense-in-itself (it could even be scattered); but an open subset of a space which is dense-in-itself must also be dense-in-itself.

XIII.2.5.2. If A is a subset of a topological space X, denote by A' the set of accumulation points of A in X. A' is called the *derived set* of A in X. A subset A of X is closed if and only if $A' \subseteq A$. We say A is a *perfect* subset of X if $A' = A$. A is perfect in X if and only if it is closed and dense-in-itself.

Note that the properties of being scattered or dense-in-itself are intrinsic properties of a space (or subspace), but being perfect is a relative property.

The next theorem, the main result of this section, is often stated in references in much less generality than necessary (the proof does not even become simpler for these special cases!) Recall (XIII.1.2.11.) that the weight $w(X)$ of a topological space X is the smallest cardinality of a base for the topology of X.

XIII.2.5.3. Theorem. [General Cantor-Bendixson Theorem] Let X be a topological space. Then there is a unique decomposition of X as a disjoint union of a perfect subset X^p and a (necessarily open) scattered subset X^s. We have $card(X^s) \leq \min(card(X), w(X))$.

Note that X^p or X^s can be empty: $X^s = \emptyset$ if (and only if) X is dense-in-itself, and $X^p = \emptyset$ if (and only if) X is scattered.

Proof: The proof will actually show more than the stated result; the additional conclusions will be formalized at the end of the proof.

Define a decreasing collection of closed subsets of X by transfinite induction as follows. Set $X^{(0)} = X$. For any ordinal α, if $X^{(\alpha)}$ has been defined, set $X^{(\alpha+1)} = (X^{(\alpha)})'$. If α is a limit ordinal, set $X^{(\alpha)} = \cap_{\beta<\alpha} X^{(\beta)}$. We have $X^{(\alpha)} \subseteq X^{(\beta)}$ if $\beta < \alpha$. Note that $X^{(1)}$ is X with the isolated points removed; since each isolated point is open, $X^{(1)}$ is closed in X. By the same argument, $X^{(\alpha+1)}$ is relatively closed in $X^{(\alpha)}$; thus by transfinite induction $X^{(\alpha)}$ is a closed subset of X for all α.

The $X^{(\alpha)}$ cannot continue to strictly decrease (it will stabilize before $card(\alpha) > card(X)$); there must be a smallest α for which $X^{(\alpha+1)} = X^{(\alpha)}$. If $X^p = X^{(\alpha)}$ for this α, then X^p is a perfect subset of X.

Set $X^s = X \setminus X^p$; then X^s is open in X. Suppose Y is a subset of X which is dense-in-itself. If $Y \cap X^s \neq \emptyset$, then there is a largest β such that $Y \subseteq X^{(\beta)}$, and $\beta < \alpha$. Thus there is a $y \in Y \cap (X^{(\beta)} \setminus X^{(\beta+1)})$. But then y is an isolated point of $X^{(\beta)}$ and hence an isolated point of Y, a contradiction. Thus every dense-in-itself subset of X is contained in X^p, and X^s is scattered.

For uniqueness, suppose $X = Z \cup W$, where Z is perfect and W scattered. We have seen that $Z \subseteq X^p$. If $X^p \cap W \neq \emptyset$, it is a relatively open subset of X^p which contains an isolated point, a contradiction.

For the cardinality assertion, let $\mathcal{B}$ be a base for the topology of X with $card(\mathcal{B}) = w(X)$. First note that if p is an isolated point of X, then $\{p\} \in \mathcal{B}$. By a similar argument, for each $x \in X^{(\alpha)} \setminus X^{(\alpha+1)}$ there is an open set $U_x \in \mathcal{B}$ with $U_x \cap X^{(\alpha)} = \{x\}$. It is clear that $x \mapsto U_x$ is an injective map from X^s to $\mathcal{B}$, so $card(X^s) \leq w(X)$. The assertion $card(X^s) \leq card(X)$ is trivial. 👍

The next corollary, or special cases of it, is usually called the "Cantor-Bendixson Theorem." Actually Cantor and Bendixson (who worked independently) only stated the result for subsets of $\mathbb{R}^n$ since the formalism of general topology had not yet been developed.

XIII.2.5.4. Corollary. [Cantor-Bendixson Theorem] Let X be a second countable topological space. Then X can be written uniquely as a disjoint union of a perfect subset X^p and a (necessarily open) scattered subset X^s. In addition, X^s is countable.

XIII.2.5.5. The set $X^{(\alpha)}$ is called the *α'th derived set* of X, and the smallest α for which $X^{(\alpha+1)} = X^{(\alpha)} = X^p$ is called the *(Cantor-Bendixson) height* $ht(X)$ of X. The chain of derived subsets and the height are homeomorphism invariants of X. The height is an ordinal invariant of X, not a cardinal invariant; if X is second countable, then $ht(X)$ is a countable ordinal, but it is not always a finite ordinal or even ω. In fact, every countable ordinal α occurs as the height of a compact subset X_α of $\mathbb{R}$ (). More generally, every ordinal occurs as the height of a compact Hausdorff space.

XIII.2.6. Exercises

XIII.2.6.1. (cf. [Dug78, III.5]) (a) Define a notion of "interior operation" on a set in analogy with the definition of a Kuratowski closure operation, using the properties from XIII.2.3.21., and prove the analog of XIII.2.4.4. that a topology can be specified by such an interior operation.

(b) Do the same for "boundary operation," using the properties of ().

(c) Do the same for the "derived set" operation $A \mapsto A'$ (XIII.2.5.2.).

XIII.2.6.2. (a) Show that if X is a T_1 topological space and $A \subseteq X$, then A' is closed in X.

(b) Show by example that if X is just T_0, A' need not be closed in X.

XIII.2.6.3. Let X be a topological space. A *block* in X is the closure of an open set in X. (Some authors call such a set *canonically closed*, but this is poor terminology since there is nothing canonical about a block.)

(a) Show that a block is the closure of its interior.

(b) Show that any finite union of blocks is a block. What about a finite intersection of blocks?

(c) An open set in X is a *regular open set* if it is the interior of its closure, i.e. the interior of a block. Show that every open set U is contained in a smallest regular open set, whose closure is the same as the closure of U.

XIII.3. Convergence

One of the most important uses of a metric is to define limits of sequences. Limits of sequences can be defined nicely in any topological space, although not all properties of limits of sequences in metric spaces carry over to this case. And it turns out that for (most – cf. XIII.3.5.2.) topological spaces which are not first countable, the theory of limits of sequences is not adequate to fully describe convergence and continuity in the space, and one must use a generalized version of sequences known as *nets*. An alternate equivalent point of view with some technical advantages uses filters in place of nets.

XIII.3.1. Limits of Sequences

Limits of sequences in topological spaces are defined just as in metric spaces, replacing open balls by open neighborhoods:

XIII.3.1.1. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, and let (x_n) be a sequence in X, and $x \in X$. Then (x_n) *converges to* x in $(X, \mathcal{T})$, denoted $x_n \to x$ or $\lim\limits_{n\to\infty} x_n = x$, if for every open neighborhood U of x there is an N such that $x_n \in U$ for all $n \geq N$. (We say that x_n is *eventually in* U.)

If there is any possible ambiguity about which topology is used on X, we write $x_n \to_{\mathcal{T}} x$ or $\mathcal{T}\text{-}\lim\limits_{n\to\infty} x_n = x$.

XIII.3.1.2. The word "open" can be deleted from this definition without changing its meaning, since every neighborhood of x contains an open neighborhood. Similarly, the existence of an N need only be shown for U's in a fixed local base for the topology at x.

XIII.3.1.3. Unlike in metric spaces, limits of sequences need not be unique: for example, in a space with the indiscrete topology every sequence converges to every point. This undesirable situation does not occur in Hausdorff spaces (XIII.7.4.1.).

Subsequences and Tails

Subsequences and tails of sequences are defined just as in (). Convergence of a sequence depends only on the tails of the sequence, and implies convergence of each subsequence, just as in the metric case (the proof is essentially identical):

XIII.3.1.4. PROPOSITION. Let X be a topological space, and (x_n) a sequence in X, $x \in X$.

(i) $x_n \to x$ if and only if some tail of (x_n) converges to x.

(ii) If $x_n \to x$, then every subsequence of (x_n) converges to x. In particular, every tail of (x_n) converges to x.

There is a sort of converse to (ii), which is almost harder to state than to prove:

XIII.3.1.5. PROPOSITION. Let X be a topological space, and (x_n) a sequence in X, $x \in X$. If (x_n) does not converge to x, then there is a subsequence of (x_n), no subsequence of which converges to x.

PROOF: If (x_n) does not converge to x, then there is an open neighborhood U of x such that $x_n \notin U$ for infinitely many n. These x_n form a subsequence with no terms in U, so no subsequence of it can converge to x. 👍

The result is often stated in the contrapositive:

XIII.3.1.6. COROLLARY. Let X be a topological space, and (x_n) a sequence in X, $x \in X$. If every subsequence of (x_n) has a subsequence which converges to x, then $x_n \to x$.

Cluster Points of Sequences

XIII.3.1.7. DEFINITION. Let X be a topological space, and (x_n) a sequence in X, $x \in X$. Then x is a *cluster point* of (x_n) if every open neighborhood U of x contains x_n for infinitely many n (we say x_n is *frequently in* U).

The proof of the next Proposition is essentially identical to the proof of ().

XIII.3.1.8. PROPOSITION. Let X be a topological space, and (x_n) a sequence in X. For each n, let F_n be the closure in X of $\{x_k : k \geq n\}$. Then $\cap_n F_n$ is exactly the set of cluster points of (x_n). In particular, the set of cluster points of a sequence is a closed set.

If $x_n \to x$, or more generally if there is a subsequence of (x_n) which converges to x, then x is a cluster point of (x_n). The converse is not true in general (), but is true if the topological space is first countable:

XIII.3.1.9. PROPOSITION. Let X be a topological space, (x_n) a sequence in X, and $x \in X$. If x is a cluster point of (x_n), and there is a countable local base for the topology at x (in particular, if X is first countable), then there is a subsequence of (x_n) which converges to x.

PROOF: Let (U_n) be a decreasing sequence of open neighborhoods of x forming a local base (). Let $x_{k_1} \in U_1$, and inductively choose $k_{n+1} > k_n$ with $x_{k_{n+1}} \in U_{n+1}$; this is possible since $x_k \in U_{n+1}$ for infinitely many k. (For definiteness and to avoid use of the Countable AC, choose k_{n+1} to be the smallest $k > k_n$ for which $x_k \in U_{n+1}$.) Then (x_{k_n}) is a subsequence of (x_n), and $x_{k_n} \in U_m$ for $n \geq m$. If V is any neighborhood of x, then $U_m \subseteq V$ for some m, so $x_{k_n} \in V$ for all $n \geq m$. Thus $x_{k_n} \to x$. 👍

XIII.3.1.10. If A is a subset of a topological space X, $x \in X$, and there is a sequence (x_n) in A with $x_n \to x$, then x is obviously a limit point of A, so $x \in \bar{A}$. The converse is not true in general (), but is true if the topological space is first countable:

XIII.3.1.11. PROPOSITION. Let A be a subset of a topological space X and $x \in X$. If there is a countable local base for the topology at x (in particular, if X is first countable), then $x \in \bar{A}$ if and only if there is a sequence (x_n) of points of A which converges to x.

PROOF: The proof is very similar to the proof of XIII.3.1.9.. One direction is noted above, and does not require a countable local base. Conversely, suppose $x \in \bar{A}$. Let (U_n) be a decreasing sequence of open neighborhoods of x forming a local base. Then $A \cap U_n$ is nonempty for every n. Choose $x_n \in A \cap U_n$ for each n. If V is any neighborhood of x, then $U_m \subseteq V$ for some m, so $x_n \in V$ for all $n \geq m$. Thus $x_n \to x$. 👍

Note that the Countable AC is needed for this proof.

XIII.3.1.12. Thus, if X is a first countable topological space, then (assuming the Countable AC) convergence of sequences tells the whole story about the topology.

Iterated Limits

There is another property of convergent sequences in first countable topological spaces:

XIII.3.1.13. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space and $\{x_{k,n} : k, n \in \mathbb{N}\}$ be a doubly indexed set of points in X. Suppose $\lim_{n\to\infty} x_{k,n} = x_k$ for each k and $\lim_{k\to\infty} x_k = x$. If x has a countable local base (in particular, if X is first countable), then there are strictly increasing sequences $(k_m), (n_m)$ in $\mathbb{N}$ such that $\lim_{m\to\infty} x_{k_m,n_m} = x$.

The sequence (x_{k_m,n_m}) is a sort of "diagonal" sequence, and this property of limits is sometimes called the *diagonal property*.

PROOF: Let (U_m) be a countable decreasing local neighborhood base at x. Then for each m there is a K_m such that $x_k \in U_m$ for all $k \geq K_m$. Then for each $k \geq K_m$ there is an $N_{k,m}$ such that $x_{k,n} \in U_m$ for all $n \geq N_{k,m}$. Inductively choose $k_1 = n_1 = 1$ and $k_m = \max(k_{m-1} + 1, K_m)$ and $n_m = \max(n_{m-1} + 1, N_{k_m,m})$ for $m > 1$. It is routine to check that $x_{k_m,n_m} \to x$. 👍

XIII.3.1.14. This result can fail if x does not have a countable local base (). But the net version holds in general ().

XIII.3.2. Nets and Limits

In a topological space which is not first countable, sequences are generally inadequate to completely describe the behavior of the topology and convergence. But there is a deceptively simple way to generalize sequences, which in complete analogy does fully describe the topology. These generalized sequences are called nets. On the surface, the results of the preceding subsection work for general topological spaces by simply substituting "net" for "sequence" and "subnet" for "subsequence"; however, below the surface there are some subtleties involved, especially with the notion of a subnet.

XIII.3.2.1. DEFINITION. Let X be a topological space (or just a set). A *net* in X is a function from a directed set $(I, \leq)$ to X. The set I is called the *index set* of the net.

A sequence is just a net whose index set is $\mathbb{N}$. If $\phi : I \to X$ is a net with index set I, we usually denote the net by $(x_i)_{i \in I}$ or just by (x_i) if I is understood, where $x_i = \phi(i)$. This notation is a simple extension of sequence notation. However, this notation does not reflect the direction on I, which is part of the definition of the net.

XIII.3.2.2. Note that the index set of a net does not have to be partially ordered. One can in principle restrict to partially ordered index sets in the theory of nets (cf. XIII.3.2.4., XIII.3.2.5., II.4.4.11.). However, it is often convenient to use nets whose index sets are not partially ordered.

XIII.3.2.3. DEFINITION. Let X be a topological space, (x_i) a net in X, and $x \in X$. Then (x_i) *converges to* x, denoted $x_i \to x$ or $\lim_i x_i = x$, if for every open neighborhood U of X there is an $i_0 \in I$ such that $x_i \in U$ whenever $i_0 \leq i$. (We say the net (x_i) is *eventually in* U.)

As for sequences, the word "open" can be deleted, and it suffices to let the U range over a local base at x.

XIII.3.2.4. If the index set of a net is not properly directed (i.e. has a largest element), convergence of the net is pretty trivial: if there is exactly one largest element i_1 in I, then the net automatically converges to x_{i_1}. If there is more than one largest element in I, then the net rarely converges (only if every neighborhood of x_i for each largest element i contains x_j for all the other largest j; this cannot happen in a T_0-space () unless $x_i = x_j$ for all largest i and j, in which case the net converges to this element). The theory of convergence of nets is only interesting if the index set is properly directed.

The convergence of a net depends on the ordering (direction) of I only up to equivalence, i.e. only on the "cofinal behavior" of the direction:

XIII.3.2.5. PROPOSITION. Let $(I, \leq)$ be a directed set, X a topological space, and $(x_i)_{i \in I}$ a net in X with index set I. Let $\precsim$ be another direction on I which is equivalent to $\leq$. If $x \in X$, then $x_i \to x$ using the direction $\leq$ on I if and only if $x_i \to x$ using the direction $\precsim$ on I.

This is obvious using the definition of equivalence of directions ().

Subnets

The definition of a subnet is not an obvious one. Experience has shown that the following definition is appropriate, and flexible enough to be useful:

XIII.3.2.6. DEFINITION. Let X be a topological space, $(I, \leq)$ a directed set, and $\phi : I \to X$ a net in X. A *subnet* of ϕ is a net $\phi \circ f$, where f is a directed function from a directed set $(J, \precsim)$ to $(I, \leq)$. The index set of the subnet is $(J, \precsim)$.

Thus the index set of a subnet of a net is not necessarily the same as the index set of the net.

XIII.3.2.7. A subnet of a sequence is not necessarily a subsequence, for two reasons. First of all, the index set of the subnet is not necessarily $\mathbb{N}$. And even if it is, the function $f : \mathbb{N} \to \mathbb{N}$ defining the subnet is not necessarily injective or increasing. A subnet of a sequence is a subsequence if and only if its index set is $\mathbb{N}$ and its function $f : \mathbb{N} \to \mathbb{N}$ is strictly increasing.

XIII.3.2.8. **Notation:** If we denote the net $\phi : I \to X$ by $(x_i)_{i\in I}$, the net $\phi \circ f$ for $f : J \to I$ will be denoted $(x_{f(j)})_{j\in J}$ or $(x_{i_j})_{j\in J}$, sometimes just $(x_j)_{j\in J}$, where $x_i = \phi(i)$ and $i_j = f(j)$. The "$j \in J$" may be omitted if J is understood. This notation is a little ambiguous, since f is not necessarily injective; we must regard $x_{i_{j_1}}$ and $x_{i_{j_2}}$ as distinct terms in the subnet if $j_1 \neq j_2$, even if $f(j_1) = f(j_2)$. (Note that terms of the original net may be "repeated" in the subnet, although only to a limited extent since f is directed. This really causes no more complication than the fact that x_{i_1} can equal x_{i_2} in the original net even if $i_1 \neq i_2$, i.e. ϕ need not be injective.)

XIII.3.2.9. The most obvious subnets of a net (x_i) are formed by letting J be a cofinal subset () of I and f the inclusion of J into I. We will call a subnet of this kind a *subtail* (a nonstandard name; there appears to be no standard one). A *tail* of a net is a subtail where J is of the form $\{i \in I : i_0 \leq i\}$ for a fixed $i_0 \in I$. A subtail of a sequence is exactly just a subsequence, and a tail of a sequence in this sense is the same as a tail as previously defined.

For purposes of describing topologies and convergence, subtails are inadequate in general, just as in the sequence case. Thus it is necessary to use the more complicated but more general and flexible notion of subnet as we have defined it.

XIII.3.2.10. Proposition. Let X be a topological space and (x_i) a net in X. Any subnet of a subnet of (x_i) is a subnet of (x_i). Any subtail of a subtail of (x_i) is a subtail, and any tail of a tail is a tail.

Proof: If $\phi : I \to X$ defines (x_i), and $f : J \to I$ the subnet, a subnet of $(x_{f(j)})$ consists of a directed set K and a directed function $g : K \to J$, with function $(\phi \circ f) \circ g$. Then $f \circ g$ is a directed function from K to I (II.4.4.8.), and $\phi \circ (f \circ g)$ gives the subsubnet as a subnet of (x_i). The last statement is obvious. 👍

The proofs of the next two results are almost identical to the sequence versions:

XIII.3.2.11. Proposition. Let X be a topological space, and (x_i) a net in X, $x \in X$.

(i) $x_i \to x$ if and only if some tail of (x_i) converges to x.

(ii) If $x_i \to x$, then every subnet of (x_i) converges to x. In particular, every tail of (x_i) converges to x.

Proof: We give the proof of (ii), leaving the simpler proof of (i) to the reader. Suppose $(x_{f(j)})_{j\in J}$ is a subnet of (x_i), and let U be an open neighborhood of x. There is an i_0 such that $x_i \in U$ for all $i \geq i_0$. Fix $j_0 \in J$ so that $f(j) \geq i_0$ for all $j \geq j_0$. Then $x_{f(j)} \in U$ for all $j \geq j_0$, so the subnet converges to x. 👍

The net version of XIII.3.1.5. is the hard-to-state "converse" to (ii):

XIII.3.2.12. PROPOSITION. Let X be a topological space, and (x_i) a net in X, $x \in X$. If (x_i) does not converge to x, then there is a subnet of (x_i) (in fact a subtail), no subnet of which converges to x.

PROOF: If (x_i) does not converge to x, then there is an open neighborhood U of x such that $J = \{i \in I : x_i \notin U\}$ is cofinal in I. This J gives a subtail with no terms in U, so no subnet of it can converge to x. 👍

The result is often stated in the contrapositive:

XIII.3.2.13. COROLLARY. Let X be a topological space, and (x_i) a net in X, $x \in X$. If every subnet of (x_i) has a subnet which converges to x, then $x_i \to x$.

Weak Subnets

We can extend XIII.3.2.12. in an interesting way.

XIII.3.2.14. DEFINITION. Let (x_i) be a net in X with index set I defined by $\phi : I \to X$. A *weak subnet* of (x_i) is a net defined by $\phi \circ f$, where f is a not necessarily directed function from a directed set J to I.

XIII.3.2.15. Any subnet is a weak subnet, but there are many weak subnets which are not subnets: for example, f could be a constant function. Even if f is strictly increasing, the weak subnet is not necessarily a subnet since the range of f may not be cofinal in I (this cannot happen for sequences).

The proof of XIII.3.2.12. actually shows the following considerably stronger result:

XIII.3.2.16. PROPOSITION. Let X be a topological space, and (x_i) a net in X, $x \in X$. If (x_i) does not converge to x, then there is a subnet of (x_i) (in fact a subtail), no weak subnet of which converges to x.

XIII.3.2.17. COROLLARY. Let X be a topological space, and (x_i) a net in X, $x \in X$. If every subnet of (x_i) has a weak subnet which converges to x, then $x_i \to x$.

Cluster Points of Nets

XIII.3.2.18. DEFINITION. Let X be a topological space, and (x_i) a net in X, $x \in X$. Then x is a *cluster point* of (x_i) if every open neighborhood U of x contains x_i for a cofinal set of i (we say x_i is *frequently in* U).

The proof of the next Proposition is essentially identical to the proof of (). This result motivates the theory of convergence via filters ().

XIII.3.2.19. PROPOSITION. Let $(x_i)_{i\in I}$ be a net in a topological space X. For each $i \in I$, set

$$T_i = \{x_j : j \geq i\} .$$

Then $\cap_{i\in I}\bar{T}_i$ is precisely the set of cluster points of the net (x_i). In particular, the set of cluster points of a net is closed.

The set T_i is called the *tail set* of the net (x_i) corresponding to i. Thus the convergence of a net depends only weakly on the actual directed set I but only on the collection of tail sets of the net.

PROOF: If x is a cluster point of (x_i), then any neighborhood of x contains point of T_i for arbitrarily large i, hence $x \in \cap_{i\in I}\bar{T}_i$. Conversely, if $x \in \cap_{i\in I}\bar{T}_i$ and U is an open neighborhood of x, then for each i we have $T_i \cap U \neq \emptyset$, and any element of this intersection is of the form x_j for some $j \geq i$. Thus x is a cluster point of (x_i). 👍

If $x_i \to x$, or more generally if there is a subnet of (x_i) which converges to x, then x is a cluster point of (x_i). Unlike in the sequence case, the converse is true in general, showing the universal applicability of nets to convergence. The proof is a good example of how nets are used in practice, and beautifully and graphically demonstrates the utility of the subnet definition.

XIII.3.2.20. THEOREM. Let X be a topological space, (x_i) a net in X, and $x \in X$. If x is a cluster point of (x_i), then there is a subnet of (x_i) which converges to x.

PROOF: Let J be the set of all pairs (U, i), where U is an open neighborhood of x, $i \in I$, and $x_i \in U$. Put a preorder $\precsim$ on J by setting $(U_1, i_1) \precsim (U_2, i_2)$ if $U_2 \subseteq U_1$ and $i_1 \leq i_2$.

We show that $(J, \precsim)$ is directed. Suppose (U_1, i_1) and (U_2, i_2) are in J. Set $V = U_1 \cap U_2$ and let i_3 be an upper bound for $\{i_1, i_2\}$ in I. Then since x is a cluster point of (x_i), there is an $i \geq i_3$ with $x_i \in V$. Then $(V, i) \in J$, and (V, i) is an upper bound for $\{(U_1, i_1), (U_2, i_2)\}$.

Define $f : J \to I$ by $f(U, i) = i$. We show that f is directed. Let $i_0 \in I$. We must find $(U_1, i_1) \in J$ such that $(V, i) \in J$, $(U_1, i_1) \precsim (V, i)$ implies $i \geq i_0$. Just take $U_1 = X$ and $i_1 = i_0$. (Note that f is in fact order-preserving.)

Finally, we show that the subnet $(x_{(U,i)})$ converges to x. Let V be an open neighborhood of x, and fix i_0 such that $x_{i_0} \in V$. Then $(V, i_0) \in J$, and $x_{(U,i)} \in V$ whenever $(V, i_0) \precsim (U, i)$. 👍

Similarly, we get a net version of XIII.3.1.11. which is valid in complete generality. The proof shows the utility of considering directed sets which are not partially ordered.

XIII.3.2.21. THEOREM. Let A be a subset of a topological space X and $x \in X$. Then $x \in \bar{A}$ if and only if there is a net (x_i) of points of A which converges to x.

PROOF: If there is a net in A converging to x, then every neighborhood of x contains a term of the net, hence a point of A. Thus $x \in \bar{A}$. Conversely, suppose $x \in \bar{A}$. Let I be the set of all pairs (U, y), where U is an open neighborhood of x and $y \in U \cap A$. (We have $U \cap A \neq \emptyset$ for all U.) Put a preorder $\leq$ on I by $(U, y) \leq (V, z)$ if $V \subseteq U$. (Note that this is rarely a partial order.)

We show that $(I, \leq)$ is directed. If (U_1, y_1) and (U_2, y_2) are in I, set $V = U_1 \cap U_2$, and let $z \in V \cap A$. Then (V, z) is an upper bound for $\{(U_1, y_1), (U_2, y_2)\}$.

Define a net $(x_i)_{i \in I}$ by setting $x_{(U,y)} = y$. Then $x_i \in A$ for all i. If U is any open neighborhood of x, let $y \in U \cap A$, and set $i_0 = (U, y) \in I$. Then for any $(V, z) \geq (U, y)$ in I, $x_{(V,z)} = z \in V \subseteq U$, i.e. if $i \geq i_0$, $x_i \in U$. Thus $x_i \to x$. 👍

Note that unlike in the sequence case, no form of Choice is needed for this result. This again shows the versatility of the theory of nets and subnets.

Interleaving Nets

XIII.3.2.22. It is technically useful to be able to "interleaf" or "intertwine" two nets into a single net. This is easily done for sequences: if (x_n) and (y_n) are sequences in X, simply alternate the terms to form the sequence

$$(x_1, y_1, x_2, y_2, \ldots, x_n, y_n, \ldots)$$

which will converge to a limit x if and only if $x_n \to x$ and $y_n \to x$ (cf. ()).

It is unclear how to do this in the case of nets. There are two technical difficulties:

(i) The two nets need not even have the same index set.

(ii) Even if the index sets are the same, it is not obvious what "alternating" terms means.

The difficulty is in the generality of the indexing for nets, but the generality and flexibility of indexing is also the key to the solution.

XIII.3.2.23. We first show how to modify two nets to new nets with the same index set. Suppose (x_i) is a net in X with index set I, and suppose J is another directed set; we will use $\prec$ for the direction in both I and J. Set $K = I \times J$, and give K the product ordering: $(i_1, j_1) \prec (i_2, j_2)$ if and only if $i_1 \prec i_2$ and $j_1 \prec j_2$ (cf. ()). For $k = (i, j)$, set $x_k = x_i$.

XIII.3.2.24. PROPOSITION. The net (x_k) converges to $x \in X$ if and only if $x_i \to x$.

PROOF: Let U be a neighborhood of x. If $x_k \to x$, there is an $k_0 = (i_0, j_0)$ such that $x_k \in U$ whenever $k_0 \prec k$; then $x_i \in U$ whenever $i_0 \prec i$. Conversely, if $x_i \to x$, there is an i_0 such that $x_i \in U$ whenever $i_0 \prec i$; if j_0 is any fixed element of J, then $x_k \in U$ whenever $k_0 = (i_0, j_0) \prec k$. 👍

In fact, the modified net (x_k) is a subnet of (x_i) via the projection from $I \times J$ onto I.

XIII.3.2.25. If (x_i) and (y_j) are nets in X with index sets I and J respectively, both can be modified to nets with index set $K = I \times J$; convergence of the nets is unaffected by the reindexing.

XIII.3.2.26. Now suppose (x_k) and (y_k) are nets with the same index set K. Define a new net $(z_{k,n})$ with index set $K \times \{1, 2\}$ with ordering $(k_1, n_1) \prec (k_2, n_2)$ if and only if $k_1 \prec k_2$, by taking $z_{k,1} = x_k$ and $z_{k,2} = y_k$ for all k (cf. ()). The net $(z_{k,n})$ is called the *interleaving* of (x_k) and (y_k). We have:

XIII.3.2.27. PROPOSITION. The interleaved net $(z_{k,n})$ converges to $x \in X$ if and only if $x_k \to x$ and $y_k \to x$.

The simple proof is almost identical to the proof of XIII.3.2.24., and is left to the reader (Exercise ()). Note that (x_k) and (y_k) are subnets of $(z_{k,n})$ via the maps $k \mapsto (k,1)$ and $k \mapsto (k,2)$ respectively, giving one direction.

Moore-Smith Convergence

XIII.3.2.28. The theory of convergence of nets is often called the theory of *Moore-Smith convergence*, since it was originated in [**?**]. The term "net" was coined by KELLEY.

The Moore of Moore-Smith was E. H. MOORE, not R. L. MOORE. Both Moores (who were not related) were very important and influential in American mathematics. Both were first-rate mathematicians, but their importance went far beyond their own mathematical work. Both served as presidents of the American Mathematical Society and were instrumental in the development of the AMS.

E. H. MOORE was probably more responsible than any other individual for the rise of American mathematics in general, and the University of Chicago in particular, to world-class status. A considerable percentage of current American mathematicians, including the author, are mathematical descendents of his (the Mathematics Genealogy Project https://genealogy.math.ndsu.nodak.edu/ lists at least 25,722 mathematical descendents of E. H. MOORE).

R. L. MOORE, in addition to being one of the principal founders of the American school of topology, was a gifted and energetic teacher who trained a large percentage of the leading American topologists of the mid-twentieth century. He also developed the "Moore method" of pedagogy (cf. I.6.7.), which is much admired and emulated, although it is not without its detractors. Unfortunately, he was also a virulent racist.

There have been a number of other mathematicians named Moore, including the author's advisor, C. C. MOORE, who is apparently unrelated to either E. H. or R. L.

XIII.3.3. Specification of a Topology via Convergence

It is often desirable to define a topology by specifying which sequences or nets converge. This is particularly convenient when considering function spaces, where we often want to discuss the "topology of pointwise convergence," the "topology of uniform convergence," etc. This can be done rather cleanly, although there are some subtleties.

This approach to topology was developed by FRÉCHET [], ALEKSANDROV and URYSOHN [], [], and BIRKHOFF [Bir37]. See [Kur66, §20] and [Kel75] for details.

XIII.3.3.1. It is clear that a topology cannot be completely determined by convergence of sequences in general in the non-first-countable case; in fact, two distinct topologies on a set can have the same convergent sequences, such as the usual topology and the discrete topology on $\beta\mathbb{N}$ (XIII.11.9.16.(v); see also () for a natural and important example, and Exercise XIII.7.10.16. for a simple one). We have a much better chance of success working with nets, but we will use sequences as much as possible since in applications it is very convenient to use sequences.

XIII.3.3.2. The general setup is: X is a set, and $\mathcal{C}$ a collection of sequences (or nets) in X which we will declare to converge, each with a specified limit. More precisely, $\mathcal{C}$ will be a set of pairs (s, x), where s is a sequence or net in X and $x \in X$; we will declare that the sequence s converges to x. We want a topology $\mathcal{T}$ on X such that $s \to x$ in the topology $\mathcal{T}$ for every pair (s, x) in $\mathcal{C}$; ideally, we want $\mathcal{C}$ to be the *complete* set of convergent sequences (or nets) with respect to $\mathcal{T}$.

We will be primarily interested in the case where a sequence cannot have more than one limit (this was an explicit assumption in [], [], [], [Bir37]), but we will not make this a blanket assumption.

XIII.3.3.3. To make the process work, we make the modest assumption on $\mathcal{C}$ that any subsequence of a sequence in $\mathcal{C}$ is also in $\mathcal{C}$, with the same limit, i.e. that if $(s, x) \in \mathcal{C}$ and s' is a subsequence of s, then $(s', x) \in \mathcal{C}$. This subsequence property must be satisfied for $\mathcal{C}$ to have any claim to completeness. ("Sequence" and "subsequence" can be replaced by "net" and "subnet" here.)

XIII.3.3.4. There are two possible approaches to defining $\mathcal{T}$:

(1) We could declare a subset E of X *closed* if, whenever $(s, x) \in \mathcal{C}$ and s is a sequence in E, then $x \in E$. (If $\mathcal{T}$ is a topology on X under which the sequences in $\mathcal{C}$ converge, then any $\mathcal{T}$-closed set has this property.)

(2) We could define a closure operation on X by saying that if $A \subseteq X$, then $\bar{A}$ consists of all $x \in X$ such that there exists $(s, x) \in \mathcal{C}$ with s a sequence in A. (If $\mathcal{T}$ is a topology on X under which the sequences in $\mathcal{C}$ converge, then any such x must be a limit point of A.) If $A \mapsto \bar{A}$ is a Kuratowski closure operator (), it defines a suitable topology $\mathcal{T}$.

Procedure (2) is more constructive. It turns out that procedure (1) works quite nicely, but procedure (2) is more problematic.

First consider procedure (1). Half the answer is almost immediate:

XIII.3.3.5. Proposition. Let X be a set and $\mathcal{C}$ a set of sequences as above, with the subsequence property. Let $\mathcal{E}$ be the collection of closed subsets of X as defined in (1). Then $\mathcal{E}$ is the collection of closed sets in a topology $\mathcal{T}$ which is the strongest topology on X for which $s \to x$ for every $(s, x) \in \mathcal{C}$.

Proof: We need to show that $\mathcal{E}$ satisfies the conditions of XIII.2.1.4.. It is obvious that $\emptyset$ and X are in $\mathcal{E}$, and that $\mathcal{E}$ is closed under arbitrary intersections. So suppose A and B are in $\mathcal{E}$ and $(s, x) \in \mathcal{C}$ with s a sequence in $A \cup B$. Then s either has a subsequence s' in A or a subsequence s'' in B. In the first case $(s', x) \in \mathcal{C}$, so $x \in A \subseteq A \cup B$ since $A \in \mathcal{E}$, and similarly in the second case $x \in B \subseteq A \cup B$. Thus $A \cup B \in \mathcal{E}$ and $\mathcal{E}$ is closed under finite unions.
It is clear from the last observation in (1) that the topology defined by $\mathcal{E}$ is the strongest possible topology on X giving the desired convergence. 👍

XIII.3.3.6. This result and its proof hold if "sequence" is replaced by "net" and "subsequence" by "subnet" throughout (i.e. the subsequence property is replaced by the subnet property, and the definition of "closed" in (1) uses nets).

XIII.3.3.7. However, it will rarely be true that $\mathcal{C}$ is the complete set of sequences which converge with respect to $\mathcal{T}$. For one thing, for $\mathcal{C}$ to be complete it must have the following property (cf. XIII.3.1.6.):

> If s is a sequence in X, $x \in X$, and every subsequence s' of s has a subsequence s'' such that $(s'', x) \in \mathcal{C}$, then $(s, x) \in \mathcal{C}$.

$\mathcal{C}$ must also contain all constant sequences (with limit equal to the constant term).

Even these additional assumptions on $\mathcal{C}$ are not enough in general. Before discussing this further, we will examine procedure (2), beginning with the net version.

XIII.3.3.8. So, given a collection $\mathcal{C}$ of nets, let us define a closure operation on X as in (2). If $\mathcal{C}$ is to be the set of convergent nets for a topology , we must have the diagonal property:

It turns out this property along with the previous ones is sufficient to obtain a Kuratowski closure operation:

XIII.3.3.9. PROPOSITION. Let X be a set and $\mathcal{C}$ a collection of pairs (s, x) where s is a net in X and $x \in X$. Suppose $\mathcal{C}$ has the following properties:

(i) If $x \in X$ and s is a constant net with all terms x, then $(s, x) \in \mathcal{C}$.

(ii) If $(s, x) \in \mathcal{C}$ and s' is a subnet of s, then $(s', x) \in \mathcal{C}$.

(iii) Diagonal property

For $A \subseteq X$, let $\bar{A}$ be the set of all $x \in X$ such that there exists a pair $(s, x) \in \mathcal{C}$ with s a net in A. Then $A \mapsto \bar{A}$ is a Kuratowski closure operation on X. The corresponding topology on X is the strongest topology $\mathcal{T}$ on X with the property that $s \to x$ in $\mathcal{T}$ for every $(s, x) \in \mathcal{C}$, and coincides with the topology defined in XIII.3.3.5..

PROOF: Obviously $\bar{\emptyset} = \emptyset$, and $\bar{X} = X$ since $\mathcal{C}$ contains all constant nets. If $A \subseteq X$, then $A \subseteq \bar{A}$ by the same argument, and $\bar{\bar{A}}$ by (iii). Now suppose $A, B \subseteq X$, and $x \in X$. If $x \in \bar{A}$ or $x \in \bar{B}$, then obviously $x \in \overline{A \cup B}$. Conversely, if $x \in \overline{A \cup B}$, let $s = (x_i)$ be a net in $A \cup B$ with $(s, x) \in S$. Then s either has a subnet s' in A or a subnet s'' in B (). In the first case $(s', x) \in \mathcal{C}$, so $x \in \bar{A}$, and in the second case $x \in \bar{B}$. In either case, $x \in \bar{A} \cup \bar{B}$. Thus $\overline{A \cup B} = \bar{A} \cup \bar{B}$. 👍

XIII.3.3.10. From either XIII.3.3.5. or XIII.3.3.9. we are left with two questions:

(1) Is the topology $\mathcal{T}$ so defined the *unique* topology on X for which $s \to x$ for all $(s, x) \in \mathcal{C}$?

(2) Is $\mathcal{C}$ the collection of *all* convergent sequences or nets with respect to $\mathcal{T}$?

The answer to (1) is clearly no in general: we have $s \to x$ for all $(x, s) \in \mathcal{C}$ for any weaker topology too, e.g. the indiscrete topology.

For question (2) to have a positive answer, it is necessary that $\mathcal{C}$ have the net version of the property of XIII.3.3.7., as well as (i)–(iii) of XIII.3.3.9.. This suffices:

XIII.3.3.11. THEOREM. (cf. [Bir37]) Let X be a set and $\mathcal{C}$ a collection of pairs (s, x) where s is a net in X and $x \in X$. Suppose $\mathcal{C}$ has the following properties:

(i) If $x \in X$ and s is a constant net with all terms x, then $(s, x) \in \mathcal{C}$.

(ii) If $(s, x) \in \mathcal{C}$ and s' is a subnet of s, then $(s', x) \in \mathcal{C}$.

(iii) Diagonal property

(iv) If s is a net in X, $x \in X$, and every subnet s' of s has a weak subnet s'' with $(s'', x) \in \mathcal{C}$, then $(s, x) \in S$.

Let $\mathcal{T}$ be the topology defined by $\mathcal{C}$ as in XIII.3.3.7. or XIII.3.3.9.. Then $\mathcal{C}$ is precisely the set of nets in X which converge in $\mathcal{T}$.

PROOF: We need only to show that if s is a net in X with $s \to x$, then $(s, x) \in \mathcal{C}$. Suppose $(s, x) \notin \mathcal{C}$. Then by (iv) there is a subnet s' of s such that $(s'', x) \notin \mathcal{C}$ for all weak subnets s'' of s'. Set

$$E = \{y \in X | (s'', y) \in \mathcal{C} \text{ for some weak subnet } s'' \text{ of } s'\} .$$

The set E is exactly the closure of the set of terms in s', hence is a closed subset of X. Since all terms of s' are in E, we cannot have $s' \to x$ and hence s does not converge to x. 👍

XIII.3.3.12. Hypotheses (iii) and (iv) may be somewhat difficult to check in some cases, but they are indeed necessary conditions for the conclusion (),(). By a somewhat more complicated argument (cf. [Kel75, p. 74]), it can be shown that hypothesis (iv) can be slightly weakened by replacing "weak subnet" with "subnet." The proof of this in [Kel75] (implicitly) uses the AC (note that the proof of XIII.3.3.11. above does not use any Choice). A class of nets satisfying (i)–(iii) and this version of (iv) is called a *convergence class*.

XIII.3.3.13. It is suggested in [**?**, p. 19] that hypothesis (iv) is redundant. However, at least some weak version of (iv) is necessary; otherwise, we could take $\mathcal{C}$ to be the class of constant nets. The topology on X determined by this $\mathcal{C}$ is the discrete topology, but if X has more than one element $\mathcal{C}$ is not the entire class of convergent nets in X (the class of eventually constant nets).

The following weak version of (iv) might suffice:

(iv′) If s is a net in X, $x \in X$, and there is a tail s' of s with $(s', x) \in \mathcal{C}$, then $(s, x) \in \mathcal{C}$.

XIII.3.3.14. There is also a set-theoretic problem: the collection of all subnets of a net does not form a set. One might be able to get around this problem by only considering nets whose index set is a subset of some fixed set whose cardinality is large compared to $|X|$, although construction of diagonal nets may cause difficulty. One really needs a notion of equivalence of nets and that every net on X is equivalent to a net with an index set of bounded cardinality. This is effectively accomplished using the theory of filter convergence (), which avoids these set-theoretic difficulties.

XIII.3.3.15. Unfortunately, there is no simple version of XIII.3.3.11. for sequences. The sequence version of (iii) does not hold in general (). If the generated topology is assumed to be first countable, the theorem holds with sequences in place of nets; but there is no simple criterion in terms of convergence of sequences which will characterize when the generated topology is first countable.

XIII.3.4. Convergence via Filters

There is another approach to convergence in a topological space which uses filters instead of nets. This approach is somewhat more complicated and less familiar at first, but has some technical advantages once set up. The theories of convergence via nets and filters are entirely equivalent logically, and there is a way of translating directly from one to the other, and it is largely a matter of taste (and, unfortunately, sometimes of national pride, since the theory of net convergence was developed in the United States and the theory of filter convergence in France) which approach is used; but some things are done more easily or efficiently from one of the points of view instead of the other, so it is valuable to be familiar with both approaches.

The theory of convergence via filters was developed by H. CARTAN for BOURBAKI around 1940. More details can be found in [**?**] and [Dug78].

XIII.3.4.1. Recall (XVI.1.8.1.) that a *filter* on X is a nonempty collection $\mathcal{F}$ of subsets of X such that

(i) $\mathcal{F}$ is closed under finite intersections.

(ii) $\mathcal{F}$ has the *superset property*: if $A \in \mathcal{F}$ and $A \subseteq B \subseteq X$, then $B \in \mathcal{F}$.

(iii) $\emptyset \notin \mathcal{F}$.

XIII.3.4.2. DEFINITION. A collection $\mathcal{D}$ of nonempty subsets of a set X is *centered* if, for every $A, B \in \mathcal{D}$ there is a $C \in \mathcal{D}$ with $C \subseteq A \cap B$. A centered collection of subsets of X is called a *filterbase* in X.

XIII.3.4.3. A collection of nonempty subsets of a set X is centered if and only if it is directed by inverse inclusion, i.e. $A \leq B$ if $B \subseteq A$. A filterbase $\mathcal{D}$ has the finite intersection property (XVI.1.8.9.), and hence is contained in a unique smallest filter $\mathcal{F}$ (XVI.1.8.10.) called the *filter generated by* $\mathcal{D}$; $\mathcal{D}$ is called a *filterbase for* $\mathcal{F}$. In fact, $\mathcal{F}$ consists precisely of all subsets of X which contain a subset in $\mathcal{D}$.

Here are the most important examples of filterbases for topology:

XIII.3.4.4. EXAMPLES. (i) Let $(X, \mathcal{T})$ be a topological space and $p \in X$. The set of all $\mathcal{T}$-neighborhoods of p is a filter, called the *neighborhood filter*, or *$\mathcal{T}$-neighborhood filter*, of p in X, denoted $\mathcal{N}_p$. The collection of open neighborhoods of p is a filterbase for $\mathcal{N}_p$; more generally, any local base for $\mathcal{T}$ at p is a filterbase for $\mathcal{N}_p$.
(ii) Let X be a set, and $(x_i)_{i \in I}$ a net in X. The tail set $\{T_i : i \in I\}$ (XIII.3.2.19.) is a filterbase in X.

XIII.3.4.5. Example XIII.3.4.4.(ii) is actually universal: if $\mathcal{D}$ is a filterbase in X, let I be the set of pairs (A, x), where $A \in \mathcal{D}$ and $x \in A$. Set $(A, x) \leq (B, y)$ if $B \subseteq A$. Then I is a directed set since $\mathcal{D}$ is centered. If $x_{(A,z)} = z$, then $(x_{(A,z)})$ is a net in X with index set I, with $T_{(A,z)} = A$ for each (A, z). There is thus a one-one correspondence between filterbases in X and equivalence classes of nets in X, where two nets are regarded as equivalent if their corresponding filterbases are identical (see Exercise XIII.3.5.3.; cf. XIII.3.3.14.).

Convergence of Filters

The definition of convergence of a filter is very simple:

XIII.3.4.6. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. A filter $\mathcal{F}$ in X *converges* to a point $p \in X$ if $\mathcal{N}_p \subseteq \mathcal{F}$.

XIII.3.4.7. Because of the superset property of filters, a filter $\mathcal{F}$ converges to a point p if and only if, whenever U is a neighborhood of p, there is an $A \in \mathcal{F}$ with $A \subseteq U$. Thus we extend the notion of filter convergence to filterbases:

XIII.3.4.8. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. A filterbase $\mathcal{D}$ in X *converges* to a point $p \in X$ if whenever U is a neighborhood of p, there is an $A \in \mathcal{D}$ with $A \subseteq U$.

Thus a filterbase converges to p if and only if the filter it generates converges to p.

XIII.3.4.9. This is consistent with convergence of nets: a net $(x_i)_{i \in I}$ converges to p if and only if the corresponding filterbase $\{T_i : i \in I\}$ converges to p.

Closure via Filters

We can give a characterization of the closure of a set using filters. It is cleanest to express it in terms of filterbases, since if A is a subset of X and $\mathcal{D}$ is a filterbase in A, $\mathcal{D}$ is also a filterbase in X (this is false for filters).

XIII.3.4.10. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space and $A \subseteq X$. The $\mathcal{T}$-closure of A in X consists precisely of all points x with the property that there is a filterbase $\mathcal{D}$ in A converging to x. In particular, A is $\mathcal{T}$-closed if and only if, whenever $\mathcal{D}$ is a filterbase in A converging to $x \in X$, then $x \in A$.

XIII.3.4.11. To state the result using filters, we make the following definition. We say a filter $\mathcal{F}$ in a set X is *eventually in* $A \subseteq X$ if $A \in \mathcal{F}$ (or, equivalently, some subset of A is in $\mathcal{F}$). We can then state:

XIII.3.4.12. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space and $A \subseteq X$. The $\mathcal{T}$-closure of A in X consists precisely of all points x with the property that there is a filter $\mathcal{F}$ which is eventually in A and converges to x. In particular, A is $\mathcal{T}$-closed if and only if, whenever $\mathcal{F}$ is a filter in X which is eventually in A and converges to $x \in X$, then $x \in A$.

XIII.3.4.13. Similarly, we say a filterbase $\mathcal{D}$ in X is *eventually in* $A \subseteq X$ if there is a $D \in \mathcal{D}$ with $D \subseteq A$.

Cluster Points of Filters

XIII.3.4.14. DEFINITION. Let $\mathcal{D}$ be a filterbase in a set X and $A \subseteq X$. We say $\mathcal{D}$ is *frequently in* A if $A \cap D \neq \emptyset$ for all $D \in \mathcal{D}$. A point x is a *cluster point* for a filterbase $\mathcal{D}$ in X if $\mathcal{D}$ is frequently in every neighborhood of x.

This terminology is consistent with the corresponding definition for nets (). There is no modification needed to replace "filterbase" by "filter" (in fact, the statement for filters is a special case since every filter is a filterbase).

XIII.3.4.15. If a filter or filterbase converges to x, then x is a cluster point. We have the following filter version of XIII.3.2.19.:

XIII.3.4.16. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space, and let $\mathcal{D}$ be a filterbase in X. A point $x \in X$ is a cluster point of $\mathcal{D}$ if and only if $x \in \cap_{D \in \mathcal{D}} \bar{D}$.

Subfilters

What is the right filter or filterbase analog of a subnet? The following definition is reasonable:

XIII.3.4.17. DEFINITION. Let $\mathcal{D}$ and $\mathcal{D}'$ be filterbases in a set X. Then $\mathcal{D}'$ is *subordinate* to $\mathcal{D}$, written $\mathcal{D}' \vdash \mathcal{D}$, if for every $A \in \mathcal{D}$ there is a $B \in \mathcal{D}'$ with $B \subseteq A$.

XIII.3.4.18. It is easily seen that if $\mathcal{F}$ and $\mathcal{F}'$ are the filters generated by $\mathcal{D}$ and $\mathcal{D}'$ respectively, then $\mathcal{D}' \vdash \mathcal{D}$ if and only if $\mathcal{F} \subseteq \mathcal{F}'$, i.e. the subordinate filter is *larger*. If $\mathcal{F} \subseteq \mathcal{F}'$, we say the filter $\mathcal{F}'$ is *finer* than $\mathcal{F}$. Thus it is not appropriate to call $\mathcal{D}'$ a "subfilterbase" of $\mathcal{D}$ (which is a mouthful anyway.) However, if $(x_i)_{i \in I}$ is a net in X and $(y_j)_{j \in J}$ is a subnet, then the tail filterbase of (y_j) is subordinate to the tail filterbase of (x_i) (the converse is false).

We have:

XIII.3.4.19. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space, and let $\mathcal{D}$, $\mathcal{D}'$ be filterbases in X, with $\mathcal{D}' \vdash \mathcal{D}$.

(i) If $\mathcal{D}$ converges to $x \in X$, then $\mathcal{D}'$ converges to x.

(ii) If $x \in X$ is a cluster point of $\mathcal{D}'$, it is a cluster point of $\mathcal{D}$.

We next obtain a generalization of XIII.3.2.20.:

XIII.3.4.20. PROPOSITION. Let $\mathcal{D}$ be a filterbase in a set X, and $(x_i)_{i \in I}$ a net in X which is frequently in D for every $D \in \mathcal{D}$. Then there is a subnet of (x_i) which is eventually in every $D \in \mathcal{D}$.

PROOF: Let J be the set of all pairs (D, i) where $D \in \mathcal{D}$, $i \in I$, and $x_i \in D$. Set $(D_1, i_1) \leq (D_2, i_2)$ if $D_2 \subseteq D_1$ and $i_1 \leq i_2$. Then J is directed: if $(D_1, i_1), (D_2, i_2) \in J$, there is a $D_3 \in \mathcal{D}$, $D_3 \subseteq D_1 \cap D_2$, and an $i_0 \in I$ with $i_1, i_2 \leq i_0$, and an $i_3 \geq i_0$ such that $x_{i_3} \in D_3$ since (x_i) is frequently in D_3, so $(D_3, i_3) \in J$ and $(D_1, i_1), (D_2, i_2) \leq (D_3, i_3)$. Define $f : J \to I$ by $f(D, i) = i$. Then f is directed: if $i_0 \in I$, fix $D_1 \in \mathcal{D}$ and $i_1 \geq i_0$ with $x_{i_1} \in D_1$; then $f(D, i) \geq i_0$ whenever $(D, i) \geq (D_1, i_1)$. If $\phi : I \to X$ gives $\phi(i) = x_i$, then $\phi \circ f$ defines a subnet $(x_{f(j)})$. If $D_0 \in \mathcal{D}$, then there is an $i_0 \in I$ with $(D_0, i_0) \in J$. If $(D_0, i_0) \leq (D, i) = j \in J$, then $x_{f(j)} \in D \subseteq D_0$, so the subnet $(x_{f(j)})$ is eventually in D_0. 👍

Rephrasing, we get:

XIII.3.4.21. Corollary. Let $\mathcal{D}$ be a filterbase in a set X, and $(x_i)_{i\in I}$ a net in X. If the tail set filterbase $\mathcal{D}'$ of (x_i) satisfies $\mathcal{D} \vdash \mathcal{D}'$, then there is a subnet of (x_i) whose tail set filterbase $\mathcal{D}''$ satisfies $\mathcal{D}'' \vdash \mathcal{D}$.

We also have the following analog of XIII.3.2.12.:

XIII.3.4.22. Proposition. Let $(X, \mathcal{T})$ be a topological space, $\mathcal{D}$ a filterbase in X, and $x \in X$. If $\mathcal{D}$ does not converge to x, then there is a filterbase $\mathcal{D}' \vdash \mathcal{D}$ such that no filterbase $\mathcal{D}'' \vdash \mathcal{D}'$ converges to x.

Ultrafilters and Maximal Filterbases

XIII.3.4.23. By XIII.3.4.18., larger means "finer" for filters. The largest filters are ultrafilters (XVI.1.8.1.), which are therefore the "finest" filters. By XVI.1.8.15. (which uses AC), every filter is contained in an ultrafilter. By XVI.1.8.12., a filter $\mathcal{F}$ in a set X is an ultrafilter if and only if for every $A \subseteq X$, either $A \in \mathcal{F}$ or $A^c \in \mathcal{F}$, i.e. $\mathcal{F}$ is either eventually in A or eventually in A^c.

XIII.3.4.24. A filterbase which generates an ultrafilter is called a *maximal filterbase*; "ultrafilterbase" would be a better name but too cumbersome ("maximal filterbase" is an abuse of language since a maximal filterbase is not maximal among all filterbases under inclusion unless it is already an ultrafilter; but see Exercise XIII.3.5.5.). A filterbase $\mathcal{D}$ in X is maximal if and only if for each $A \subseteq X$, there is a $D \in \mathcal{D}$ such that either $D \subseteq A$ or $D \subseteq A^c$.

XIII.3.4.25. Proposition. Let $(X, \mathcal{T})$ be a topological space, $\mathcal{D}$ a maximal filterbase in X, and $x \in X$. If x is a cluster point of $\mathcal{D}$, then $\mathcal{D}$ converges to x.

Proof: Let U be a neighborhood of x. Then there is a $D \in \mathcal{D}$ with either $D \subseteq U$ or $D \subseteq U^c$. But $D \cap U \neq \emptyset$ since x is a cluster point of $\mathcal{D}$, so $D \subseteq U^c$ is impossible and $D \subseteq U$. 👍

XIII.3.4.26. Proposition. Let X and Y be sets, $f : X \to Y$ any function, and $\mathcal{D}$ a maximal filterbase in X. Then $f(\mathcal{D}) = \{f(D) : D \in \mathcal{D}\}$ is a maximal filterbase in Y.

Proof: It is clear that $f(\mathcal{D})$ is a filterbase in Y. If $B \subseteq Y$ and $A = f^{-1}(B)$, then $A^c = f^{-1}(B^c)$. Since $\mathcal{D}$ is maximal, there is a $D \in \mathcal{D}$ such that either $D \subseteq f^{-1}(B)$ or $D \subseteq f^{-1}(B^c)$, hence either $f(D) \subseteq B$ or $f(D) \subseteq B^c$. Thus $f(\mathcal{D})$ is maximal. 👍

Universal Nets

XIII.3.4.27. Definition. A net (x_i) in a set X is a *universal net* if its tail set filterbase is maximal.

The simple verification of the next proposition is left to the reader.

XIII.3.4.28. PROPOSITION. A net (x_i) in a set X is universal if and only if, for every subset A of X, either (x_i) is eventually in A or (x_i) is eventually in A^c.

XIII.3.4.29. The existence of universal nets is an immediate consequence of the existence of ultrafilters. But we can do better:

XIII.3.4.30. THEOREM. Let X be a set, and (x_i) a net in X. Then (x_i) has a universal subnet.

PROOF: Let $\mathcal{D}$ be the tail set filterbase corresponding to (x_i). Let $\mathcal{F}$ be an ultrafilter containing $\mathcal{D}$ (i.e. containing the filter generated by $\mathcal{D}$). Since $\mathcal{D} \subseteq \mathcal{F}$, if $F \in \mathcal{F}$ we have $F \cap D \neq \emptyset$ for all $D \in \mathcal{D}$, and it follows that (x_i) is frequently in $\mathcal{F}$. Since $\mathcal{F}$ is itself a filterbase, by XIII.3.4.20. there is a subnet $(x_{f(j)})$ of (x_i) which is eventually in F for any $F \in \mathcal{F}$. If $A \subseteq X$, then either $A \in \mathcal{F}$ or $A^c \in \mathcal{F}$, so $(x_{f(j)})$ is either eventually in A or eventually in A^c. Thus $(x_{f(j)})$ is a universal net. 👍

This result requires some version of Choice, and is not a theorem of ZF. In fact, it is equivalent to the fact that every filter on a set is contained in an ultrafilter, hence to BPI ().

XIII.3.5. Exercises

XIII.3.5.1. Let X be a topological space.

(a) Let (x_n) be a sequence in X. If (y_n) is a rearrangement of (x_n), i.e. $y_n = x_{\phi(n)}$ for a bijection (permutation) $\phi : \mathbb{N} \to \mathbb{N}$, show that (y_n) converges if and only if (x_n) converges, and $\lim_{n\to\infty} y_n = \lim_{n\to\infty} x_n$ (if X is not Hausdorff, then (x_n) and (y_n) converge to the same limits).

(b) Show that the analogous statement for nets is false in general.

XIII.3.5.2. Let X be a topological space. Make the following definitions:

(i) A subset A of X is *sequentially closed* if, whenever (x_n) is a sequence in A with $x_n \to x \in X$, then $x \in A$.

(ii) If A is a subset of X, then the *sequential closure* of A is the set of all $x \in X$ for which there is a sequence (x_n) in A with $x_n \to x$.

These definitions, from [Fra65], are standard, but somewhat inconsistent: the sequential closure of a set is not sequentially closed in general. Sequential closure can be applied repeatedly, even transfinitely. Every closed set is sequentially closed, and the sequential closure of a set is contained in its closure. ✋

(iii) The space X is *sequential* if every sequentially closed subset of X is closed.

(iv) The space X is a *Fréchet-Urysohn space* if the sequential closure of every subset of X equals its closure (i.e. is closed).

(a) Show that every first countable topological space is Fréchet-Urysohn, and every Fréchet-Urysohn space is sequential. The simplest example of a Fréchet-Urysohn space which is not first countable is the one-point compactification of an uncountable discrete space (so even a compact Hausdorff Fréchet-Urysohn space need

not be first countable). See XIII.9.5.7. for another example, and [Eng89, 1.6.19] for a sequential space which is not Fréchet-Urysohn.

(b) Show that the space in XIII.11.14.10. is not sequential, even though every singleton subset is a G_δ. (In fact, "most" topological spaces which are not first countable fail to be sequential.)

(c) Show that every quotient of a sequential space is sequential. In particular, every quotient of a first countable space is sequential.

(d) Show that every sequential Hausdorff space is a quotient of a locally compact metrizable space. [Consider the separated union of the spaces consisting of convergent sequences in the space and their limits, i.e. a separated union of copies of $\mathbb{N}^\dagger$ and singletons.] This is a strong converse to (c).

(e) Show that a topological space X is sequential if and only if the sequential criterion for continuity holds for functions from X, i.e. if Y is any topological space and $f : X \to Y$ is a function, then f is continuous if and only if, whenever $x_n \to x$ in X, we have $f(x_n) \to f(x)$.

(f) If X is a Fréchet-Urysohn space, (x_n) is a sequence in X, and $x \in X$, then x is a cluster point of (x_n) if and only if there is a subsequence of (x_n) converging to X. Show that this property characterizes Fréchet-Urysohn spaces among sequential spaces (but not among all topological spaces).

(g) Show that a topological space X is Fréchet-Urysohn if and only if every subspace of X is sequential in the relative topology.

For more on these spaces, see [Eng89] (where Fréchet-Urysohn spaces are called *Fréchet spaces*). Both notions (sequential and Fréchet-Urysohn) are reasonable precise versions of the condition "the topology of X is completely determined by convergence of sequences."

XIII.3.5.3. Let $(x_i)_{i\in I}$ and $(y_j)_{j\in J}$ be two nets in a set X. Here are various possible notions of equivalence for the nets. Show each is an equivalence relation and compare them.

(a) The nets generate the same filterbase (XIII.3.4.5.).

(b) The corresponding filterbases generate the same filter.

(c) Each net is a subnet of the other.

(d) Every subnet of each net has a subnet which is also a subnet of the other net.

[To compare (a) with (c) and (d), consider constant nets with index sets $\mathbb{N}$ and the first uncountable ordinal.]

XIII.3.5.4. In what sense is XIII.3.4.20. a generalization of XIII.3.2.20.?

XIII.3.5.5. Let X be a set, and $\mathcal{D}$, $\mathcal{D}'$ filterbases in X.

(a) Show that we can have $\mathcal{D}' \vdash \mathcal{D}$ and $\mathcal{D} \vdash \mathcal{D}'$ even if $\mathcal{D} \neq \mathcal{D}'$. Thus $\vdash$ is only a preorder and not a partial order on filterbases.

(b) If $\mathcal{D}$ is a maximal filterbase and $\mathcal{D}' \vdash \mathcal{D}$, then also $\mathcal{D} \vdash \mathcal{D}'$, and any $\mathcal{D}$ with this property is maximal.

XIII.4. Continuous Functions

XIII.4.1. Definition and Properties

XIII.4.2. Topological Continuity vs. Metric Continuity

XIII.5. Comparison of Topologies

XIII.5.1. Stronger and Weaker Topologies

XIII.5.1.1. DEFINITION. Let $\mathcal{T}$ and $\mathcal{S}$ be topologies on the same set X. Then $\mathcal{S}$ is *stronger* (*larger, finer*) than $\mathcal{T}$, and $\mathcal{T}$ is *weaker* (*smaller, coarser*) than $\mathcal{S}$, if $\mathcal{T} \subseteq \mathcal{S}$, i.e. if every $\mathcal{T}$-open set is also $\mathcal{S}$-open.

The philosophy behind the names reflects the intuitive idea that a topology gives a notion of "closeness" for points, and a stronger or finer topology gives a more discriminating notion of "closeness" than a weaker or coarser one, i.e. makes it easier to distinguish between points using open sets.

A topology is stronger than itself, and also weaker than itself. In fact, a common and useful way to prove that two potentially different topologies are the same is to show that each is stronger than the other. We sometimes use the terms *strictly stronger*, *strictly weaker*, etc., when we mean to specify that the topologies are distinct.

XIII.5.1.2. It can happen, and often does, that if $\mathcal{T}$ and $\mathcal{S}$ are topologies on a set X, then neither topology is stronger than the other. In this case the topologies are said to be *incomparable*.

XIII.5.1.3. EXAMPLES. (i) On every set X, the discrete topology is the strongest topology: every topology on X is weaker than the discrete topology. Similarly, every topology on X is stronger than the indiscrete topology, so the indiscrete topology on X is the weakest topology on X.

(ii) The finite complement topology (XIII.1.2.8.(iv)) on a set X is the weakest T_1 topology () on X. For in any T_1 topology on X, the singleton subsets are closed, hence finite sets are closed, so complements of finite subsets must be open. In particular, the finite complement topology on $\mathbb{R}$ is weaker than the ordinary topology.

(iii) Let $X = \{a, b\}$, and let $\mathcal{T} = \{\emptyset, \{a\}, X\}$, $\mathcal{S} = \{\emptyset, \{b\}, X\}$. Then $\mathcal{T}$ and $\mathcal{S}$ are incomparable.

(iv) On $C([0,1])$, the topology of pointwise convergence () is strictly weaker than the topology of uniform convergence (). The topology of mean convergence () is strictly weaker than uniform convergence but incomparable with pointwise convergence. In fact, for every p, $0 < p \leq \infty$, there is an L^p topology on $C([0,1])$ (), with $p = 1$ giving mean convergence and $p = \infty$ uniform convergence; the L^p topology is weaker than the L^r topology if and only if $p \leq r$ (). So there is an uncountable family of natural topologies on $C([0,1])$, all distinct and all comparable; only the strongest of these (uniform convergence) is comparable with pointwise convergence. (There are very many other topologies on $C([0,1])$ too, most of which are incomparable to these.)

(v) On $C_c(\mathbb{R})$, the set of real-valued continuous functions on $\mathbb{R}$ of compact support, i.e. zero outside a finite interval, there is also an L^p topology for $0 < p \leq \infty$ (). In this case no two of the topologies are comparable ().

XIII.5.1.4. When the topology is changed, all the ancillary things associated with a topological space potentially change. If the topology on a space X is made stronger, i.e. with more open sets, then

> Since closed sets are complements of open sets, there are *more* closed sets. (Recall that being closed is not the opposite of being open!) Thus it is *easier* for a subset of X to be closed.

It is *harder* for a point x to be a limit point of a subset E, since there are *more* open neighborhoods of x which must have nonempty intersection with E. Thus E has *fewer* limit points and the closure of E is *smaller*.

Similarly, it is *harder* for a point x to be the limit of a sequence or net (x_i), since there are *more* open neighborhoods of x which must eventually contain the sequence or net. Thus there are *fewer* convergent sequences and nets. Similarly, a sequence or net has *fewer* cluster points.

It is *harder* for a subset F of X to be compact, since there are *more* open covers of F which must each have finite subcovers.

It is *easier* for X to be T_0, T_1, or Hausdorff, since there are *more* open sets to choose from.

However, it is unclear whether regularity or normality becomes easier when the topology becomes stronger, since on the one hand for a given point p and closed set F it is easier to find disjoint neighborhoods, but on the other hand there are more pairs (p, F) to consider. Similarly, normality does not obviously become easier. In fact, regularity and normality are not preserved under making the topology stronger or weaker (cf. XIII.7.10.17.; for the other direction, consider weakening the discrete topology). In a similar way, one might expect paracompactness, like compactness, to be preserved when weakening a topology, since there are *fewer* open covers to consider; but on the other hand an open cover has *fewer* refinements. Paracompactness is in fact not preserved under either strengthening or weakening the topology (XIII.7.10.17. and weakening the discrete topology are again counterexamples).

It is *harder* for a subset of X to be connected, since there are more potential separations.

If Y is another space and $f : X \to Y$, and the topology on Y is kept fixed and the topology on X strengthened, it is *easier* for f to be continuous since it is easier for $f^{-1}(V)$ to be open for any given open set $V \subseteq Y$. However, if the topology on X is kept fixed and the topology on Y is strengthened, it is *harder* for f to be continuous since the preimages of more open sets must be considered. If the topologies on both X and Y are strengthened, it may be neither easier nor harder for f to be continuous. In particular, if the topology on X is strengthened, functions from X to X may change from continuous to discontinuous or vice versa; the set of continuous functions from X to X will normally be different, although neither larger nor smaller (cf. Exercise XIII.7.10.15.).

Of course, there are analogous statements if the topology on X is weakened.

XIII.5.1.5. The following chart summarizes the main features of stronger and weaker topologies:

Stronger (larger, finer) topology	Weaker (smaller, coarser) topology
More open sets	Fewer open sets
More closed sets	Fewer closed sets
Fewer limit points	More limit points
Smaller closures	Larger closures
Fewer compact subsets	More compact subsets
Fewer connected subsets	More connected subsets
Fewer convergent sequences (nets)	More convergent sequences (nets)
More continuous functions out	Fewer continuous functions out
Fewer continuous functions in	More continuous functions in

The easiest way to remember what happens under strengthening or weakening of the topology is to think of the extreme cases of the discrete or indiscrete topology. (In fact, one good way to analyze many mathematical situations and guess what happens is to pass to the most extreme case.)

XIII.5.1.6. Our use of "stronger" and "weaker" in comparing topologies is the standard one, especially in analysis, particularly functional analysis (weak and weak-* topologies (), strong and weak operator topologies ()). However, it should be noted that a few topology references use these terms in the opposite sense.

XIII.5.1.7. Another sometimes convenient way to describe stronger and weaker topologies is: if $\mathcal{S}$ and $\mathcal{T}$ are topologies on a set X, then $\mathcal{S}$ is stronger than $\mathcal{T}$ if and only if the identity map from $(X, \mathcal{S})$ to $(X, \mathcal{T})$ is continuous.

Strongest and Weakest Topologies Defined by a Family of Functions

XIII.5.1.8. If X is a set, and $(Y, \mathcal{S})$ is a topological space, and $f : X \to Y$ is a function, then there is a weakest topology on X such that f is continuous. In fact,

$$\mathcal{T} = \{f^{-1}(V) : V \in \mathcal{S}\}$$

is a topology on X, as is easily checked since preimages respect unions and intersections, and $\mathcal{T}$ is obviously the weakest topology on X making f continuous.

XIII.5.1.9. More generally, if X is a set, $\{(Y_j, \mathcal{S}_j) : j \in J\}$ is a family of topological spaces (not necessarily distinct), and $f_j : X \to Y_j$ is a function for each $j \in J$, then there is a weakest topology on X making all the f_j continuous. The set

$$\{f_j^{-1}(V) : j \in J,\ V \in \mathcal{S}_j\}$$

is not usually a topology on X; however, it is a subbase for a topology $\mathcal{T}$ which is obviously the weakest topology on X making all the f_j continuous.

XIII.5.1.10. The topology $\mathcal{T}$ can also be characterized as the intersection of all topologies on X making each f_j continuous (there is always at least one such topology, the discrete topology). However, it is usually better to describe $\mathcal{T}$ constructively as in XIII.5.1.9..

XIII.5.1.11. The topology $\mathcal{T}$ can be described in terms of convergence: a net (x_i) in X converges to x in $(X, \mathcal{T})$ if and only if $f_j(x_i) \to f_j(x)$ in $(Y_j, \mathcal{S}_j)$ for every $j \in J$. Thus $\mathcal{T}$ is often called the *weak topology* on X defined by the f_j.

XIII.5.1.12. One of the most important examples of a weak topology defined by a family of functions is the product topology (), i.e. the case where $X = \prod_j Y_j$ and the f_j are the coordinate projections. The general situation can also be reduced to the case of one target space using the product topology: if X is a set, $\{(Y_j, \mathcal{S}_j) : j \in J\}$ is a family of topological spaces (not necessarily distinct), and $f_j : X \to Y_j$ is a function for each $j \in J$, let $Y = \prod_j Y_j$ and $f : X \to Y$ the induced map, i.e. $(f(x))_j = f_j(x)$ for all j. Give Y the product topology. Then the weak topology on X from the f_j is the weak topology on X from f ().

XIII.5.1.13. If all the Y_j are Hausdorff and the f_j separate points of X, then the weak topology on X defined by the f_j is also Hausdorff: if $p, q \in X$, $p \neq q$, choose j with $f_j(p) \neq f_j(q)$, and choose disjoint neighborhoods U and V of $f_j(p)$ and $f_j(q)$ in Y_j; then $f_j^{-1}(U)$ and $f_j^{-1}(V)$ are disjoint neighborhoods of p and q. The same statement is true if all the Y_j are regular or completely regular ().

XIII.5.1.14. Similarly, if Y is a set, $\{(X_j, \mathcal{T}_j) : j \in J\}$ is a set of topological spaces, and $f_j : X_j \to Y$ is a function for each j, then there is a strongest topology on Y making all the f_j continuous:

$$\mathcal{S} = \{V \subseteq Y : f_j^{-1}(V) \in \mathcal{T}_j \text{ for all } j\} .$$

It is easy to verify that $\mathcal{S}$ is a topology on Y, and it is obviously the strongest one making all the f_j continuous. However, this topology is often very weak (frequently even the indiscrete topology). The quotient topology (XIII.8.1.1.) is an example of this construction, probably the most important example.

The Lattice of Topologies

Fix a set X. We consider all topologies on X, partially ordered by containment.

XIII.5.1.15. If $\{\mathcal{T}_i : i \in I\}$ is a set of topologies on X, then $\bigcap_{i\in I} \mathcal{T}_i$ is a topology on X, which is the strongest topology on X weaker than all the $\mathcal{T}_i$. Denote this topology by $\bigwedge_{i\in I} \mathcal{T}_i$. If $I = \{1, \ldots, n\}$, write $\mathcal{T}_1 \wedge \cdots \wedge \mathcal{T}_n$ for $\bigwedge_{i\in I} \mathcal{T}_i$.

Define a family $(X_i, \mathcal{T}_i)$ of topological spaces, where $X_i = X$ for all i. For each i let f_i be the identity map on X, regarded as a map from X_i to X. Then $\bigwedge_{i\in I} \mathcal{T}_i$ is the strongest topology on X making all the f_i continuous.

XIII.5.1.16. If $\{\mathcal{T}_i : i \in I\}$ is a set of topologies on X, then $\bigcup_{i\in I} \mathcal{T}_i$ is not usually a topology on X. But it is a subbase for a topology on X, which is the weakest topology on X stronger than all the $\mathcal{T}_i$. Denote this topology by $\bigvee_{i\in I} \mathcal{T}_i$. If $I = \{1, \ldots, n\}$, write $\mathcal{T}_1 \vee \cdots \vee \mathcal{T}_n$ for $\bigvee_{i\in I} \mathcal{T}_i$. If $\mathcal{S}_i$ is a subbase for $\mathcal{T}_i$ for each i, then $\bigcup_{i\in I} \mathcal{S}_i$ is a subbase for $\bigvee_{i\in I} \mathcal{T}_i$.

Define a family $(X_i, \mathcal{T}_i)$ of topological spaces, where $X_i = X$ for all i. For each i let g_i be the identity map on X, regarded as a map from X to X_i. Then $\bigvee_{i\in I} \mathcal{T}_i$ is the weakest topology on X making all the g_i continuous.

XIII.5.1.17. Thus the set of all topologies on X is a complete lattice under containment. There is a largest element, the discrete topology, and a smallest element, the indiscrete topology. It turns out that this lattice is complemented: for every topology $\mathcal{T}$ there is a unique topology $\mathcal{T}'$ such that $\mathcal{T} \vee \mathcal{T}'$ is the discrete topology and $\mathcal{T} \wedge \mathcal{T}'$ the indiscrete topology. This was not proved until 1966 ([Ste66], cf. [vR68]).

XIII.6. Products and the Product Topology

One of the most important constructions in topology is the topological product.

XIII.6.1. Finite Products

We begin with the product of two topological spaces, which by iteration gives a product for finitely many.

XIII.6.1.1. Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be topological spaces. If $U \in \mathcal{T}$ and $V \in \mathcal{S}$, then $U \times V \subseteq X \times Y$ should be an open set in $X \times Y$ in any reasonable compatible topology; such a subset of $X \times Y$ is called an *open rectangle*. In particular, all sets of the form $U \times Y$ or $X \times V$ ($U \in \mathcal{T}$, $V \in \mathcal{S}$) are open rectangles.

"Compatible topology" is not well defined, but one reasonable requirement is that the coordinate projections $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ should be continuous. Since if $U \subseteq X$ and $V \subseteq Y$, then $\pi_X^{-1}(U) = U \times Y$ and $\pi_Y^{-1}(V) = X \times V$, so for π_X to be continuous $U \times X$ must be open in $X \times Y$ for every $U \in \mathcal{T}$, and similarly $X \times V$ must be open in $X \times Y$ for every $V \in \mathcal{S}$; since $U \times V = (U \times Y) \cap (X \times V)$, every open rectangle must be open in $X \times Y$ (and conversely).

XIII.6.1.2. Definition. Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be topological spaces. The *product topology* $\mathcal{T} \times \mathcal{S}$ is the set of all unions of open rectangles in $X \times Y$.

XIII.6.1.3. Since the intersection of two open rectangles is another open rectangle, it is easily checked that $\mathcal{T} \times \mathcal{S}$ is a topology, and is the smallest topology on $X \times Y$ making π_X and π_Y continuous.

Note that the union of two open rectangles is rarely an open rectangle, however; thus $\mathcal{T} \times \mathcal{S}$ consists of far more than just the open rectangles, or even finite unions of open rectangles, in general. We should really write $\mathcal{T} \otimes \mathcal{S}$ for the product topology in analogy with () and (); however, the notation $\mathcal{T} \times \mathcal{S}$ is standard.

The set of open rectangles does form a base for the topology $\mathcal{T} \times \mathcal{S}$, and thus open rectangles are often called *basic open sets*. More generally, we have:

XIII.6.1.4. Proposition. Let $\mathcal{B}$ be a base for the topology $\mathcal{T}$ on X, and $\mathcal{C}$ a base for the topology $\mathcal{S}$ on Y. Then

$$\mathcal{B} \times \mathcal{C} = \{B \times C : B \in \mathcal{B}, C \in \mathcal{C}\}$$

is a base for the topology $\mathcal{T} \times \mathcal{S}$ on $X \times Y$.

Proof: Let $(p, q) \in X \times Y$ and $W \in \mathcal{T} \times \mathcal{S}$ with $(p, q) \in W$. Then there are $U \in \mathcal{T}$ and $V \in \mathcal{S}$ with $p \in U$, $q \in V$, and $U \times V \subseteq W$. There is a $B \in \mathcal{B}$ with $p \in B \subseteq U$, and $C \in \mathcal{C}$ with $q \in C \subseteq V$. Then

$$(p, q) \in B \times C \subseteq U \times V \subseteq W \ .$$

XIII.6.1.5. EXAMPLES. (i) Let X and Y be sets, and $\mathcal{T} = \mathcal{P}(X)$ and $\mathcal{S} = \mathcal{P}(Y)$ be the discrete topologies on X and Y. Then the singleton subsets of X and Y form a base for the discrete topologies on X and Y. The product of two singleton sets is a singleton set and an open rectangle. Every subset of $X \times Y$ is a union of singletons, hence in $\mathcal{T} \times \mathcal{S}$; thus $\mathcal{T} \times \mathcal{S}$ is the discrete topology on $X \times Y$.

This shows the problem with using the notation $\mathcal{T} \times \mathcal{S}$ for the product topology: we get $\mathcal{P}(X) \times \mathcal{P}(Y) = \mathcal{P}(X \times Y)$ in this notation, but this equation only holds in usual set-theoretic notation if X or Y is itself a singleton.

(ii) Let $\mathcal{T}$ be the usual topology on $\mathbb{R}$, i.e. the set of unions of open intervals. Then the product topology $\mathcal{T}^{(2)} = \mathcal{T} \times \mathcal{T}$ on $\mathbb{R}^2$ consists of all unions of open rectangles (in the usual geometric sense) with sides parallel to the coordinate axes. Every open set in $\mathbb{R}^2$ in the usual sense is a union of open rectangles, although "most" open sets in $\mathbb{R}^2$ (e.g. an open disk) are not open rectangles or even finite unions of open rectangles. Conversely, any union of open rectangles is an open set in the usual sense. Thus $\mathcal{T}^{(2)}$ is the usual topology on $\mathbb{R}^2$, defined by the usual metric.

More generally, we have:

XIII.6.1.6. PROPOSITION. Let (X, ρ) and (Y, σ) be metric spaces, and $\mathcal{T}$ and $\mathcal{S}$ the topologies defined by ρ and σ respectively. Then $\mathcal{T} \times \mathcal{S}$ is the topology defined by any of the metrics $\rho \times_1 \sigma$, $\rho \times_2 \sigma$, $\rho \times_\infty \sigma$ on $X \times Y$, where

$$\rho \times_1 \sigma((x,y),(x',y')) = \rho(x,x') + \sigma(y,y')$$

$$\rho \times_2 \sigma((x,y),(x',y')) = [\rho(x,x')^2 + \sigma(y,y')^2]^{1/2}$$

$$\rho \times_\infty \sigma((x,y),(x',y')) = \max\{\rho(x,x'), \sigma(y,y')\}$$

For $x, x' \in X$, $y, y' \in Y$.

PROOF: It is easily checked that these are metrics (although the triangle inequality for $\rho \times_2 \sigma$ takes some work, cf. ()), and that they are equivalent, so it suffices to work with $\rho \times_\infty \sigma$. The open $(\rho \times_\infty \sigma)$-balls are open rectangles, hence in $\mathcal{T} \times \mathcal{S}$; conversely, if $(x,y) \in X \times Y$ and $W \in \mathcal{T} \times \mathcal{S}$ with $(x,y) \in W$, there are $U \in \mathcal{T}$ and $V \in \mathcal{S}$ with $(x,y) \in U \times V \subseteq W$, and there is an $\epsilon > 0$ such that $B^\rho_\epsilon(x) \subseteq U$ and $B^\sigma_\epsilon(y) \subseteq V$; thus

$$B^{\rho \times_\infty \sigma}_\epsilon((x,y)) = B^\rho_\epsilon(x) \times B^\sigma_\epsilon(y) \subseteq U \times V \subseteq W$$

so the open $(\rho \times_\infty \sigma)$-balls are a base for $\mathcal{T} \times \mathcal{S}$. 👍

Convergence

Since open rectangles form a base for the product topology, the following is obvious:

XIII.6.1.7. PROPOSITION. Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be topological spaces, and $((x_i, y_i)_{i \in I})$ be a net in $X \times Y$, $(x,y) \in X \times Y$. Then $(x_i, y_i) \to (x,y)$ in $(X \times Y, \mathcal{T} \times \mathcal{S})$ if and only if $x_i \to x$ in $(X, \mathcal{T})$ and $y_i \to y$ in $(Y, \mathcal{S})$.

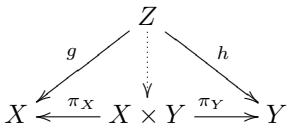
Because of this, convergence in the product topology is often called *coordinatewise convergence*.

Universal Property

Because $\mathcal{T} \times \mathcal{S}$ is the weakest topology on $X \times Y$ for which the coordinate projections are continuous, the following universal property is obvious:

XIII.6.1.8. PROPOSITION. Let $(X, \mathcal{T})$, $(Y, \mathcal{S})$, $(Z, \mathcal{R})$ be topological spaces, and $f : Z \to X \times Y$ a function. Then f is $(\mathcal{R}, \mathcal{T} \times \mathcal{S})$-continuous if and only if $\pi_X \circ f$ is $(\mathcal{R}, \mathcal{T})$-continuous and $\pi_Y \circ f$ is $(\mathcal{R}, \mathcal{S})$-continuous. Conversely, if $g : Z \to X$ is $(\mathcal{R}, \mathcal{T})$-continuous and $h : R \to Y$ is $(\mathcal{R}, \mathcal{S})$-continuous, then $(g, h) : Z \to X \times Y$ is $(\mathcal{R}, \mathcal{T} \times \mathcal{S})$-continuous.

XIII.6.1.9. In other words, the following diagram has a unique completion in the category $\mathcal{C}$ of topological spaces and continuous functions:



i.e. $(X \times Y, \mathcal{T} \times \mathcal{S})$ is a *product* of $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ in $\mathcal{C}$ ().

Finite Products

XIII.6.1.10. All the results of this subsection hold equally for products of finitely many topological spaces $\{(X_k, \mathcal{T}_k) : 1 \leq k \leq n\}$. Products of open sets are called *open rectangles*, *open boxes*, or *basic open sets*. The product topology on $\prod_{k=1}^n X_k$ is conventionally denoted $\prod_{k=1}^n \mathcal{T}_k$, although the notation $\bigotimes_{k=1}^n \mathcal{T}_k$ would be better. Details are left to the reader.

XIII.6.2. General Products

We now consider the topological product of an arbitrary indexed family of topological spaces $\{(X_i, \mathcal{T}_i) : i \in I\}$.

XIII.6.2.1. In the absence of the Axiom of Choice, there may be a set-theoretic problem: we could have $\prod_{i \in I} X_i = \emptyset$. But if the product is nonempty, it is fairly large (). (We implicitly assume that each X_i is nonempty, and that I is nonempty.) We will not assume the AC, but we will assume, usually without explicit mention, that the product is nonempty.

XIII.6.2.2. In analogy with the case of finitely many X_i, we could expect we want any set of the form $\prod_{i \in I} U_i$ to be open in $\prod_i X_i$, where $U_i \in \mathcal{T}_i$ for each i. Such a subset of $\prod_i X_i$ is called an *open box*. The set of all unions of open boxes is a topology on $\prod_i X_i$, called the *box topology*. However, it turns out that if I is infinite, the box topology has pathological properties (cf. XIII.6.3.8.) and is not very useful, although it is occasionally considered.

XIII.6.2.3. The proper definition of the product topology is a little less obvious. We say that an open box $\prod_{i \in I} U_i$ is an *open rectangle*, or *basic open set*, if U_i is the whole space X_i for all but finitely many i. (This means that membership in an open rectangle depends on only finitely many coordinates.)

XIII.6.2.4. DEFINITION. Let $\{(X_i, \mathcal{T}_i) : i \in I\}$ be an indexed collection of topological spaces, with $\prod_{i\in I} X_i \neq \emptyset$. The *product topology* on $\prod_i X_i$ is the set of unions of open rectangles.

The product topology on $\prod_{i\in I} X_i$ is conventionally denoted $\prod_{i\in I} \mathcal{T}_i$, although the notation $\bigotimes_{i\in I} \mathcal{T}_i$ would be better.

XIII.6.2.5. This definition makes more sense if coordinate projections are considered. Let $X = \prod_{i\in I} X_i$, and fix $i \in I$, and let π_i be the coordinate projection from X to X_i. If $U \in \mathcal{T}_i$, then $\pi_i^{-1}(U) = U \times \prod_{j\neq i} X_j$, so for π_i to be continuous such sets must be open in X. These sets (for various i) form a subbase for the weakest topology on X making each π_i continuous. Open rectangles are just finite intersections of such sets, and hence form a base for this topology which is therefore exactly the product topology. In the language of XIII.5.1.9., the product topology is the weak topology on X defined by the π_i.

XIII.6.2.6. The product topology is weaker than the box topology. The two coincide if I is finite, but not usually if I is infinite. The product topology is much more useful than the box topology in general, and when products are considered we will always assume we use the product topology without explicit mention to the contrary.

XIII.6.2.7. PROPOSITION. Let $\{(X_i, \mathcal{T}_i) : i \in I\}$ be an indexed collection of topological spaces, with $X = \prod_i X_i \neq \emptyset$, and for each i let $\mathcal{B}_i$ be a base for $\mathcal{T}_i$. Let $\mathcal{B}$ be the collection of all subsets of X of the form $\prod_{i\in F} B_i \times \prod_{i\in I\setminus F} X_i$, where F is a finite subset of I and $B_i \in \mathcal{B}_i$ for $i \in F$. Then $\mathcal{B}$ is a base for the product topology on X.

Convergence

The analog of XIII.6.1.7. holds for general products (i.e. the product topology is the topology of coordinate-wise convergence, or "pointwise convergence on I"), immediate from the fact that open rectangles form a base:

XIII.6.2.8. PROPOSITION. Let $\{(X_i, \mathcal{T}_i) : i \in I\}$ be an indexed collection of topological spaces, with $X = \prod_i X_i \neq \emptyset$. Give X the product topology $\mathcal{T}$. Let $(x_\lambda)_{\lambda\in\Lambda}$ be a net in X, and for each i and λ let $x_{\lambda,i} = \pi_i(x_\lambda)$, where $\pi_i : X \to X_i$ is the i'th coordinate projection. If $x = (\cdots x_i \cdots) \in X$, then $x_\lambda \to x$ in $(X, \mathcal{T})$ if and only if the net $(x_{\lambda,i})$ converges to x_i in $(X_i, \mathcal{T}_i)$ for each $i \in I$.

The analog of XIII.6.1.8. holds:

XIII.6.2.9. PROPOSITION. Let $\{(X_i, \mathcal{T}_i) : i \in I\}$, $(Z, \mathcal{R})$ be topological spaces, with $X = \prod_{i\in I} X_i \neq \emptyset$, and $f : Z \to X$ a function. Give X the product topology $\mathcal{T}$. Then f is $(\mathcal{R}, \mathcal{T})$-continuous if and only if $\pi_i \circ f$ is $(\mathcal{R}, \mathcal{T}_i)$-continuous for each i. Conversely, if $g_i : Z \to X_i$ is $(\mathcal{R}, \mathcal{T}_i)$-continuous for each i, then $(\cdots g_i \cdots) : Z \to X$ is $(\mathcal{R}, \mathcal{T})$-continuous.

XIII.6.3. Exercises

XIII.6.3.1. Let $(X, \mathcal{T})$ be the Sierpinski space (). What are the open sets in $(X \times X, \mathcal{T} \times \mathcal{T})$? Is every open set an open rectangle? [There are six open sets.]

XIII.6.3.2. Let I be an uncountable index set, and $\{X_i : i \in I\}$ a collection of topological spaces. Suppose for each i there is a point $p_i \in X_i$ which has an open neighborhood U_i which is not all of X_i (there will be such a p_i unless X_i has the indiscrete topology). Let $X = \prod_{i\in I} X_i$ with the product topology, and $p = (\cdots p_i \cdots) \in X$. Show that p does not have a countable neighborhood base, and hence X is not first countable. [If $\{V_n : n \in \mathbb{N}\}$ is a sequence of open neighborhoods of p, then for each n there is a finite subset F_n of I and nonempty open sets W_{ni} such that V_n contains $\prod_{i\in F_n} W_{ni} \times \prod_{i\notin F_n} X_i$. If $j \in I \setminus (\cup_{n=1}^{\infty} F_n)$, then $U_j \times \prod_{i\neq j} X_i$ does not contain any V_n.] This result requires the AC.

XIII.6.3.3. Let $\{X_i : i \in I\}$ be a collection of topological spaces, and let $X = \prod_{i\in I} X_i$ with the product topology. Suppose each X_i is separable and $card(I) \leq 2^{\aleph_0}$. Show that X is separable. [Let $\{p_{in} : n \in \mathbb{N}\}$ be a countable dense set in X_i for each i. Identify I with a subset of $\mathbb{R}$. For each finite set $S = \{(a_1, b_1), \ldots, (a_m, b_m)\}$ of disjoint open intervals in $\mathbb{R}$ with rational endpoints, and every corresponding tuple $N = (n_1, \ldots, n_m)$ of natural numbers, let $p_{(S,N)}$ be the point of x whose i'th coordinate is p_{in_k} if $i \in (a_k, b_k)$ and p_{i1} otherwise. Show that the set of all $p_{(S,N)}$ is a countable dense set in X.]

This result requires the AC for index sets of cardinality $2^{\aleph_0}$.

XIII.6.3.4. Let D be a countable dense set in $[0, 1]^{\mathfrak{c}}$. Then D, with the relative topology, is a countable normal topological space in which no point has a countable local base.

XIII.6.3.5. Let $\{X_i : i \in I\}$ be a collection of topological spaces, and let $X = \prod_{i\in I} X_i$ with the product topology. Suppose each X_i has two disjoint nonempty open sets U_i and V_i (this will happen if X_i is Hausdorff and has more than one point), and that X is separable. Show that $card(I) \leq 2^{\aleph_0}$. [If $\{p_n : n \in \mathbb{N}\}$ is dense in X, for each i let $S(i)$ be the set of all n such that the i'th coordinate of p_n is in U_i. Show that $S(i)$ and $S(j)$ are distinct subsets of $\mathbb{N}$ for $i \neq j$.]

XIII.6.3.6. Let X be an arbitrary product of copies of $\mathbb{N}$ with the finite complement topology. Show that X is separable in the product topology. [For each $n \in \mathbb{N}$ let p_n be the point of X with all coordinates n.]

XIII.6.3.7. (a) ([Sto48]; cf. [Eng89, 2.3.E]) Show that $\mathbb{N}^{\aleph_1}$ is not normal. [For $j = 1, 2$ let A_j be the set of all points for which, for each $k \neq j$, at most one coordinate is k. Show that A_1 and A_2 are closed and disjoint, but do not have disjoint neighborhoods.]

(b) Conclude from XIII.11.14.11. that if I is an index set with $card(I) \geq \aleph_1$, and, for each $i \in I$, X_i is a topological space which is not countably compact, then $\prod_{i\in I} X_i$ is not normal. In particular, if X is a topological space which is not countably compact, then $X^{\aleph_1}$ is not normal.

XIII.6.3.8. Let X be the product of countably many copies of $\mathbb{R}$, and $\mathcal{T}$ the box topology on X from the ordinary topology on $\mathbb{R}$. For each $t \in \mathbb{R}$, let $\boldsymbol{t}$ be the point of X with all coordinates t, and $Y = \{\boldsymbol{t} : t \in \mathbb{R}\} \subseteq X$. Then $\phi : \mathbb{R} \to Y$, $\phi(t) = \boldsymbol{t}$, is a bijection. Show that the relative box topology on Y is the discrete topology, and ϕ is not continuous.

XIII.6.3.9. Let $\{X_i : i \in i\}$ be an indexed collection of sets, with $X = \prod_i X_i \neq \emptyset$. For each i let $\mathcal{T}_i$ be the discrete topology on X_i. Show that the box topology on X is the discrete topology, but the product topology is not unless I is finite or all but finitely many X_i are singletons.

XIII.7. Separation Axioms

The notion of topological space is so general that practically nothing can be said about all topological spaces. All nontrivial results in topology require additional restrictions, usually including *separation axioms*. Separation axioms roughly insure that there are "enough" open sets in the topology. There are several important separation axioms of varying strength, described in this section. Virtually all topological spaces arising in applications, particularly in analysis, including any metrizable space, satisfy all these separation axioms.

Note that the separation axioms have nothing to do with separability in the sense of XIII.2.3.11..

XIII.7.1. The T-Axioms

The first five separation axioms first appeared together in [AH74], where they were called T_0–T_4 (T for *Trennungsaxiom*); the authors attributed each property to someone else (Kolmogorov, Fréchet, Hausdorff, Vietoris, and Tietze respectively), and most date back to the early 1920s (cf. [Tie23], [Ury25a]). Alternate names for T_2–T_4 predate [AH74] and are usually used today. There are also a $T_{3\frac{1}{2}}$ (XIII.7.8.3.), and a T_5, discussed in XIII.10.1.2.; some add $T_{2\frac{1}{2}}$ (XIII.7.10.21.) and T_6 (XIII.10.6.11.).

XIII.7.1.1. Definition. Let X be a topological space. Then

(o) X is T_0 if, whenever p and q are distinct points of X, there is an open neighborhood of at least one of the points not containing the other.

(i) X is T_1 if each singleton subset of X is closed.

(ii) X is T_2 if any two distinct points of X have disjoint open neighborhoods.

(iii) X is T_3 if whenever E is a closed set in X and p a point of X not in E, then p and E have disjoint open neighborhoods.

(iv) X is T_4 if disjoint closed sets in X have disjoint open neighborhoods.

X is *Hausdorff* if it is T_2.
X is *regular* if it is T_3 and Hausdorff.
X is *normal* if it is T_4 and Hausdorff.

Each of these separation axioms is discussed in more detail below.

XIII.7.1.2. There is some nonuniformity of use of the terminology in this definition. Some references define "regular" and "normal" to mean what we have called T_3 and T_4, and define T_3 to mean "T_2 and regular", and similarly for T_4. Our definition reflects the original historical usage of the terms. We will regard "T_3" and "regular" as essentially synonymous since they differ only in the case of spaces which are not T_0 (XIII.7.5.2.), and similarly "T_4" and "normal" are essentially equivalent. (In fact, Hausdorff's definition of topological space in [Hau49] included the T_2 axiom, and his definition was generally used by others in the 1920s.)

Some references use the term *separated* as a synonym for "Hausdorff." But this term is too easily confused with "separable" as defined in XIII.2.3.11..

XIII.7.2. T_0-Spaces

XIII.7.2.1. PROPOSITION. Let X be a topological space, and p and q distinct points of X. The following are equivalent:

(i) $\{p\}$ and $\{q\}$ have identical closures.

(ii) p is in the closure of $\{q\}$ and q is in the closure of $\{p\}$.

(iii) Any open set containing p also contains q, and vice versa.

X is T_0 if and only if none of these occur for any distinct p and q.

PROOF: (i) $\Rightarrow$ (iii): if U is an open set containing p but not q, then U^c is a closed set containing q and hence containing the closure of $\{q\}$, so the closure of $\{q\}$ is not the same as the closure of $\{p\}$ (in fact, does not contain p). The same argument works with p and q interchanged.

(iii) $\Rightarrow$ (ii): If p is not in the closure of $\{q\}$, the complement of the closure of $\{q\}$ is an open set containing p but not q. The same argument works if p and q are interchanged.

(ii) $\Rightarrow$ (i) is trivial.

It is obvious that the T_0 condition is precisely the negation of (iii). 👍

XIII.7.2.2. COROLLARY. A T_1 topological space is T_0.

There are T_0 spaces which are not T_1, e.g. the Sierpiński space.

The T_0 property is a local property. Say that a topological space X is *locally* T_0 if every point of X has an open neighborhood which is T_0 in the relative topology. Every T_0 space is obviously locally T_0. The converse is also true:

XIII.7.2.3. PROPOSITION. A locally T_0 space is T_0.

PROOF: Let X be a locally T_0 space, and p, q distinct points of X. We must find an open set containing one of p and q and not the other. Let U be an open neighborhood of p which is T_0 in the relative topology. If $q \notin U$, then U is the desired set. If $q \in U$, there is a relatively open subset V of U containing one of p and q but not the other. Since U is open in X, V is also open in X (), and is the desired subset. 👍

XIII.7.2.4. Every topological space has a maximal T_0 quotient, which captures all the topological information of the space (XIII.8.3.10.).

XIII.7.2.5. There seems to be no reason to consider topological spaces which are not at least T_0 (especially in light of XIII.8.3.10.). Many authors feel the same about restricting to T_1 or T_2 spaces; indeed, topological spaces which are not Hausdorff are rather bizarre and are rarely encountered. In analysis almost all spaces encountered are even completely regular. However, there are topological spaces arising in applications in algebra and logic, such as spectra of rings (XIII.7.10.20., XX.16.4.39.), which are T_0 but not necessarily even T_1, so such spaces cannot be completely ignored.

And there is one situation in analysis where spaces which are not T_0 arise naturally: certain natural topological groups () and topological vector spaces () are not T_0. In this case, taking the maximal T_0 quotient amounts to dividing out by the closure of the identity (), hence the T_0 quotient is also naturally a topological group or topological vector space.

XIII.7.3. T_1-Spaces

The T_1 property has an alternate characterization, which was actually the original definition:

XIII.7.3.1. Proposition. A topological space X is T_1 if and only if, whenever p and q are distinct points of X, each of p and q has an open neighborhood not containing the other.

Proof: If X is T_1 and p and q are distinct points of X, then $X\setminus\{q\}$ and $X\setminus\{p\}$ are the desired neighborhoods of p and q. Conversely, let $p \in X$. For each $q \in X\setminus\{p\}$, there is an open neighborhood U_q of q not containing p. The union of the U_q is open, and equals $X \setminus \{p\}$, so $\{p\}$ is closed. 👍

This characterization gives a simpler proof that a T_1 space is T_0, and also immediately implies that a Hausdorff space is T_1. We used the definition we did for T_1 since it is the version which is usually verified and used in applications.

XIII.7.3.2. Note that the condition in XIII.7.3.1. can be rephrased: whenever p and q are distinct points of X, then p has an open neighborhood not containing q. For if this statement is true for any p and q (in that order), it is also true for q and p, i.e. each of p and q has a neighborhood not containing the other.

XIII.7.3.3. Corollary. A topological space X is T_1 if and only if every singleton subset of X is an intersection of open sets.

XIII.7.3.4. Like the T_0 property, the T_1 property is a local property. Say that a topological space X is *locally* T_1 if every point of X has an open neighborhood which is T_1 in the relative topology. Every T_1 space is obviously locally T_1. The converse is also true:

XIII.7.3.5. Proposition. A locally T_1 space is T_1.

The proof is a slightly simplified version of the argument in the proof of XIII.7.2.3., and is left to the reader.

XIII.7.4. Hausdorff Spaces

The Hausdorff property is probably the most widely used separation axiom. Here is one of the main reasons. One of the most basic facts about limits in $\mathbb{R}$ or $\mathbb{R}^n$, or in general metric spaces, is that limits are unique (). This property only holds in Hausdorff spaces:

XIII.7.4.1. PROPOSITION. Let X be a topological space. Then X is Hausdorff if and only if every convergent net in X has a unique limit.

PROOF: Suppose X is Hausdorff, and (x_i) is a net in X with $x_i \to x$ and $x_i \to y$ with $x \neq y$. Let U and V be disjoint neighborhoods of x and y respectively. Then there is an i_1 such that $x_i \in U$ for all $i \geq i_1$, and there is an i_2 such that $x_i \in V$ for all $i \geq i_2$. If $i \geq i_1$ and $i \geq i_2$ (there is such an i since the index set is directed), then $x_i \in U \cap V$, a contradiction. Thus the limit of a convergent net is unique.

Conversely, suppose X is not Hausdorff, and let x and y be distinct points which do not have disjoint neighborhoods. Let $I = \{(U, V, z)\}$, where U is a neighborhood of x, V is a neighborhood of y, and $z \in U \cap V$. Put a preorder on I by setting $(U_1, V_1, z_1) \leq (U_2, V_2, z_2)$ if $U_2 \subseteq U_1$ and $V_2 \subseteq V_1$. Then I is a directed set. Set $x_{(U,V,z)} = z$; this defines a net with index set I, and this net converges to both x and y. So limits of convergent nets are not unique. 👍

XIII.7.4.2. Note that the "if" part can be false if "net" is replaced by "sequence" if the space is not first countable (Exercise XIII.7.10.16.). In a first countable space, the sequence version is true ().

The (equivalent) filter version of XIII.7.4.1. is especially easy to prove:

XIII.7.4.3. PROPOSITION. Let X be a topological space. Then X is Hausdorff if and only if every convergent filter in X has a unique limit.

PROOF: If X is Hausdorff and $x, y \in X$, $x \neq y$, then the neighborhood filters of x and y cannot be contained in a single filter since x and y have disjoint neighborhoods. Conversely, if X is not Hausdorff, let x and y be distinct points of X which do not have disjoint neighborhoods, and let $\mathcal{D}$ be the collection of all sets of the form $U \cap V$, where U is a neighborhood of x and V a neighborhood of y. Then $\mathcal{D}$ is a filterbase in X which converges to both x and y. 👍

XIII.7.4.4. Unlike the T_0 and T_1 properties, the Hausdorff property is not a local property. Say that a topological space X is *locally Hausdorff* if every point of X has an open neighborhood which is Hausdorff in the relative topology. A Hausdorff space is obviously locally Hausdorff. But a locally Hausdorff space need not be Hausdorff (XIII.20.1.3.(v)), although it is T_1 (XIII.7.3.5.). However, see XIII.7.10.5..

XIII.7.5. Regular Spaces

The T_3 separation axiom has a simple reformulation which reinforces the utility of the axiom:

XIII.7.5.1. PROPOSITION. Let X be a topological space. Then X is a T_3 space if and only if, whenever p is a point of X and U is an open neighborhood of p, there is an open neighborhood V of p with $\bar{V} \subseteq U$.

PROOF: Suppose X is T_3, and U is an open neighborhood of p. Then U^c is a closed set not containing p, so p and U^c have disjoint open neighborhoods V and W. Then $\bar{V}$ is contained in the closed set W^c, hence in U. Conversely, suppose X has the property in the statement, and E is a closed set in X and p is a point of X not in E. Then $U = E^c$ is an open neighborhood of p, so there is an open neighborhood V of p with $\bar{V} \subseteq U$. Then V and $(\bar{V})^c$ are disjoint open neighborhoods of p and E, so X is a T_3 space. 👍

In the presence of T_3, only the T_0 axiom is needed to give the Hausdorff property and thus regularity:

XIII.7.5.2. PROPOSITION. Let X be a topological space which is T_3 and T_0. Then X is Hausdorff, and thus regular.

PROOF: Let p and q be distinct points of X. Suppose p is not in the closure E of $\{q\}$. Then p and E have disjoint open neighborhoods, which are disjoint open neighborhoods of p and q. The same proof works if q is not in the closure of $\{p\}$. Apply XIII.7.2.1.. 👍

A set with the indiscrete topology and more than one point is an example of a T_3 space which is not T_0.

XIII.7.6. Normal Spaces

XIII.7.6.1. In the presence of T_4, the T_1 condition gives T_2, hence normality, as is obvious. However, unlike with regularity, T_0 is not enough: the Sierpiński space is T_0 and vacuously T_4, but not T_1. It is also obvious that normality implies regularity, although T_4 does not imply T_3 in the absence of T_1 (the Sierpiński space is again a counterexample).

There is an analog of XIII.7.5.1. for normal spaces. The proof is virtually identical.

XIII.7.6.2. PROPOSITION. Let X be a topological space. Then X is a T_4 space if and only if, whenever E is a closed set in X and U is an open neighborhood of E, there is an open neighborhood V of E with $\bar{V} \subseteq U$.

XIII.7.6.3. PROPOSITION. A closed subspace of a T_4 space is T_4. Hence a closed subspace of a normal space is normal.

PROOF: If A is a closed subspace of a T_4 space X, and Y, Z are (relatively) closed subsets of A, then Y and Z are closed in X, hence have disjoint neighborhoods U, V in X. Then $A \cap U$ and $A \cap V$ are disjoint neighborhoods in A. 👍

A general subspace of a normal space is not necessarily normal (XIII.11.14.7.). However, an F_σ in a normal space is normal (XIII.7.10.8.).

An alternate way of phrasing normality is using partitions:

XIII.7.6.4. Definition. Let X be a topological space, and Y, Z disjoint closed subsets of X. A *partition* between Y and Z in X is a (necessarily closed) subset P of X such that there are open sets $U, V \subseteq X$ with $Y \subseteq U$, $Z \subseteq V$, $U \cap V = \emptyset$, and $X \setminus (U \cup V) = P$.

Note that there can be many partitions between disjoint subsets. A partition can be empty: $\emptyset$ is a partition between Y and Z in X if (and only if) there is a clopen set U in X with $Y \subseteq U$ and $Z \subseteq (X \setminus U)$.

XIII.7.6.5. Proposition. Let X be a topological space. Then X is T_4 if and only if every pair of disjoint closed subsets of X has a partition in X.

Proof: If every pair of disjoint closed subsets of X has a partition in X, and Y, Z are disjoint closed subsets of X, let P be a partition between Y and Z in X; the associated U and V are disjoint neighborhoods of Y and Z. Thus X is T_4. Conversely, suppose X is T_4 and Y and Z are disjoint closed subsets of X. Let U and V be disjoint neighborhoods of Y and Z in X; then $P = X \setminus (U \cup V)$ is a partition between Y and Z in X. 👍

XIII.7.6.6. Proposition. A metrizable topological space is normal.

See XIII.7.10.25.. In fact, a metrizable space has a stronger property called *complete normality*. See XIII.10.1.1. for a definition and proof. In particular, a metrizable space is regular and Hausdorff.

Every compact Hausdorff space is also normal (XIII.11.4.3.).

Here is a useful refinement of the definition of normality:

XIII.7.6.7. Definition. Let X be a topological space, and $\mathcal{U} = \{U_i : i \in I\}$ an open cover of X. A *shrinking* of $\mathcal{U}$ is an open cover $\mathcal{V} = \{V_i : i \in I\}$ (same index set) with $\bar{V}_i \subseteq U_i$ for all i.

XIII.7.6.8. Proposition. Let X be a topological space. Then X is T_4 if and only if every finite open cover of X has a shrinking.

Proof: If X is not T_4, let Y and Z be disjoint closed sets in X which do not have disjoint neighborhoods. Then the open cover $\{X \setminus Y, X \setminus Z\}$ does not have a shrinking.

Conversely, suppose X is T_4. We show by induction on the number of sets in the cover that every finite open cover of X has a shrinking. First suppose $\mathcal{U} = \{U_1, U_2\}$ is an open cover of X. Then $X \setminus U_1$ and $X \setminus U_2$ are disjoint closed sets in X, and so have disjoint open neighborhoods V_1, V_2. By XIII.7.6.2. there are open neighborhoods W_1, W_2 with $\bar{W}_k \subseteq V_k$. Then $\{X \setminus \bar{W}_1, X \setminus \bar{W}_2\}$ is a shrinking of $\mathcal{U}$.

Now suppose every open cover with n sets has a shrinking, and let $\mathcal{U} = \{U_1, \ldots, U_{n+1}\}$ be an open cover of X. There is a shrinking $\{V_1, \ldots, V_{n-1}, W\}$ for the open cover $\{U_1, \ldots, U_{n-1}, U_n \cup U_{n+1}\}$. Then $F = \bar{W}$ is a closed subset of X, hence T_4, so the open cover $\{F \cap U_n, F \cap U_{n+1}\}$ of F has a shrinking $\{W_n, W_{n+1}\}$. Set $V_n = W \cap W_n$, $V_{n+1} = W \cap W_{n+1}$. Then $\{V_1, \ldots, V_{n+1}\}$ is a shrinking of $\mathcal{U}$. 👍

XIII.7.6.9. Proposition XIII.7.6.8. uses no form of Choice, i.e. is a theorem of ZF. Using the AC (Well-ordering), one direction can be extended to show that every point-finite open cover of a normal (T_4) space has a shrinking (Exercise XIII.7.10.24.). This is used e.g. in XIII.11.10.6..

There is a nice "dual" fact about normal spaces. We first make a definition:

XIII.7.6.10. Definition. Let X be a topological space, and $A_1, \ldots, A_n$ closed subsets of X. A *swelling* of $\{A_1, \ldots, A_n\}$ is a collection $\{U_1, \ldots, U_n\}$ of open sets in X with $A_k \subseteq U_k$ for each k and $U_{k_1} \cap U_{k_2} \cap \cdots \cap U_{k_m} = \emptyset$ for any $k_1, \ldots, k_m$ with $A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m} = \emptyset$.

XIII.7.6.11. Proposition. Let X be a T_4 topological space, and $A_1, \ldots, A_n$ closed subsets of X. Then $\{A_1, \ldots, A_n\}$ has a swelling.

Proof: Let B_1 be the union of all sets of the form $A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m}$ for which

$$A_1 \cap A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_m} = \emptyset \ .$$

Since this is a finite union, B_1 is closed; and A_1 and B_1 are disjoint. By normality there are disjoint open neighborhoods U_1 and V_1 of A_1 and B_1, so that $\bar{U}_1$ is disjoint from B_1.

Now repeat the process with A_2, with A_1 replaced by $\bar{U}_1$: let B_2 be the union of all finite intersections of $\{\bar{U}_1, A_3, \ldots, A_n\}$ which are disjoint from A_2, and obtain an open neighborhood U_2 of A_2 with $\bar{U}_2$ disjoint from B_2. Repeating the process n times (at the k'th step A_j is replaced by $\bar{U}_j$ for all $j < k$), we obtain a swelling $\{U_1, \ldots, U_n\}$. 👍

There are two important theorems about normality with nontrivial proofs. The first is due to ():

XIII.7.6.12. Theorem. A T_3 Lindelöf space is T_4. Thus a regular Lindelöf space is normal.

In fact, a stronger result holds: every regular Lindelöf space is paracompact (), and every paracompact space is normal (). But we give a straightforward proof of XIII.7.6.12. which is simpler than the proofs of these two stronger results.

Proof: Let X be a T_3 Lindelöf space, and E and F disjoint closed sets in X. For each $p \in E$, there is an open neighborhood W_p of p such that $\bar{W}_p \cap F = \emptyset$ (XIII.7.5.1.). Then $\{W_p : p \in E\} \cup \{E^c\}$ is an open cover of X, so has a countable subcover $\{E^c, W_1, W_2, \ldots\}$. Similarly, there is a countable cover $\{F^c, Z_1, Z_2, \ldots\}$ of X with $\bar{Z}_n \cap E = \emptyset$ for all n. We have $E \subseteq \cup_n W_n$ and $F \subseteq \cup_n Z_n$. Inductively define $U_1 = W_1$, $V_1 = Z_1 \setminus \bar{U}_1$, $U_2 = W_2 \setminus \bar{V}_1$, $V_2 = Z_2 \setminus (\bar{U}_1 \cup \bar{U}_2)$, $U_3 = W_3 \setminus (\bar{V}_1 \cup \bar{V}_2)$, etc. Each U_n and V_n is open, and $U_n \cap V_m = \emptyset$ for all n and m. If $p \in E$, then $p \in W_n$ for some n, hence $p \in U_n$ since $p \notin \bar{V}_m \subseteq \bar{Z}_m$ for any m. Similarly, if $q \in F$, then $q \in V_m$ for some m. Thus $U = \cup_n U_n$ and $V = \cup_n V_n$ are disjoint neighborhoods of E and F. 👍

The second theorem is a key result showing that normal spaces can be completely studied by analytic methods (specifically, real-valued continuous functions):

XIII.7.6.13. THEOREM. [URYSOHN'S LEMMA] Let X be a normal topological space. If E_0 and E_1 are disjoint closed sets in X, then there is a continuous function $f : X \to [0,1]$ with $f(x) = 0$ for all $x \in E_0$ and $f(x) = 1$ for all $x \in E_1$.

This result is a special case of the Tietze Extension Theorem, but we will use it in the proof of the Tietze Theorem.

PROOF: We will make repeated use of XIII.7.6.2.. Let $U_0 = E_0^c$. Then U_0 is an open neighborhood of E_1. Let $U_{1/2}$ be an open neighborhood of E_1 with $\bar{U}_{1/2} \subseteq U_0$. Let $U_{3/4}$ be an open neighborhood of E_1 with $\bar{U}_{3/4} \subseteq U_{1/2}$. U_0 is an open neighborhood of $\bar{U}_{1/2}$; let $U_{1/4}$ be an open neighborhood of $\bar{U}_{1/2}$ with $\bar{U}_{1/4} \subseteq U_0$. Continuing in this way, for each dyadic rational t with $0 \leq t < 1$ choose an open neighborhood U_t of E_1 such that $\bar{U}_t \subseteq U_s$ if $s < t$.

For $x \in X$ define $f(x) = \sup\{t : x \in U_t\}$ if $x \in U_0$ and $f(x) = 0$ if $x \in E_0$. Then $0 \leq f(x) \leq 1$ for all $x \in X$, and $f(x) = 1$ if $x \in E_1$. We show f is continuous. If (x_i) is a net in X with $x_i \to x$, we show that $f(x_i) \to f(x)$. Let $t = f(x)$. If $t > 0$, then for any dyadic rational s, $0 \leq s < t$, we have $x \in U_s$; since U_s is open, $x_i \in U_s$ for sufficiently large i, i.e. $f(x_i) \geq s$ for sufficiently large i. This shows that $f(x_i) \to f(x)$ if $t = 1$. Now suppose $t < 1$. If s is any dyadic rational with $t < s < 1$, choose a dyadic rational r with $t < r < s$. Then $x \notin U_r$. Since $\bar{U}_s \subseteq U_r$, x is in the open set $(\bar{U}_s)^c$, so $x_i \in (\bar{U}_s)^c$ for sufficiently large i, i.e. $f(x_i) \leq s$ for sufficiently large i. Thus $f(x_i) \to f(x)$. 👍

The Axiom of Dependent Choice is used in this proof. Some Choice is necessary for the result (XIII.11.14.20.). No Choice is necessary for metrizable spaces (XIII.7.10.25.).

XIII.7.6.14. Normality is a nice separation condition to have, but it is somewhat more delicate than the other conditions: for example, unlike all the other separation axioms it does not pass to subspaces or products in general (see XIII.11.14.7. and XIII.7.10.14. for counterexamples; cf. XIII.10.1.2.).

XIII.7.7. Fully Normal Spaces

An important tightening of normality is full normality. This notion apparently first explicitly appeared in the literature in [Tuk40]; TUKEY said the concept goes back at least to URYSOHN.

Star Refinements

XIII.7.7.1. DEFINITION. Let X be a topological space and $\mathcal{U}$ an open cover of X. For each $x \in X$, the *star* of $\mathcal{U}$ at x is

$$St(x,\mathcal{U}) = \bigcup\{U \in \mathcal{U} : x \in U\}$$

and if $A \subseteq X$, the *star* of $\mathcal{U}$ at A is

$$St(A,\mathcal{U}) = \bigcup\{U \in \mathcal{U} : A \cap U \neq \emptyset\} = \bigcup_{x \in A} St(x,\mathcal{U}) \ .$$

Note that $\{St(x,\mathcal{U}) : x \in X\}$ is an open cover of X, and that $St(A,\mathcal{U})$ is open for any $A \subseteq X$.

XIII.7.7.2. DEFINITION. Let X be a topological space and $\mathcal{U}, \mathcal{V}$ open covers of X.

(i) The cover $\mathcal{V}$ *barycenter-refines* or Δ-*refines* $\mathcal{U}$, or is a *barycenter refinement* or Δ-*refinement* of $\mathcal{U}$, if the open cover $\{St(x, \mathcal{V}) : x \in X\}$ refines $\mathcal{U}$, i.e. for every $x \in X$ there is a $U \in \mathcal{U}$ with $St(x, \mathcal{V}) \subseteq U$.

(ii) The cover $\mathcal{V}$ *star-refines* $\mathcal{U}$, or is a *star refinement* of $\mathcal{U}$, written $\mathcal{V} <^* \mathcal{U}$, if for every $V \in \mathcal{V}$ there is a $U \in \mathcal{U}$ with $St(V, \mathcal{V}) \subseteq U$.

XIII.7.7.3. A star refinement of $\mathcal{U}$ is a barycentric refinement of $\mathcal{U}$, and a barycentric refinement of $\mathcal{U}$ is a refinement of $\mathcal{U}$, but the converses are far from true in general. But it is easily checked that a barycentric refinement of a barycentric refinement of $\mathcal{U}$ is a star refinement of $\mathcal{U}$.

XIII.7.7.4. EXAMPLE. Let (X, ρ) be a metric space, and for $\epsilon > 0$ let $\mathcal{U}_\epsilon$ be the open cover of X consisting of all open balls of radius ϵ. Then $\mathcal{U}_{\epsilon/2}$ is a star-refinement of $\mathcal{U}_\epsilon$ for any $\epsilon > 0$.

XIII.7.7.5. DEFINITION. A Hausdorff topological space X is *fully normal* every open cover of X has a star refinement.

Because of XIII.7.7.3., a Hausdorff space is fully normal if and only if every open cover has a barycentric refinement.

XIII.7.7.6. PROPOSITION. A fully normal topological space is normal.

PROOF: Let X be fully normal. Then X is Hausdorff. Let A and B be disjoint closed sets in X, and set $\mathcal{U} = \{X \setminus A, X \setminus B\}$. Then $\mathcal{U}$ is an open cover of X. Let $\mathcal{V}$ be a star refinement of $\mathcal{U}$. Then $St(A, \mathcal{V})$ and $St(B, \mathcal{V})$ are disjoint neighborhoods of A and B respectively. 👍

XIII.7.7.7. There are many normal spaces which are not fully normal. In fact, a space is fully normal if and only if it is paracompact (XIII.11.10.6.). We give the essential part of the easier direction here:

XIII.7.7.8. THEOREM. Let X be a topological space, and $\mathcal{U} = \{U_i : i \in I\}$ a locally finite open cover of X which has a shrinking $\mathcal{V} = \{V_i : i \in I\}$. Then $\mathcal{U}$ has a barycentric refinement.

PROOF: The open cover $\mathcal{V}$ is also locally finite. For $x \in X$, set $A_x = \bigcap\{U_i : x \in \bar{V}_i\}$ and $B_x = \bigcup\{\bar{V}_i : x \notin \bar{V}_i\}$. Since $\mathcal{V}$ is locally finite, these are really finite unions and intersections, so A_x is open and B_x is closed; and $x \in A_x$, $x \notin B_x$. Let $W_x = A_x \setminus B_x$. Then W_x is an open neighborhood of x, so $\mathcal{W} = \{W_x : x \in X\}$ is an open cover of X.

We claim $\mathcal{W}$ is a barycentric refinement of $\mathcal{U}$. Let $y \in X$, and fix i with $y \in \bar{V}_i$. For any $x \in X$ with $y \in W_x$, we have $x \in \bar{V}_i$ since otherwise $\bar{V}_i \subseteq B_x$ contradicting that $y \in W_x$. Thus $W_x \subseteq A_x \subseteq U_i$. This is true for every x such that $y \in W_x$, so $St(y, \mathcal{W}) \subseteq U_i$. 👍

XIII.7.7.9. This theorem does not require any Choice. If the AC is assumed and X is normal, the hypothesis of the shrinking is automatic (XIII.7.6.9.). But there is an important consequence which does not require any Choice (XIII.11.4.5.).

XIII.7.8. Completely Regular Spaces

Another separation axiom is commonly added to the hierarchy:

XIII.7.8.1. DEFINITION. Let X be a topological space. Then X is *completely regular* if it is T_0 and, whenever E is a closed set in X and p a point of X not in E, there is a continuous function $f : X \to [0,1]$ such that $f(p) = 1$ and $f(x) = 0$ for all $x \in E$.

XIII.7.8.2. PROPOSITION. A completely regular space is regular (and in particular Hausdorff).

PROOF: Suppose X is completely regular. If E is a closed set in X and p a point of X not in E, and f is a continuous function from X to $[0,1]$ as in XIII.7.8.1., then $f^{-1}((1/2,1])$ and $f^{-1}([0,1/2))$ are disjoint open neighborhoods of p and E, so X is T_3. Since X is T_0, it is regular (). 👍

XIII.7.8.3. A normal space is completely regular by Urysohn's Lemma. Thus the completely regular condition is a separation axiom between T_3 and T_4, and is sometimes called $T_{3\frac{1}{2}}$ (with the usual ambiguity about whether the space is required to be Hausdorff). A completely regular space is also sometimes called a *Tikhonov space*.

There are regular spaces which are not completely regular (see [SS95]), and completely regular spaces which are not normal, e.g. XIII.20.5.8..

XIII.7.8.4. THEOREM. Let X be a topological space. Then X is homeomorphic to a subset of a compact Hausdorff space if and only if X is completely regular.

PROOF: Since compact Hausdorff spaces are normal (XIII.11.4.3.) and hence completely regular, and complete regularity obviously passes to subspaces, the complete regularity condition is necessary. Conversely, let X be completely regular. Let $\mathcal{C}$ be a set of continuous functions from X to $[0,1]$ such that for every closed set E in X and $p \in X \setminus E$, there is an $f \in \mathcal{C}$ with $f(p) = 1$ and $f = 0$ on E, and form the space $Y = [0,1]^{\mathcal{C}}$ with the product topology. Y is Hausdorff, and compact by Tikhonov's theorem (). Define a function $\phi : X \to Y$ by letting $\phi(x)$ be the point whose f'th coordinate is $f(x)$, for each $f \in \mathcal{C}$. ϕ is continuous since its composition with each coordinate projection is continuous. We need only show that it is a homeomorphism, e.g. if (x_i) is a net in X, and $x \in X$ with $\phi(x_i) \to \phi(x)$, then $x_i \to x$. We have $\phi(x_i) \to \phi(x)$ if and only if $f(x_i) \to f(x)$ for every $f \in \mathcal{C}$. If $x_i \not\to x$, there is a neighborhood U of x and a subnet (x_j) such that $x_j \notin U$ for all j. There is an $f \in \mathcal{C}$ such that $f(x) = 1$ and $f(y) = 0$ for all $y \in U^c$; for this f, $f(x_j) = 0$ for all j, so $f(x_i) \not\to f(x)$, a contradiction. 👍

XIII.7.8.5. A good case can be made that complete regularity is the most natural restriction for topological spaces, and that completely regular spaces form the proper class of spaces to consider, especially for applications in analysis. This is the general thesis of [GJ76], for example. Completely regular spaces are precisely the spaces characterized by each of the following conditions (under some mild Choice assumptions):

(i) Spaces which can be embedded in compact Hausdorff spaces.

(ii) Spaces with a topology arising from a uniform structure.

(iii) Spaces whose topology is defined by a (separating) family of pseudometrics.

(iv) Spaces whose structure can be completely analyzed by study of real-valued continuous functions (i.e. analytic methods).

XIII.7.9. The Tietze Extension Theorem

The Tietze Extension Theorem is an important extension of Urysohn's Lemma.

XIII.7.9.1. THEOREM. [TIETZE EXTENSION THEOREM] Let X be a normal topological space, and E a closed subspace. If I is any interval in $\mathbb{R}$, and f is a continuous function from E to I, then f extends to a continuous function from X to I. Conversely, if X is a T_1 topological space and every real-valued continuous function on a closed subspace of X extends to X, then X is normal.

PROOF: ($\Rightarrow$) Since every interval in $\mathbb{R}$ is homeomorphic to $[-1,1]$, $(-1,1)$, or $(-1,1]$, it suffices to prove the statement for these intervals.

First suppose $I = [-1,1]$. If $f : E \to I$ is continuous, let $A_1 = f^{-1}\left(\left[\frac{1}{3},1\right]\right)$ and $B_1 = f^{-1}\left(\left[-1,-\frac{1}{3}\right]\right)$. Then A_1 and B_1 are disjoint closed subsets of E, hence of X, so by Urysohn's Lemma there is a continuous function $g_1 : X \to \left[-\frac{1}{3},\frac{1}{3}\right]$ with $g_1(x) = \frac{1}{3}$ for $x \in A_1$ and $g_1(x) = -\frac{1}{3}$ for $x \in B_1$. Then, for all $x \in E$,

$$|f(x) - g_1(x)| \leq \frac{2}{3}$$

for all $x \in E$. Thus $f - g_1 : E \to \left[-\frac{2}{3},\frac{2}{3}\right]$. Repeat the process with $f - g_1$ in place of f and $\frac{2}{9}$ in place of $\frac{1}{3}$ to get $g_2 : X \to \left[-\frac{2}{9},\frac{2}{9}\right]$ with

$$|(f(x) - g_1(x)) - g_2(x)| \leq \frac{4}{9}$$

for all $x \in E$. Continuing inductively, for each n there is a continuous function $g_n : X \to \left[-\frac{2^{n-1}}{3^n},\frac{2^{n-1}}{3^n}\right]$ with

$$\left| f(x) - \sum_{k=1}^{n} g_k(x) \right| \leq \left[\frac{2}{3}\right]^n$$

for all $x \in E$.

Since $|g_k(x)| \leq \frac{2^{k-1}}{3^k}$ for all k and $x \in X$, the series $\sum_{k=1}^{\infty} g_k(x)$ converges uniformly on X to a function $g : X \to \mathbb{R}$ by the Weierstrass M-test (), and g is continuous by (). The function g takes values in $[-r,r]$, where

$$r = \sum_{k=1}^{\infty} \frac{2^{k-1}}{3^k} = \frac{1}{2} \cdot \frac{\frac{2}{3}}{1-\frac{2}{3}} = 1 \ .$$

We clearly have $g(x) = f(x)$ for $x \in E$.

Now suppose $I = (1,1)$ and $f : E \to I$. There is an extension of f to a continuous function $g : X \to [-1,1]$. If $B = g^{-1}(\{-1,1\})$, then B is a closed set in X disjoint from E, so there is a continuous function $\phi : X \to [0,1]$ with $\phi(x) = 1$ for $x \in E$ and $\phi(x) = 0$ for $x \in B$. Set $h = \phi \cdot g$. Then $h : X \to (-1,1)$ and $h = f$ on E. The proof if $I = (-1,1]$ is almost identical.

Conversely, suppose X has the Tietze extension property and A, B are disjoint closed subsets of X. Then $A \cup B$ is closed. Let $f : A \cup B \to [0,1]$ with $f = 0$ on A and $f = 1$ on B. Extend f to a continuous $g : X \to [0,1]$, and set $U = g^{-1}\left(\left[0, \frac{1}{2}\right)\right)$ and $V = g^{-1}\left(\left(\frac{1}{2}, 1\right]\right)$. Then U and V are disjoint open neighborhoods of A and B. So X is normal. 👍

XIII.7.9.2. This theorem requires Dependent Choice, both for this argument and Urysohn's Lemma itself. However, if X is metrizable, there is a canonical choice of Urysohn function for any pair of disjoint closed subspaces (XIII.7.10.25.), so the construction of the extension is recursive and no Choice is needed.

XIII.7.10. Exercises

XIII.7.10.1. Let $(X, \mathcal{T})$ be a topological space, and $Y \subseteq X$. Give Y the relative topology $\mathcal{T}_Y$.

(a) If $(X, \mathcal{T})$ is T_0, then $(Y, \mathcal{T}_Y)$ is T_0.

(b) If $(X, \mathcal{T})$ is T_1, then $(Y, \mathcal{T}_Y)$ is T_1.

(c) If $(X, \mathcal{T})$ is Hausdorff, then $(Y, \mathcal{T}_Y)$ is Hausdorff.

(d) If $(X, \mathcal{T})$ is regular, then $(Y, \mathcal{T}_Y)$ is regular.

(e) If $(X, \mathcal{T})$ is completely regular, then $(Y, \mathcal{T}_Y)$ is completely regular.

The corresponding statement for normal spaces is false (XIII.11.14.7.).

XIII.7.10.2. Let X be a topological space. Show that the following are equivalent:

(i) Every singleton subset of X is an intersection of open sets.

(ii) Every subset of X is an intersection of open sets.

(iii) X is T_1.

XIII.7.10.3. Let $X = \{a, b\}$, and let $\mathcal{T} = \{\emptyset, \{a\}, X\}$.

(a) Show that $\mathcal{T}$ is a topology on X.

(b) $\{b\}$ is closed, but $\{a\}$ is not closed; $\{a\}$ is open, and the closure of $\{a\}$ is X (i.e. $\{a\}$ is dense in X).

(c) $(X, \mathcal{T})$ satisfies T_0 and T_4, but not T_1, T_2, or T_3.

(d) If $x_n = a$ for all n, then $x_n \to a$ and $x_n \to b$.

(e) X is compact () and connected (). In fact, X is path-connected ().

$(X, \mathcal{T})$ is called the *Sierpiński space*.

XIII.7.10.4. (a) Let X be a T_0 space, and U a finite nonempty open subset of X. Show that U contains an open singleton subset of X. [Consider a minimal nonempty open subset of U.]

(b) Let X be a T_1 space, and U a finite nonempty open subset of X. Show that every point of U is an isolated point of X. In particular, a finite T_1 space has the discrete topology.

A T_0 space containing a nonempty finite open set need not have any isolated points, and not every singleton subset of a finite open set is open in general. The Sierpiński space is a counterexample to both.

XIII.7.10.5. Let X be a topological space.

(a) If each point of X has a closed neighborhood which is Hausdorff in the relative topology, show that X is Hausdorff.

(b) If each point of X has a closed neighborhood which is T_3 in the relative topology, show that X is T_3.

(c) (cf. [Eng89, 3.3.1]) If each point of X has a closed neighborhood which is $T_{3\frac{1}{2}}$ in the relative topology, show that X is $T_{3\frac{1}{2}}$.

(d) If each point of X has a closed neighborhood which is [completely] regular in the relative topology, show that X is [completely] regular.

Example XIII.20.1.3.(v) shows that the word "closed" cannot be removed from any of the statements. Since there are locally compact Hausdorff spaces which are not normal (XIII.11.14.7.), the corresponding statements with "T_4" or "normal" do not hold.

XIII.7.10.6. (a) Let X be a Hausdorff space. Suppose X has a dense subspace D of cardinality κ. Show that $card(X) \leq 2^{2^\kappa}$. Thus $card(X) \leq 2^{2^{d(X)}}$ (XIII.2.3.16.). [For each $p \in X$, let

$$\mathcal{A}_p = \{U \cap D : U \text{ an open neighborhood of } p\} \in \mathcal{P}(\mathcal{P}(D)) \ .$$

Show that $\mathcal{A}_p \neq \mathcal{A}_q$ if $p \neq q$.] In particular, a separable Hausdorff space has cardinality $\leq 2^{2^{\aleph_0}}$.

There are separable Hausdorff spaces of cardinality $2^{2^{\aleph_0}}$, even compact ones, e.g. $\beta\mathbb{N}$ () and $[0,1]^{\mathfrak{c}}$ (). In fact, if Y is a discrete space of infinite cardinality κ, then the Stone-Čech compactification βY is a compact Hausdorff space of cardinality 2^{2^κ}.

(b) Show that if X is any set with the finite complement topology (), then any infinite subset of X is dense, and hence X is separable if $card(X) \geq \aleph_0$. Thus the conclusion of (a) fails in general for compact T_1 spaces, and there is no bound at all on the cardinality of such spaces.

(c) If X is a separable, first countable Hausdorff space, show that X has cardinality $\leq 2^{\aleph_0}$. [If D is a countable dense set in X, for each $x \in X$ choose a sequence in D converging to x.] In general, if X is Hausdorff, we have $card(X) \leq d(X)^{\chi(X)}$ (XIII.1.2.11., XIII.2.3.16.).

Note that unlike for (a), the Axiom of Choice is needed for (c) [the AC is needed for the examples at the end of (a)].

XIII.7.10.7. (a) Let X be a regular space. Suppose X has a dense subspace D of cardinality κ. Show there is a base for the topology of cardinality $\leq 2^\kappa$. Thus $w(X) \leq 2^{d(X)}$ (XIII.1.2.11., XIII.2.3.16.). [If $\{U_i : i \in I\}$ is a base, for each i set $V_i = [(U_i \cap D)^-]^o$. Note that $U_i \subseteq V_i \subseteq \bar{U}_i$. Show that $\{V_i : i \in I\}$ is a base of cardinality $\leq 2^\kappa$.]

(b) If X is a T_0-space, and there is a base for the topology of cardinality κ, then $card(X) \leq 2^\kappa$. Thus $card(X) \leq 2^{w(X)}$ (XIII.1.2.11.). [If $\mathcal{U}$ is the base, associate to each p the collection $\mathcal{U}_p$ of all elements of $\mathcal{U}$ containing p.]

(c) Combine (a) and (b) to give an alternate proof of XIII.7.10.6.(a) for regular spaces.

XIII.7.10.8. (a) Let X be a T_1 topological space. Show that X is normal if and only if, whenever E is a closed set in X and U an open set in X containing E, there is a sequence (V_n) of open sets in X with

$\bar{V}_n \subseteq U$ for all n and $E \subseteq \cup_{n=1}^{\infty} V_n$. [If A and B are disjoint closed subsets of X, choose such sequences (V_n) for A in $X \setminus B$ and (W_n) for B in $X \setminus A$. Then

$$\bigcup_{n=1}^{\infty} \left(V_n \setminus \left[\bigcup_{k \leq n} \bar{W}_k \right] \right) \text{ and } \bigcup_{n=1}^{\infty} \left(W_n \setminus \left[\bigcup_{k \leq n} \bar{V}_k \right] \right)$$

are disjoint neighborhoods of A and B.]

(b) Use (a) to show that every F_σ in a normal space is normal.

XIII.7.10.9. Let X be a normal space.

(a) Use Urysohn's Lemma to prove that if E is a closed set in X and U is an open neighborhood of E, then there is an open F_σ V with $E \subseteq V \subseteq U$.

(b) Give an alternate proof of (a) not using Urysohn's Lemma by repeatedly applying XIII.7.6.2. to obtain a sequence (V_n) of open sets with $E \subseteq V_n \subseteq \bar{V}_n \subseteq V_{n+1} \subseteq U$.

(c) Extend the argument of (b) to show that if F is an F_σ in X and U an open neighborhood of F, there is an open F_σ V with $F \subseteq V \subseteq U$. [Write $F = \cup_n F_n$ with F_n closed, and successively apply XIII.7.6.2. to $F_n \cup \bar{V}_n$.]

(d) Extend the argument of (b) to show that if E is a closed set in X and U a neighborhood of E, there is an open F_σ V with $E \subseteq V \subseteq \bar{V} \subseteq U$. [First find an open set W with $E \subseteq W \subseteq \bar{W} \subseteq U$ and do the construction within W.] Does (c) similarly extend? [Consider the case where F is an open F_σ and $U = F$.]

Compare with XIII.22.8.2..

XIII.7.10.10. Let $\{X_i : i \in I\}$ be an indexed collection of (nonempty) sets with $X = \prod_{i \in I} X_i \neq \emptyset$. Prove the following without using the AC:

(o) X is T_0 if and only if each X_i is T_0.

(i) X is T_1 if and only if each X_i is T_1.

(ii) X is Hausdorff if and only if each X_i is Hausdorff.

(iii) X is regular if and only if each X_i is regular. [Use XIII.7.5.1..]

(iiis) X is completely regular if and only if each X_i is completely regular.

The corresponding statement for normal spaces is false (XIII.7.10.14.(h)).

XIII.7.10.11. Let X be a T_3 space, Y a dense subspace of X, and $p \in Y$. Suppose $\{V_i : i \in I\}$ is a local base for the relative topology of Y consisting of relatively open sets, and for each i U_i is an open set in X with $U_i \cap Y = V_i$ (to avoid use of the AC we may choose the U_i as in ()). Show that $\{U_i : i \in I\}$ is a local base for p in X, as follows.

(a) If U is an open set in X, then $U \cap Y$ is dense in U, so $U \subseteq Cl_X(U \cap Y)$. In particular, $U_i \subseteq Cl_X(V_i)$ for all i.

(b) Let W be an open neighborhood of p in X. Then there is an open neighborhood Z of p in X with $Cl_X(Z) \subseteq W$ (XIII.7.5.1.). There is an i such that $V_i \subseteq Z \cap Y$.

(c) For this i we have $U_i \subseteq Cl_X(V_i) \subseteq Cl_X(Z \cap Y) = Cl_X(Z) \subseteq W$.

(d) In particular, if p has a countable local base in Y, then it has a countable local base in X.

(e) Result (d) can fail if X is not T_3, even if it is T_1. [Let X be an uncountable set with the finite complement topology, and Y a countably infinite subset; cf. XIII.1.2.8.(iv).]

XIII.7.10.12. [Pea75] Make the following definitions:

DEFINITION. Let X be a topological space.

A subset Y of X is a *generalized* F_σ in X if, whenever U is an open neighborhood of Y in X, there is an F which is an F_σ in X with $Y \subseteq F \subseteq U$.

A subset Y of X is *normally situated* in X if, whenever U is an open neighborhood of Y in X, there is a locally finite cover $\{V_i : i \in I\}$ of Y with each V_i an open F_σ in X contained in U.

X is *totally normal* if X is normal and every subset of X is normally situated in X.

(a) Show that if X is normal, then every generalized F_σ in X is normally situated in X. [Use ().]

(b) Show that if X is normal and Y is normally situated in X, then Y is normal.

(c) Show that every subspace of a totally normal space is totally normal.

(d) Show that every totally normal space is completely normal (XIII.10.1.2.).

(e) Show that the space of ordinals less than or equal to the first uncountable ordinal (which is completely normal by XIII.16.1.1.) is not totally normal (cf. XIII.11.14.12.).

(f) Show that every perfectly normal space (XIII.10.6.11.) is totally normal. Give an example of a totally normal space which is not perfectly normal.

***(g) What is the relationship between total normality and (hereditary) paracompactness (XIII.11.10.3.)?

XIII.7.10.13. Show that a separable normal space cannot have a closed discrete subspace of cardinality $2^{\aleph_0}$. [Count the number of real-valued continuous functions and apply Urysohn's Lemma.] More generally, if X is normal and has a dense set of cardinality κ, and a closed discrete subspace of cardinality λ, then $2^\lambda \leq 2^\kappa$; in particular, X cannot have a closed discrete subspace of cardinality 2^κ (Jones's Lemma).

XIII.7.10.14. The Sorgenfrey Line. Put a topology $\mathcal{S}$ on $\mathbb{R}$ by taking $\{(a, b] : a, b \in \mathbb{R}, a < b\}$ as a base. (Many references use base sets of the form $[a, b)$, which gives a different but homeomorphic topology.) $(\mathbb{R}, \mathcal{S})$ is called the *Sorgenfrey line*, denoted S.

(a) Show that $\mathcal{S}$ is stronger than the usual topology on $\mathbb{R}$, and thus is Hausdorff.

(b) Show that S is zero-dimensional, hence normal. [Show that $(a, b]$ is clopen for all $a < b$.]

(c) Show that any bounded nondecreasing sequence in $\mathbb{R}$ converges in S to its limit in $\mathbb{R}$, but that no strictly decreasing sequence converges in S.

(d) Show that S is first countable and separable.

(e) Show that every subset A of S is separable and Lindelöf. [Show that

$$\{x \in A : (x - \epsilon, x) \cap A = \emptyset \text{ for some } \epsilon > 0\}$$

is countable. Then show that if $\mathcal{U}$ is an open cover of A, and V is the union of all intervals (a, b) for which $(a, b]$ is contained in some element of $\mathcal{U}$, then $A \setminus V$ is countable.] In particular, S is (hereditarily) paracompact.

(f) Show that every closed subset of S is a G_δ. Thus S is perfectly normal (XIII.10.6.11.), hence completely normal.

(g) Show that a subset of S is compact if and only if it is closed in the usual topology of $\mathbb{R}$, bounded above, and has no strictly decreasing sequences (i.e. is well ordered in the usual ordering). Conclude that every compact subset of S is countable (cf. ()). In particular, S is not locally compact or σ-compact.

(h) Show that $S \times S$ is not normal. [The set $\{(x, -x) : x \in S\}$ is a closed discrete subset of $S \times S$ of cardinality $2^{\aleph_0}$. Apply Exercise XIII.7.10.13.. Alternatively, show using the Baire Category Theorem that $\{(x, -x) : x \in \mathbb{Q}\}$ and $\{(x, -x) : x \in \mathbb{J}\}$ do not have disjoint neighborhoods.] So $S \times S$ is not paracompact or Lindelöf. Note that $S \times S$ is separable, first countable, and completely regular.

(i) Show that S is not second countable. [If it were, it would be metrizable, so $S \times S$ would also be metrizable. There is also a simple direct proof.]

(j) Show that S is a Baire space ().

(k) Give $\mathbb{R} \times \{0, 1\}$ the lexicographic ordering and the order topology. Show that S is homeomorphic to the subspace $\mathbb{R} \times \{0\}$ with the relative topology. Thus S is a suborderable space (), and hence completely collectionwise normal (XIII.16.1.1.). But S is not an orderable space.

The Sorgenfrey line provides a simple counterexample for many seemingly plausible statements, particularly about product spaces. If T is the subset of S consisting of irrational numbers, with the relative topology, it is plausible that T is homeomorphic to S since S is zero-dimensional; but in fact T is not homeomorphic to S [**?**].

XIII.7.10.15. Let $\mathcal{T}$ be the usual topology on $\mathbb{R}$ and $\mathcal{S}$ the Sorgenfrey topology (XIII.7.10.14.).

(a) Which functions from $\mathbb{R}$ to $\mathbb{R}$ are continuous from $(\mathbb{R}, \mathcal{S})$ to $(\mathbb{R}, \mathcal{T})$?

(b) Which functions from $\mathbb{R}$ to $\mathbb{R}$ are continuous from $(\mathbb{R}, \mathcal{T})$ to $(\mathbb{R}, \mathcal{S})$?

(c) Which functions from $\mathbb{R}$ to $\mathbb{R}$ are continuous from $(\mathbb{R}, \mathcal{S})$ to $(\mathbb{R}, \mathcal{S})$?

XIII.7.10.16. Let X be an uncountable set with the countable complement topology (XIII.1.6.4.).

(a) Show that X is T_1 but not Hausdorff.

(b) Show that a sequence in X converges if and only if it is eventually constant. Thus the convergent sequences in this topology are the same as the convergent sequences in the discrete topology.

(c) Every convergent sequence in X has a unique limit. Contrast with XIII.7.4.1..

XIII.7.10.17. **Smirnov's Deleted Sequence Topology** [SS95, #64]. Let $A = \left\{\frac{1}{n} : n \in \mathbb{N}\right\} \subseteq \mathbb{R}$. Let $\mathcal{T}$ be the usual topology on $\mathbb{R}$. Define a topology $\mathcal{S}$ on $\mathbb{R}$ by

$$\mathcal{S} = \{U \setminus B : U \in \mathcal{T}, B \subseteq A\} .$$

(a) Show that $\mathcal{S}$ is a topology on $\mathbb{R}$ which is stronger than $\mathcal{T}$.

(b) Show that $(\mathbb{R}, \mathcal{S})$ is Hausdorff.

(c) Show that $(\mathbb{R}, \mathcal{S})$ is not regular. [A is a closed set, $0 \notin A$, but 0 and A do not have disjoint neighborhoods.]

(d) Repeat the construction with A replaced by any nonclosed set with dense complement, e.g. $\mathbb{Q}$.

XIII.7.10.18. **The Tangent Disk Space** [SS95, Example 82]. Let X be the closed upper half plane in $\mathbb{R}^2$. Let $\mathcal{T}$ be the collection of all unions of subsets of X of the following types:

(i) Any subset of the open upper half plane which is open in the ordinary topology.

(ii) The union of $\{(a, 0)\}$ and an open disk of radius r $(r > 0)$ centered at (a, r), for some $a \in \mathbb{R}$ (an open disk in the upper half plane tangent to the x-axis at a).

(a) Show that $\mathcal{T}$ is a topology on X which is stronger than the relative Euclidean topology.

(b) Show that $(X, \mathcal{T})$ is separable and first countable.

(c) Show that $(X, \mathcal{T})$ is completely regular.

(d) Show that $(X, \mathcal{T})$ is not normal. [Show that the relative topology on the x-axis is the discrete topology, and apply XIII.7.10.13..]

This example is due to NIEMYTZKI [] and is called the *Niemytzki tangent disk space* or *Niemytzki plane*. Some references call it the *Moore plane* since it was apparently discovered independently by R. L. MOORE, but this name can be confused with the Moore manifold XIII.20.5.8., which is not the same (to make matters worse, the Moore manifold has been occasionally called the Niemytzki plane! But they are closely related: the Niemytzki plane is a quotient of the Moore manifold.)

XIII.7.10.19. Show that every infinite Hausdorff space contains a sequence of pairwise disjoint nonempty open sets. (There is a simple argument using DC. Some form of Choice is needed; cf. **??**.)

XIII.7.10.20. Let R be a unital ring (III.3.6.3.). Let $\mathrm{Spec}(R)$ be the set of prime ideals (II.6.4.12.) in R. Put a topology on $\mathrm{Spec}(R)$ as follows. If $E \subseteq \mathrm{Spec}(R)$, set

$$\ker(E) = \bigcap_{P \in E} P$$

and define

$$\bar{E} = \{P \in \mathrm{Spec}(R) : \ker(E) \subseteq P\} \ .$$

More generally, if $A \subseteq R$, define

$$\mathrm{hull}(A) = \{P \in \mathrm{Spec}(R) : A \subseteq P\}$$

so that $\bar{E} = \mathrm{hull}(\ker(E))$.

(a) Show that $E \mapsto \bar{E}$ is a Kuratowski closure operator (XIII.2.4.1.), hence defines a topology on $\mathrm{Spec}(R)$, called the *hull-kernel* or *Jacobson topology* on $\mathrm{Spec}(R)$ (if R is commutative, in a slightly different context it is also called the *Zariski topology*).

(b) Show that the hull-kernel topology is always T_0.

(c) Show that the hull-kernel topology on $\mathrm{Spec}(\mathbb{Z})$ is not T_1.

(d) Show that $\mathrm{Spec}(R)$ is compact in the hull-kernel topology.

(e) Let $\mathrm{Maxspec}(R)$ be the subset of $\mathrm{Spec}(R)$ consisting of the maximal ideals (II.6.4.12.) of R, with the hull-kernel topology (which is the relative topology). Show that $\mathrm{Maxspec}(R)$ is closed in $\mathrm{Spec}(R)$, hence also compact, and always T_1. Show that $\mathrm{Maxspec}(\mathbb{Z})$ is not Hausdorff.

(f) Explain why prime ideals must be used, i.e. what goes wrong if we try to take $\mathrm{Spec}(R)$ to be the set of all (proper) ideals in R?

$\mathrm{Spec}(R)$, or certain natural subsets like $\mathrm{Maxspec}(R)$, with the hull-kernel topology, is a fundamental object of study in several parts of mathematics, including Operator Algebras () and Algebraic Geometry ().

XIII.7.10.21. Other Separation Axioms. Several other separation axioms have been considered. Here are two of them.

DEFINITION. Let X be a topological space.

X is *completely Hausdorff*, or $T_{2\frac{1}{2}}$, if distinct points of X have disjoint closed neighborhoods.

X is *Urysohn* if, whenever $p, q \in X$, $p \neq q$, there is a continuous function $f : X \to [0, 1]$ with $f(p) = 0$ and $f(q) = 1$ (equivalently, there is a continuous $g : X \to \mathbb{R}$ with $g(p) \neq g(q)$).

(a) Show that the two conditions in the definition of an Urysohn space are equivalent.

(b) Show that a regular space is completely Hausdorff (use XIII.7.5.1.).

(c) Show that an Urysohn space is completely Hausdorff.

(d) Show that any topology stronger than a completely Hausdorff topology [*resp.* Urysohn topology] is completely Hausdorff [*resp.* Urysohn].

A completely Hausdorff space is obviously Hausdorff, and a completely regular space is Urysohn. There are Hausdorff spaces which are not completely Hausdorff (XIII.7.10.23.), in fact in which no two points have disjoint closed neighborhoods [SS95, Example 75]. There are Urysohn spaces which are not regular, hence not completely regular (XIII.7.10.17.), completely Hausdorff spaces which are neither regular nor Urysohn [SS95, Example 80], regular spaces which are not Urysohn [SS95, Example 90], and regular Urysohn spaces which are not completely regular [SS95, Example 91].

Both properties were introduced (the completely Hausdorff property somewhat more explicitly, but neither property was named) by URYSOHN in [Ury25a], where Urysohn's Lemma was proved and Urysohn's Metrization Theorem (XIII.10.1.3.) first stated.

There has been a bizarre reversal in terminology concerning these two axioms. Our definitions are the ones from [SS95], which were apparently universally used until that time (see e.g. [Bou98]). For unknown reasons, the definitions were reversed in [Wil04]: what we call "completely Hausdorff" is called "Urysohn" there, and vice versa. Since the publication of [Wil04], usage of the terms has been inconsistent but has tended to follow [Wil04]. The earlier definitions make more sense to me since "Urysohn functions" separating points or points and closed sets are standard terminology (although using "completely Hausdorff" in WILLARD'S sense would be in analogy with the term "completely regular," note that the "completely" in "completely normal" means something completely different). My guess is that WILLARD simply made a mistake, which has been perpetuated since; if he had deliberately changed established terminology he probably would have said so and given a reason, and I find it hard to believe he would worry about doing so in such a relatively minor side thread of topology. I have thus reverted to the historical definitions. It would be good if topologists would definitively decide what these definitions should be. There is one additional term sometimes used: spaces which are Urysohn according to our definition are sometimes called *functionally Hausdorff*. (In analogy, completely regular spaces could be called *functionally regular*; "functionally normal" is the same as normal by Urysohn's Lemma.)

XIII.7.10.22. Let $(X, \mathcal{T})$ be a topological space. Show (without using AC) that the following are equivalent:

(i) $(X, \mathcal{T})$ is Urysohn.

(ii) There is a completely regular topology $\mathcal{S}$ on X which is weaker than $\mathcal{T}$.

(iii) There is a strongest completely regular topology $\mathcal{S}$ on X which is weaker than $\mathcal{T}$.

[For (i) ⇒ (iii), mimic the argument of XIII.11.9.12. to embed X in a product Y of copies of $[0,1]$. Note that without AC, Y is not necessarily compact, but it is completely regular by XIII.7.10.10.(iiis).]

XIII.7.10.23. **Double Origin Topology** (Aleksandrov and Urysohn, cf. [SS95, Example 74]). Let $X = \mathbb{R}^2 \cup \{\mathbf{0}'\}$. Give X a topology $\mathcal{T}$ consisting of all unions of the following types of sets: any subset of $X \setminus \{\mathbf{0}, \mathbf{0}'\}$ which is open in $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ in the usual topology; the union of $\{\mathbf{0}\}$ and an open half-disk in the upper half plane centered at $\mathbf{0}$; the union of $\{\mathbf{0}'\}$ and an open half-disk in the lower half plane centered at $\mathbf{0}$.

(a) Show that $\mathcal{T}$ is a topology on X.

(b) Show that $(X, \mathcal{T})$ is Hausdorff.

(c) Show that $\mathbf{0}$ and $\mathbf{0}'$ do not have disjoint closed neighborhoods. Thus $(X, \mathcal{T})$ is not completely Hausdorff.

(d) Show that $(X, \mathcal{T})$ is path-connected.

XIII.7.10.24. Using AC, prove the following generalization of XIII.7.6.8.:

Theorem. Let X be a topological space. The following are equivalent:

(i) X is T_4.

(ii) Every point-finite open cover of X has a shrinking.

(iii) Every finite open cover of X has a shrinking.

(iv) Every finite open cover of X has a locally finite closed refinement.

(Note that (i) ⟺ (iii) is XIII.7.6.8..)

(a) For (i) ⇒ (ii), well-order the index set and use the proof of XIII.7.6.8. by transfinite induction. (This is where the AC is used, so the AC is not needed to prove that (i), (iii), and (iv) are equivalent.)

(b) For (iv) ⇒ (i), let A and B be disjoint closed subsets of X. Take a locally finite closed refinement $\mathcal{V}$ of $\{X \setminus A, X \setminus B\}$, and let E [resp. F] be the union of the sets in $\mathcal{V}$ disjoint from A [resp. B]. Show that E and F are closed, and take $U = X \setminus E$, $V = X \setminus F$. Show that U and V are disjoint neighborhoods of A and B.

(c) Deduce the following corollary:

Corollary. Let X and Y be topological spaces, and $f : X \to Y$ a continuous closed surjection. If X is T_4, then Y is also T_4. If X is normal, Y is also normal.

(d) Show that the result of (c) is not necessarily true if f is not closed. [Let Y be any topological space, and let X be the set Y with the discrete topology.]

XIII.7.10.25. (a) Let (X, ρ) be a metric space, and A, B disjoint closed subsets of X. Show that

$$f(x) = \frac{\rho(x, A)}{\rho(x, A) + \rho(x, B)}$$

is a well-defined continuous function from X to $[0,1]$ with $A = f^{-1}(0)$ and $B = f^{-1}(1)$.

(b) Show that Urysohn's Lemma for metrizable spaces can be proved without any Choice, and that a metrizable space is normal.

XIII.7.10.26. Say a topological space $(X, \mathcal{T})$ is $T_{1/2}$ (nonstandard definition) if every singleton subset of X is a Borel set (). A T_1 space is trivially $T_{1/2}$.

(a) If p and q are distinct points of X with the same open neighborhoods, show that any Borel set containing one of them contains the other [show the collection of such sets is a σ-algebra]. Conclude that a $T_{1/2}$ space is T_0.

(b) If X is T_0 and p and q are distinct points of X, show that there is a Borel set (either open or closed) containing p but not q. Conclude that every countable T_0 space is $T_{1/2}$. In particular, the Sierpiński space () is $T_{1/2}$ but not T_1.

(c) Show that every second countable T_0 space X is $T_{1/2}$. [If (U_n) is a countable base of open sets, and $p \in X$, for each n either $p \in U_n$ or $p \in U_n^c$. Let E_n be the one containing p. Then $\cap_n E_n = \{p\}$.]

(d) Put a topology on $\mathbb{R}$ as follows. Let $Y = \mathbb{R} \setminus \{0\}$. Let $\mathcal{T}$ consist of $\emptyset$ and all subsets of $\mathbb{R}$ containing 0 and all but finitely many points of Y. Then $\mathcal{T}$ is a T_0 topology on $\mathbb{R}$. The Borel sets in $(\mathbb{R}, \mathcal{T})$ are all countable subsets of Y and all sets containing 0 and all but countably many points of Y; thus $\{0\}$ is not a Borel set and $(\mathbb{R}, \mathcal{T})$ is not $T_{1/2}$.

(e) Is there a simple characterization of $T_{1/2}$ not using Borel sets or σ-algebras? Is a first countable T_0 space necessarily $T_{1/2}$, or conversely?

XIII.8. Quotient Spaces and the Quotient Topology

In this section, we will examine what happens when we take a topological space X and "collapse" various subsets of X to points.

Recall (II.3.3.9.) that there are three equivalent points of view for identifying subsets of a set X:

(i) Equivalence relations on X.

(ii) Partitions of X.

(iii) Functions from X onto another set Y.

We suppose the set X has a given topology, and want to topologize the set of equivalence classes, or the set Y, in an appropriate way. This topological space will be called the *quotient space* of X by the relation.

Many constructions in topology amount to forming a quotient space.

XIII.8.1. The Quotient Topology

We first take the point of view of having a function f from our set X onto another set Y. If X has a specified topology $\mathcal{T}$, we want to put a topology $\mathcal{T}_f$ on Y which makes f continuous, and which is "universal", i.e. the strongest possible. There is an obvious way to do this, since for f to be continuous $f^{-1}(V)$ must be in $\mathcal{T}$ for every $V \in \mathcal{T}_f$ (cf. XIII.5.1.14.).

XIII.8.1.1. DEFINITION. Let $(X, \mathcal{T})$ be a topological space, and $f : X \to Y$ a surjective function. The *quotient topology* on Y from f is the topology

$$\mathcal{T}_f = \{V \subseteq Y : f^{-1}(V) \in \mathcal{T}\} .$$

In other words, a subset of Y is open in the quotient topology if and only if its preimage in X is open. Equivalently, a subset of Y is closed in the quotient topology if and only if its preimage in X is closed.

It is easily checked () that the quotient topology is indeed a topology. The quotient topology on Y has a universal property. If g is a function from X to a set Z, and $g(x_1) = g(x_2)$ whenever $x_1, x_2 \in X$ and $f(x_1) = f(x_2)$, then there is an induced function $\bar{g} : Y \to Z$ with $g = \bar{g} \circ f$. The next result is obvious from the definition of the quotient topology.

XIII.8.1.2. PROPOSITION. Let X be a topological space, and f a surjective function from X to Y. Give Y the quotient topology. If Z is a topological space and $g : X \to Z$ factors through f, i.e. $g = \bar{g} \circ f$, then g is continuous if and only if $\bar{g}$ is continuous.

XIII.8.1.3. EXAMPLES. (i) Let π_1 be the projection of $\mathbb{R}^2$ onto the x-axis, regarded as $\mathbb{R}$, i.e. $\pi_1(x, y) = x$. It is easily seen that the quotient topology on $\mathbb{R}$ is the ordinary topology (preimages of sets in $\mathbb{R}$ are unions of vertical lines through points of the set). More generally, if X and Y are topological spaces, the projection from $X \times Y$ onto X (or onto Y) is a quotient map, i.e. the quotient topology on X is the given topology.

(ii) Let $X = [0, 1]$, Y the unit circle in $\mathbb{R}^2$, and $f(t) = (\cos 2\pi t, \sin 2\pi t)$. Then f is surjective, and one-to-one except at the endpoints. It is easy to check that the quotient topology on Y is the usual topology (relative topology from $\mathbb{R}^2$). Thus we can regard a circle as the quotient of a closed interval with the endpoints identified.

Not all quotient spaces are this nice (cf. XIII.8.1.11.).

XIII.8.1.4. The quotient topology is not the only topology on Y making f continuous in general; any weaker topology (e.g. the indiscrete topology) also will. But by XIII.8.1.2. (taking $Z = Y$ and $g = f$), the quotient topology is the *strongest* topology on Y making f continuous. For example, if $f : X \to X$ is the identity function, the quotient topology on X from f is precisely the original topology on X.

Quotients by Equivalence Relations

XIII.8.1.5. Now switch to the equivalent point of view of an equivalence relation $\sim$ on X. The set Y can be identified with the set of equivalence classes, and the function $f : X \to Y$ is the function $f(x) = [x]$. We will write $X/\sim$ for Y from this point of view, and call f the *projection*, *quotient map*, or *identification map*, of X onto $X/\sim$, which is called the *quotient space* or *identification space* of X by $\sim$. We give $X/\sim$ the quotient topology.

XIII.8.1.6. DEFINITION. Let $\sim$ be an equivalence relation on a set X.

(i) A subset A of X is *saturated* if it is closed under $\sim$, i.e. if $x \in A$ and $y \sim x$, then $y \in A$.

(ii) If $A \subseteq X$, the *saturation* of A is the set of all elements of X equivalent to an element of A.

XIII.8.1.7. It is obvious that the saturation of a subset of X is saturated, and is the smallest saturated set containing A. If $f : X \to X/\sim$ is the projection from X to $X/\sim$, the saturation of A is precisely $f^{-1}(f(A))$. More generally, a subset A of X is saturated if and only if $A = f^{-1}(B)$ for some subset B of $X/\sim$. In fact, $A = f^{-1}(f(A))$ if A is saturated.

A saturated set is just a union of equivalence classes. An arbitrary union, intersection, or complement of saturated sets is saturated.

XIII.8.1.8. The quotient map $f : X \to X/\sim$, besides being continuous, has the additional property that if U is a saturated open set in X, then $f(U)$ is open in $X/\sim$, or equivalently (since the complement of a saturated open set is a saturated closed set, and vice versa) if E is a saturated closed set in X, then $f(E)$ is closed in $X/\sim$. A surjective continuous map from a topological space X to Y with the quotient topology has this property for the equivalence relation defined by $x_1 \sim x_2$ if $f(x_1) = f(x_2)$.

It would be natural to give this type of function a name. Three names have been proposed and used in various references: (1) *strongly continuous*, first used in [AH74]; (2) *bicontinuous*, first used in [Kur66];

(3) *quasi-compact*, first used in [Why50]. None of the names have become generally established, and none is satisfactory since, among other reasons, they are all also sometimes used to mean other things. The term *quasi-compact* is the most unsatisfactory, since this term is sometimes used for spaces which are compact but not necessarily Hausdorff, and a mapping of the above type has nothing to do with quasi-compactness, and anyway the term *quasi-compact mapping* is commonly used to mean something else in algebraic geometry (more appropriately related to quasi-compactness). The term *strongly continuous* is also not appropriate, since it is too easily confused with the common use of "strongly continuous" in functional analysis (), and anyway it does not have the right intuitive connotation: a better term (which we do not seriously advocate) would be *barely continuous*, since the topology on Y is "just barely" weak enough to make the function continuous. The term *bicontinuous* is the least objectionable and we will use it, although it is also used to refer to simultaneous continuity with respect to two pairs of topologies (cf. [Kel63]). The term "bicontinuous" is especially appropriate in our context in the case where f is also injective: f is then bicontinuous if and only if both f and f^{-1} are continuous (i.e. f is a homeomorphism).

It is easily seen that, conversely, if X and Y are topological spaces and f is a surjective continuous map which is bicontinuous in the above sense, then the topology on Y is necessarily the quotient topology. Thus we obtain:

XIII.8.1.9. Proposition. Let X and Y be topological spaces and f a surjective function from X to Y. Then the topology on Y is the quotient topology from f (i.e. f is a quotient map) if and only if f is bicontinuous.

XIII.8.1.10. This proposition characterizes when a surjective continuous function is a quotient map. Since it is obvious that a surjective continuous function which is either open or closed is bicontinuous, it follows that any such map is a quotient map. But a bicontinuous map need not be either open or closed (XIII.8.1.11.).

A quotient space can be badly behaved:

XIII.8.1.11. Example. Let $\sim$ be the equivalence relation on $\mathbb{R}$ defined by $x \sim y$ if $x - y \in \mathbb{Q}$ (). Then each equivalence class is dense in $\mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}/\sim$ be the projection. If E is a nonempty closed set in $\mathbb{R}/\sim$, then $f^{-1}(E)$ is a saturated closed subset of $\mathbb{R}$. Since each equivalence class is dense, $f^{-1}(E)$ is dense in $\mathbb{R}$, hence $f^{-1}(E) = \mathbb{R}$, $E = \mathbb{R}/\sim$. Thus the quotient topology on $\mathbb{R}/\sim$ is the indiscrete topology.

The same argument works for any X and $\sim$ if all equivalence classes are dense in X.

XIII.8.2. Semicontinuous Decompositions

We want some conditions on $\sim$ which will give that the quotient map f is closed or open. These are easily describable:

XIII.8.2.1. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space, $\sim$ an equivalence relation on X, and $f : X \to X/\sim$ the quotient map. The following are equivalent:

(i) The saturation of every closed set in X is closed.

(ii) If p is any point of X, and U is an open neighborhood of $[p]$, then there is a saturated open set $V \subseteq X$ with $[p] \subseteq V \subseteq U$.

(iii) If A is any saturated subset of X, and U is an open neighborhood of A, then there is a saturated open set $V \subseteq X$ with $A \subseteq V \subseteq U$.

(iv) If U is an open set in X, the union of all saturated subsets of U is also open.

(v) f is a closed mapping.

An equivalence relation or partition of X satisfying these conditions is called an *upper semicontinuous (usc) decomposition* of X.

PROOF: (iii) $\Rightarrow$ (ii) is trivial, and for the converse, suppose A is saturated and U is an open neighborhood of A. For each $p \in A$ let V_p be a saturated open neighborhood of $[p]$ contained in U. Then $V = \cup_p V_p$ is a saturated open neighborhood of A contained in U.

(i) $\Rightarrow$ (ii): Let $p \in X$ and U an open neighborhood of $[p]$. Then U^c is closed, so the saturation B of U^c is closed. If $p \in B$, then there is a $q \in X$, $q \sim p$, with $q \in U^c$, contradicting that $[p] \subseteq U$. Thus $V = B^c$ is a saturated open set containing $[p]$ and contained in U.

(iii) $\Rightarrow$ (i): Let E be a closed subset of X, and B its saturation. Then $U = E^c$ is an open neighborhood of the saturated set $A = B^c$, so there is a saturated open neighborhood V of A contained in U. The saturated closed set V^c contains $U^c = E$ and is contained in $A^c = B$; since B is the saturation of E, we have $V^c = B$ and B is closed.

(i) $\Longleftrightarrow$ (iv): Let U be open. Then the saturation of U^c is precisely the complement of the union of all saturated subsets of U.

(i) $\Longleftrightarrow$ (v): If E is a closed set in X, and B is the saturation of E, then B is closed if and only if $f(B) = f(E)$ is closed. ☝

There is a similar characterization of when a quotient map is open, although there does not appear to be a clean analog of conditions (ii)–(iv).

XIII.8.2.2. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space, $\sim$ an equivalence relation on X, and $f : X \to X/\sim$ the quotient map. Then f is an open mapping if and only if the saturation of every open set in X is open.

An equivalence relation or partition of X satisfying these conditions is called a *lower semicontinuous (lsc) decomposition* of X.

XIII.8.2.3. More generally, if X and Y are topological spaces and 2^X is given the topology of (), a similar definition can be made for a function from Y to 2^X to be upper or lower semicontinuous. There is a connection with upper and lower semicontinuous functions from $\mathbb{R}$ to $\mathbb{R}$ (XIII.8.5.16.), which probably motivated the terminology for decompositions. See [Kur66, p. 174] or [Kel75, 3F].

A decomposition which is both upper and lower semicontinuous is called *continuous*. A decomposition is continuous if and only if the quotient map is both open and closed.

Combining these characterizations with XIII.8.1.10., we obtain:

XIII.8.2.4. Proposition. Let X and Y be topological spaces, and $f : X \to Y$ a continuous surjection. If f is closed [open], then f is a quotient map and the corresponding decomposition of X is usc [lsc].

XIII.8.2.5. Examples. (i) The decomposition of XIII.8.1.3.(i) is lower semicontinuous but not upper semicontinuous: if E is the closed set $\{(x, y) : xy = 1\}$ of $\mathbb{R}^2$, the saturation of E is $\mathbb{R}^2$ with the y-axis removed, and $\pi_1(E) = \mathbb{R} \setminus \{0\}$, which are not closed. The projection π_1 is open ().

(ii) The decomposition of XIII.8.1.3.(ii) is upper semicontinuous but not lower semicontinuous. Upper semicontinuity is obvious since only one equivalence class has more than one point. But the saturation and image of the open set $\left[0, \frac{1}{2}\right)$ is not open.

(iii) Let π_1 be projection of $[0, 1]^2$ onto the first coordinate. Then the corresponding decomposition of $[0, 1]^2$ into vertical line segments is continuous (), ().

(iv) The decomposition of XIII.8.1.11. is neither upper nor lower semicontinuous.

XIII.8.3. Separation Axioms for Quotient Spaces

Since a primary use of quotient spaces is to construct new topological spaces, it is important to know how properties of X, especially separation properties, pass to $X/\sim$. Example XIII.8.1.11. shows that nothing is preserved in general; however, under reasonable and verifiable conditions separation properties do pass to quotients.

We begin with a simple observation:

XIII.8.3.1. Proposition. Let $\sim$ be an equivalence relation on a topological space X. Then $X/\sim$ is a T_1 space if and only if each equivalence class is closed.

XIII.8.3.2. Corollary. Let $\sim$ be an equivalence relation on a T_1 topological space X, and $f : X \to X/\sim$ the quotient map. If f is closed (i.e. if the decomposition is usc), then $X/\sim$ is T_1.

Proof: If $p \in X$, then $\{p\}$ is closed in X, and hence its saturation $[p]$ is also closed. 👍

An important variation is that the equivalence relation is closed:

XIII.8.3.3. Definition. Let X be a topological space. A relation R on X is *closed* if it is a closed subset of $X \times X$ with the product topology.

More generally, we can define closed relations from one topological space to another.

XIII.8.3.4. If $\sim$ is an equivalence relation, the condition that $\sim$ is closed implies that each equivalence class is closed. But the conditions are not equivalent: for example, if X is a T_1 space which is not Hausdorff, the equality relation on X has closed equivalence classes (singletons) but is not a closed relation (i.e. the diagonal is not closed in $X \times X$, cf. ()).

This example shows that for an equivalence relation, being closed is not closely related to being upper or lower semicontinuous: the equality relation on a topological space is always both upper and lower semicontinuous, but only closed if the space is Hausdorff. Conversely, the equivalence relations of XIII.8.2.5.(i)-(ii) are closed, but the decompositions fail to be lsc and usc respectively.

XIII.8.3.5. Proposition. Let $\sim$ be an equivalence relation on a topological space X. If $X/\sim$ is Hausdorff, then $\sim$ is closed.

Proof: Suppose $\sim$ is not closed. Then there is a net $((x_i, y_i))$ in $X \times X$ converging to an $(x, y) \in X \times X$ with $x_i \sim y_i$ for all i but $x \not\sim y$. If $f : X \to X/\sim$ is the quotient map, then the net $(f(x_i)) = (f(y_i))$ in $X/\sim$ in $X/\sim$ converges to both $f(x)$ and $f(y)$, so $X/\sim$ is not Hausdorff.

The converse of XIII.8.3.5. is not true in general. But we have:

XIII.8.3.6. Proposition. Let $\sim$ be an equivalence relation on a topological space X, and $f : X \to X/\sim$ the quotient map. If f is open (i.e. the decomposition is lsc) and $\sim$ is closed, then $X/\sim$ is Hausdorff.

Proof: Let $y_1, y_2 \in X/\sim$ and $x_1, x_2 \in X$ with $f(x_1) = y_1$, $f(x_2) = y_2$. Then $x_1 \not\sim x_2$, so, since $\sim$ is closed, there are open neighborhoods V and W of x_1 and x_2 such that $V \times W$ is disjoint from the relation $\sim$. Then $f(V)$ and $f(W)$ are disjoint open neighborhoods of y_1 and y_2.

XIII.8.3.7. The converse of XIII.8.3.6. is also false in general: in XIII.8.2.5.(ii), the quotient space is Hausdorff but the decomposition is not lsc. We also have:

XIII.8.3.8. Proposition. Let $\sim$ be an equivalence relation on a T_4 topological space X, and $f : X \to X/\sim$ the quotient map. If f is closed (i.e. if the decomposition is usc), then $X/\sim$ is T_4.

Proof: Let E_1 and E_2 be disjoint closed subsets of $X/\sim$. Then $f^{-1}(E_1)$ and $f^{-1}(E_2)$ are disjoint closed subsets of X, so they have disjoint open neighborhoods V and W. We have that V^c and W^c are closed and $V^c \cup W^c = X$, so $f(V^c)$ and $f(W^c)$ are closed in $X/\sim$ and $f(V^c) \cup f(W^c) = X/\sim$. Thus $U_1 = [f(V^c)]^c$ and $U_2 = [f(W^c)]^c$ are disjoint open sets. Since $f^{-1}(E_1) \subseteq V$, $f(V^c)$ is disjoint from E_1 and $E_1 \subseteq U_1$. Similarly, $E_2 \subseteq U_2$.

XIII.8.3.9. Corollary. Let $\sim$ be an equivalence relation on a normal topological space X, and $f : X \to X/\sim$ the quotient map. If f is closed (i.e. if the decomposition is usc), then $X/\sim$ is normal and $\sim$ is closed.

Proof: Combine XIII.8.3.8. and XIII.8.3.2. to conclude that $X/\sim$ is normal. Since $X/\sim$ is Hausdorff, $\sim$ is closed by XIII.8.3.5.. 👍

The Maximal T_0 Quotient

We have so far skipped the T_0 condition. There is an important result here:

XIII.8.3.10. Theorem. Let X be a topological space. Then there is a maximal T_0 quotient $\bar{X}$, a quotient space of X, such that:

(i) $\bar{X}$ is T_0.

(ii) If Y is any T_0 space and $f : X \to Y$ a continuous function, then there is a unique continuous function $\bar{f} : \bar{X} \to Y$ such that $f = \bar{f} \circ \pi_0$, where $\pi_0 : X \to \bar{X}$ is the quotient map.

(iii) The quotient map π_0 is open and closed, i.e. the partition of X is continuous.

(iv) If A is any subset of X and $B = \pi_0^{-1}(\pi_0(A))$ is the saturation of A, then

$$\bar{A} = \bar{B} = \pi_0^{-1}(\overline{\pi_0(A)}) = \pi_0^{-1}(\pi_0(\bar{A}))$$

and $\overline{\pi_0(A)} = \pi_0(\bar{A})$. Every closed set and every open set in X is saturated.

The quotient space $\bar{X}$ is uniquely determined up to bijective equivalence by conditions (i) and (ii). The space $\bar{X}$ is $X/\sim$, where $\sim$ is the equivalence relation on X defined by $x \sim y$ if $\overline{\{x\}} = \overline{\{y\}}$.

Proof: We verify that the quotient by the given equivalence relation has properties (i)–(iv). Let $A \subseteq X$, and $B = \pi_0^{-1}(\pi_0(A))$. Let $p \in \bar{B}$. If U is any open neighborhood of p, then there is a $y \in U \cap B$. Then $y \sim x$ for some $x \in A$, so $y \in \overline{\{x\}}$. Thus any neighborhood of y contains x, and in particular $x \in U$. Thus every neighborhood of p contains a point of A, so $p \in \bar{A}$, i.e. $\bar{B} \subseteq \bar{A}$, and the opposite inclusion is trivial. Thus $\bar{A} = \bar{B}$.

If A is closed, then we have $A = \bar{A} = \bar{B}$, so $B \subseteq A$, and the opposite inclusion is trivial, so A is saturated. Since the complement of a saturated set is saturated, every open set is saturated. Thus (iii) is an immediate consequence.

If A is again a general subset of X, then $\pi_0(\bar{A})$ is a closed set containing $\pi_0(A)$, so $\overline{\pi_0(A)} \subseteq \pi_0(\bar{A})$, and the opposite inclusion holds by continuity. Since $\bar{A}$ is saturated, $\bar{A} = \pi_0^{-1}(\pi_0(\bar{A}))$. Thus (iv) is proved.

We then easily obtain (i): if $x, y \in X$ and $\overline{\{\pi_0(x)\}} = \overline{\{\pi_0(y)\}}$, we have

$$\overline{\{x\}} = \pi_0^{-1}(\overline{\{\pi_0(x)\}}) = \pi_0^{-1}(\overline{\{\pi_0(y)\}}) = \overline{\{y\}}$$

so $x \sim y$ and $\pi_0(x) = \pi_0(y)$. Apply XIII.7.2.1..

To show (ii), suppose f is a continuous function from X to a T_0 space Y. It suffices to show that if $x_1, x_2 \in X$ and $x_1 \sim x_2$, then $f(x_1) = f(x_2)$. Suppose $f(x_1) \neq f(x_2)$. Since Y is T_0, we may assume without

loss of generality that $f(x_1)$ has an open neighborhood V not containing $f(x_2)$. But then $f^{-1}(V)$ is an open neighborhood of x_1 not containing x_2, so $x_1 \not\sim x_2$.

Finally, we verify uniqueness; in fact, we show slightly more. We use the standard uniqueness argument for universal properties (). Let $\tilde{X}$ be a T_0 space, and $\tilde{\pi} : X \to \tilde{X}$ a continuous function with the universal property of (ii). Then there is a (unique) $f : \bar{X} \to \tilde{X}$ such that $\tilde{\pi} = f \circ \pi_0$, and a (unique) $g : \tilde{X} \to \bar{X}$ such that $\pi_0 = g \circ \tilde{\pi}$. But then $(g \circ f) \circ \pi_0 = \pi_0$, so $g \circ f$ is the unique $h : \bar{X} \to \bar{X}$ such that $\pi_0 = h \circ \pi_0$. But the identity map $\iota_{\bar{X}}$ also has this property, so $g \circ f = \iota_{\bar{X}}$. Similarly, $f \circ g = \iota_{\tilde{X}}$, so $g = f^{-1}$ is a homeomorphism. 👍

This result is due to KOLMOGOROV, and the maximal T_0 quotient is sometimes called the *Kolmogorov quotient*. There is also a maximal Hausdorff quotient (XIII.8.5.6.).

Other Properties Preserved in Quotients

We now discuss some properties other than separation properties.

XIII.8.3.11. Any property preserved under all surjective continuous maps, e.g. separability and compactness, is of course preserved under arbitrary quotient maps.

For first and second countability, we need the decomposition to be lsc, i.e. the quotient map is open:

XIII.8.3.12. PROPOSITION. Let $\sim$ be an equivalence relation on a topological space X, and $f : X \to X/\sim$ the quotient map. If f is open (i.e. the decomposition is lsc) and X is first countable [second countable], then $X/\sim$ is also first countable [second countable].

PROOF: Let $p \in X$, and $\{U_n : n \in \mathbb{N}\}$ a countable base at p. Then $\{f(U_n)\}$ is a collection of open neighborhoods of $f(p)$. If V is an open neighborhood of $f(p)$, then $f^{-1}(V)$ is an open neighborhood of p, so there is an n such that $U_n \subseteq f^{-1}(V)$. Then $f(U_n) \subseteq V$ and $\{f(U_n) : n \in \mathbb{N}\}$ is a countable base at $f(p)$.

If $\{U_n : n \in \mathbb{N}\}$ is a countable base for the topology of X, then by a nearly identical argument $\{f(U_n) : n \in \mathbb{N}\}$ is a countable base for the topology of $X/\sim$. 👍

Compactness and Local Compactness

XIII.8.3.13. If X is compact and Y is Hausdorff, and $f : X \to Y$ is continuous, then f is automatically a closed mapping (XIII.11.5.1.); so if f is surjective it is a usc quotient map (XIII.8.2.4.) and Y is also compact. But f need not be lsc (i.e. an open map); see XIII.11.14.30.. Conversely, if X is compact Hausdorff and $X \to Y$ is a usc quotient map, then Y is compact Hausdorff (XIII.8.3.8.); see XIII.9.1.2. for an application.

XIII.8.4. Quotient Metrics

A construction related to, but not in general identical to, the quotient space construction is the construction of a quotient (pseudo)metric.

XIII.8.4.1. Let (X, ρ) be a metric space and $\sim$ an equivalence relation on X, and let $Y = X/\sim$ be the set-theoretic quotient. We want to put a "quotient metric" $\bar{\rho}$ on Y. As a first try, we could consider setting

$$\bar{\rho}([x],[y]) = \inf\{\rho(x', y') : x' \in [x], y' \in [y]\} .$$

This function is well-defined, nonnegative, and symmetric, but easy examples (e.g. ()) show that $\bar{\rho}$ defined this way usually does not satisfy the triangle inequality (and it is not definite in general, i.e. could only be a pseudometric).

There is a way to modify this definition to give a pseudometric (Exercise XIII.8.5.15.), but there is a very slick way to define what is obviously the right candidate for the quotient metric:

XIII.8.4.2. Definition. Let (X, ρ) be a metric space, and $\sim$ an equivalence relation on X. Define $\beta : X \times X \to [0, +\infty)$ by

$$\beta(x, y) = \begin{cases} \rho(x, y) & \text{if} \quad x \not\sim y \\ 0 & \text{if} \quad x \sim y \end{cases} .$$

Let ρ_β be the maximum pseudometric of β (VI.2.5.7.), and set $\bar{\rho}([x],[y]) = \rho_\beta(x, y)$. Then $\bar{\rho}$ is the *quotient pseudometric* on $X/\sim$.

It is not hard to see that $\bar{\rho}$ is well defined (Exercise XIII.8.5.16.), and it is then obviously a pseudometric.

XIII.8.4.3. Examples. (i) Let $X = [0, 1]$ with the usual metric, and $\sim$ the equivalence relation obtained by identifying 0 and 1 (i.e. $0 \sim 1$ and the other equivalence classes are singletons). Then $X/\sim$ is a circle of circumference 1, and the quotient pseudometric is the arc length metric.

(ii) Let $X = [0, +\infty)$ with the usual metric ρ, and $\sim$ the equivalence relation obtained by identifying $\mathbb{N} \cup \{0\}$ to a point, i.e. the equivalence classes are $\mathbb{N} \cup \{0\}$ and singletons. $(X/\sim, \bar{\rho})$ is homeomorphic to the space of XIII.19.5.9..

(iii) Let $X = (0, 1]$ with the usual metric ρ, and $\sim$ the equivalence relation obtained by identifying $N = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$ to a point, i.e. the equivalence classes are N and singletons. $(X/\sim, \bar{\rho})$ is homeomorphic to the Hawaiian earring (XIII.19.5.8.).

(iv) Let $X = \mathbb{R}$ with the usual metric ρ, and $\sim$ the equivalence relation whose classes are the cosets of $\mathbb{Q}$. Every equivalence class is dense, so $\bar{\rho}$ is the zero pseudometric.

(v) Let $X = (0, 1)$ with the usual metric ρ. Let $\sim$ be the equivalence relation obtained by identifying x with $\frac{x}{2}$ for each $x \in (0, 1)$. Then each equivalence class is closed, but we still have that $\bar{\rho}$ is the zero pseudometric.

XIII.8.4.4. Examples (iv) and (v) show that $\bar{\rho}$ is not always a metric if ρ is, even if all equivalence classes are closed. Since $\bar{\rho}([x],[y]) \leq \rho(x, y)$ for all $x, y \in X$, the quotient map from (X, ρ) to $(X/\sim, \bar{\rho})$ is continuous, i.e. the $\bar{\rho}$ topology on $X/\sim$ is weaker than the quotient topology. It is the quotient topology in good cases (e.g. (i), (iv)), but strictly weaker in general (e.g. in (ii), (iii), (v)). The $\bar{\rho}$ topology can depend on the exact metric and not just on the topology: examples (ii) and (iii) are homeomorphic respecting the equivalence relation, but the quotient metric topologies are not the same.

XIII.8.4.5. But if X is metrizable and $X/\sim$ is also metrizable in the quotient topology (cf. ()), there is a metric on X giving its topology such that the quotient metric gives the quotient topology. For if σ is a metric on X giving its topology and τ a metric on $X/\sim$ giving its topology, and $\pi : X \to X/\sim$ is the quotient map, then $\rho = \sigma + (\tau \circ \pi)$ is a metric on X which is equivalent to σ, and the $\bar{\rho}$-topology on $X/\sim$ is stronger than than the τ-topology and weaker than the quotient topology, hence they all coincide, i.e. $\bar{\rho}$ is equivalent to τ.

XIII.8.5. Exercises

XIII.8.5.1. Let X, Y, Z be topological spaces, and $f : X \to Y$ and $g : Y \to Z$ surjective continuous functions. Set $h = g \circ f : X \to Z$.

(a) Show that if f and g are quotient maps, then h is also a quotient map.

(b) Show that if h is a quotient map, then g is also a quotient map. Is f necessarily a quotient map?

XIII.8.5.2. (a) Let X and Y be topological spaces, and let π_1 be projection of $X \times Y$ onto X. Show that π_1 is a quotient map and that the corresponding decomposition of $X \times Y$ is lsc.

(a) Let $\{X_i : i \in I\}$ be a family of topological spaces, and let $X = \prod_{i\in I} X_i$. For $j \in I$, let π_j be projection of X onto X_j. Show that π_j is a quotient map and that the corresponding decomposition of X is lsc.

XIII.8.5.3. Let $\sim$ be an equivalence relation on a topological space X, and $f : X \to X/\sim$ the quotient map. If $Z \subseteq X$, let $\sim_Z$ be the restriction of $\sim$ to Z. Is $Z/\sim_Z$ homeomorphic to $f(Z)$ in general? What if Z is open or closed in X? What if the decomposition is usc or lsc?

XIII.8.5.4. Let X_1, X_2, Y_1, Y_2 be topological spaces, and $f : X_1 \to Y_1$ and $g : X_2 \to Y_2$ be quotient maps. Define $h : X_1 \times X_2 \to Y_1 \times Y_2$ by $h(x_1, x_2) = (f(x_1), g(x_2))$. Is h necessarily a quotient map? What if f and g are open or closed? (cf. [Mun75, p. 141-143].)

XIII.8.5.5. Show that the maximal T_0 quotient of a topological space X can be described alternately as follows.

(a) Define an equivalence relation $\approx$ on X by $x_1 \approx x_2$ if $f(x_1) = f(x_2)$ for every continuous function f from X to a T_0 space Y_f (which can depend on f). Show that $X/\approx$ is T_0. Conclude that $\approx = \sim$.

(b) Describe $\bar{X}$ explicitly by considering the map π from X to $\prod_f Y_f$ by $(\pi(x))_f = f(x)$ (show that this can be done without set-theoretic difficulties by choosing a suitable representative set of functions). Set $Y = \pi(X)$. Show that $Y \cong \bar{X}$ and that π_0 can be identified with π, regarded as a map from X to Y.

XIII.8.5.6. **The Maximal Hausdorff Quotient**. Let X be a topological space. Prove the following theorem:

THEOREM. Let X be a topological space. Then there is a maximal Hausdorff quotient $\tilde{X}$, a quotient space of X, such that:

(i) $\tilde{X}$ is Hausdorff.

(ii) If Y is any Hausdorff space and $f : X \to Y$ a continuous function, then there is a unique continuous function $\tilde{f} : \tilde{X} \to Y$ such that $f = \tilde{f} \circ \pi_2$, where $\pi_2 : X \to \tilde{X}$ is the quotient map.

The quotient space $\tilde{X}$ is uniquely determined up to bijective equivalence by conditions (i) and (ii).

(a) Define an equivalence relation $\approx$ on X by $x_1 \approx x_2$ if $f(x_1) = f(x_2)$ for every continuous function f from X to a Hausdorff space Y_f (which can depend on f). Show that $\tilde{X} = X/\approx$ is Hausdorff. Describe $\tilde{X}$ as in XIII.8.5.5.(b). Give the standard proof of uniqueness.

(b) It is not easy to explicitly describe the equivalence relation $\approx$. As a first approximation, define $\sim$ on X by $x \sim y$ if x and y do not have disjoint neighborhoods. Is $\sim$ an equivalence relation?

(c) Extend $\sim$ to an equivalence relation by taking its transitive closure. Is $X/\sim$ necessarily Hausdorff? If it is, it must be (up to bijective equivalence) the $\tilde{X}$ of (a).

(d) If $X/\sim$ is not Hausdorff, iterate the construction. Let $\sim'$ be the relation on $\mathcal{X}/\sim$ defined in the same way as $\sim$. Write $\sim_1 = \sim$ on X, and let $\sim_2$ be the equivalence relation on X corresponding to the relation $\sim'$ on $X/\sim$. Describe $\sim_2$ explicitly.

(e) Transfinitely define an increasing collection $\sim_\alpha$ of equivalence relations on X in this way. Show that this collection stabilizes at some β, and that $X/\sim_\beta$ is Hausdorff and is bijectively equivalent to $\tilde{X}$.

The smallest such β is called the *Hausdorff height* of X and is an ordinal invariant of X.

The maximal Hausdorff quotient of X is sometimes called the *Hausdorffization* or *Hausdorffification* of X. See () for details.

XIII.8.5.7. (a) Is there a maximal T_1 quotient of a topological space X along the same lines as the maximal T_0 quotient and the maximal Hausdorff quotient?

(b) Show there is no maximal regular, completely regular, normal, or completely normal quotient in general. [There are nonregular topologies on $\mathbb{R}$ which are stronger than the usual topology (e.g. XIII.7.10.17.).]

XIII.8.5.8. There are $2^{\aleph_0}$ pairwise disjoint dense subsets of $\mathbb{R}$ [this can be proved without any form of choice, e.g. by taking rational translates of the von Neumann numbers (III.8.4.13.). See also XIII.8.5.9..] Let $\{A_t : t \in \mathbb{R}\}$ be such a collection. We may assume $\cup_t A_t = \mathbb{R}$ by expanding one of the sets if necessary. Set

$$X = \{(t, y) \in \mathbb{R}^2 : y \in A_t\} \subseteq \mathbb{R}^2 \ .$$

Then projection π_2 of X onto the second coordinate is a bijection from X onto $\mathbb{R}$. The projection π_1 is a surjective but not injective function from X to $\mathbb{R}$. Let $\mathcal{E}$ denote the usual topology on $\mathbb{R}$.

(a) Let $\mathcal{T}$ be any topology on $\mathbb{R}$. Let $\mathcal{T}_X$ be the relative topology on X of the product topology $\mathcal{T} \otimes \mathcal{E}$ on $\mathbb{R} \times \mathbb{R}$. Show that $\mathcal{T}_X$ is a Hausdorff topology on X. [Distinct points of X have distinct second coordinates.]

(b) If $\mathcal{T}$ is any topology on $\mathbb{R}$, let $\mathcal{H}(\mathcal{T})$ be the topology on $\mathbb{R}$ obtained by transferring $\mathcal{T}_X$ to $\mathbb{R}$ via the bijection π_2. Then $\mathcal{H}(\mathcal{T})$ is a Hausdorff topology on $\mathbb{R}$ which is stronger than $\mathcal{E}$. If $\mathcal{T}_1$ and $\mathcal{T}_2$ are different topologies on $\mathbb{R}$, then $\mathcal{H}(\mathcal{T}_1)$ and $\mathcal{H}(\mathcal{T}_2)$ are different topologies on $\mathbb{R}$. Thus there are enormously many distinct (Hausdorff) topologies on $\mathbb{R}$ stronger than $\mathcal{E}$.

(c) The surjective map $\pi_1 \circ \pi_2^{-1}$ is continuous and open from $(\mathbb{R}, \mathcal{H}(\mathcal{T}))$ to $(\mathbb{R}, \mathcal{T})$.

(d) Show that $(\mathbb{R}, \mathcal{T})$ has the quotient topology from $(\mathbb{R}, \mathcal{H}(\mathcal{T}))$ under this map.

(e) If Y is any topological space of cardinality $\leq 2^{\aleph_0}$, then Y can be identified as a subset of $\mathbb{R}$ with some relative topology. Thus Y is a continuous open image of a Hausdorff space.

This can be extended to topological spaces of arbitrary cardinality (XIII.8.5.9.).

XIII.8.5.9. (a) Let κ be a cardinal. Then there is a compact Hausdorff space Z which has a collection of κ pairwise disjoint dense subsets. [Let $Z = \{0,1\}^\kappa$, and consider the equivalence relation on Z where two points are equivalent if they differ in only finitely many coordinates. Use II.9.9.14.]

(b) If Y is any topological space of cardinality κ, let Z be as in (a) with a collection $\{A_y : y \in Y\}$ of pairwise disjoint dense subsets with $\cup_y A_y = Z$. Set

$$X = \{(y,z) : z \in A_y\} \subseteq Y \times Z \ .$$

Repeat problem XIII.8.5.8. with this X to show that Y is a continuous open image (quotient) of a Hausdorff space.

XIII.8.5.10. Let Y be the Sierpiński space. Define $f : [0,1] \to Y$ by $f(t) = a$ if $t < 1$ and $f(1) = b$. Show that f is a quotient map. However, f is neither open nor closed. Use XIII.8.5.8. to write Y as a quotient of a Hausdorff space by an open map; see XIII.7.10.17.(d). (It cannot be done by a closed map since the quotient of a T_1 space by a closed map is T_1.)

XIII.8.5.11. Let X and Y be topological spaces, and $f : X \to Y$ a surjective continuous map. Say f is *semiclosed* if the image of the closure of every saturated set in X is closed in Y.

(a) Show that f is semiclosed if and only if, for every subset B of Y, we have

$$\bar{B} = f(\overline{f^{-1}(B)}) \ .$$

(b) Show that every semiclosed map is a quotient map. In fact, f is semiclosed if and only if, for every subset B of Y, $f|_{f^{-1}(B)}$ is a quotient map from $f^{-1}(B)$ onto B. Thus a semiclosed map is also called a *hereditary quotient map*.

(c) Show that every (surjective) closed map is semiclosed.

(d) Show that all the examples of quotient maps in this section (except XIII.8.5.12.) are semiclosed. In particular, a semiclosed map need not be closed.

(e) If $f : X \to Y$ is a quotient map and Y is Hausdorff and first countable, then f is semiclosed.

(f) Not every quotient map is semiclosed (XIII.8.5.12.).

(g) Show that if X and Y are first countable and Y is T_1, $f : X \to Y$ is semiclosed, (y_n) is a sequence in Y, and $y \in Y$, then $y_n \to y$ if and only if for every subsequence of (y_n) there is a subsubsequence (y_{k_n}) for which there is a sequence (x_{k_n}) in X and an $x \in X$ with $x_{k_n} \to x$, $f(x_{k_n}) = y_{k_n}$ for all n, and $f(x) = y$. In other words, convergent sequences in Y have subsequences which can be lifted to convergent sequences in X.

This result is quite useful in identifying and describing quotient topologies. The restriction that Y be T_1, or something like it, is necessary (XIII.8.5.13.). And even if X and Y are compact metrizable spaces (where the result in (g) is almost trivial), and f is closed, passage to subsequences is necessary in general (XIII.8.5.14.).

XIII.8.5.12. Let $X = \{a_1, a_2, b_1, b_2, c_1, c_2, x_1, x_2\}$, $Y = \{a, b, c, y\}$, and $f : X \to Y$ with $f(a_j) = a$, $f(b_j) = b$, $f(c_j) = c$, $f(x_j) = y$ for $j = 1, 2$. Give X the topology generated by $\{b_1, x_1\}$ and $\{c_2, x_2\}$.

(a) Show that the quotient topology on Y is the indiscrete topology.

(b) Show that y is in the closure of $\{a\}$ in Y, but neither x_1 nor x_2 is in the closure in X of $f^{-1}(\{a\}) = \{a_1, a_2\}$. Thus f is not semiclosed.

XIII.8.5.13. Let $X = \{x_n, z_n : n \in \mathbb{N}\}$, $Y = \{y\} \cup \{y_n : n \in \mathbb{N}\}$, $f : X \to Y$ with $f(x_n) = y_n$ and $f(z_n) = y$ for all n. Give X the topology generated by $\{\{x_n\}, \{x_n, z_n\} : n \in \mathbb{N}\}$, and give Y the quotient topology.

(a) Show that X and Y are first countable but not T_1, and that the quotient map f is semiclosed.

(b) Show that $y_n \to y$ in Y, but no subsequence of (y_n) lifts to a convergent sequence in X.

XIII.8.5.14. Let X be the compact subset

$$\left\{\left(\frac{1}{n}, 1\right) : n \in \mathbb{N} \text{ odd}\right\} \cup \left\{\left(\frac{1}{n}, 0\right) : n \in \mathbb{N} \text{ even}\right\} \cup \{(0,0), (0,1)\}$$

of $\mathbb{R}^2$, and $Y = \{0\} \cup \left\{\frac{1}{n} : n \in \mathbb{N}\right\} \subseteq \mathbb{R}$. Define $f : X \to Y$ by $f(x, y) = x$.

(a) Show that f is a quotient map which is closed, hence semiclosed.

(b) Show that the convergent sequence $y_n = \frac{1}{n}$ in Y does not lift to a convergent sequence in X.

XIII.8.5.15. Let (X, ρ) be a metric space, and $\sim$ an equivalence relation on X. Show that the quotient metric $\bar{\rho}$ can be explicitly described as

$$\bar{\rho}(a, b) = \inf \sum_{k=1}^{n} \rho(x_k, y_k)$$

where the infimum is over all finite sets $\{x_1, \ldots, x_{n+1}, y_1, \ldots, y_n\}$, where $x_1 = a$, $x_{n+1} = b$, and $y_k \sim x_{k+1}$ for $1 \leq k \leq n$ (n may vary). The definition does not change if we only require that $x_1 \sim a$, along with the other conditions.

XIII.8.5.16. Let $X, \rho, \sim, \beta$ be as in XIII.8.4.2.. Use the triangle inequality to show that if σ is a pseudometric on X with $\sigma \leq \rho_\beta$, then σ drops to a well-defined pseudometric on $X/\sim$. In particular, $\bar{\rho}$ is well defined.

XIII.8.5.17. Let $f : \mathbb{R} \to \mathbb{R}$ be a function, and $X = \{(x, y) \in \mathbb{R}^2 : y \leq f(x)\}$ be the subgraph. Give X the relative topology from $\mathbb{R}^2$. Let $\sim$ be the equivalence relation on X from projection onto the X-axis (i.e. the decomposition partitions X into vertical rays). Show that the decomposition is usc [resp. lsc] if and only if f is upper semicontinuous [resp. lower semicontinuous] in the sense of (). [XIII.11.14.19. is helpful.]

XIII.9. Identification Spaces

Perhaps the most important method of construction of topological spaces is by identification or "gluing." There are a few variations, which all fall under the quotient space construction.

XIII.9.1. Collapsing a Subset to a Point

XIII.9.1.1. Let X be a topological space, and A a subset. We can "collapse A to a point" by taking the equivalence relation $\sim$ for which A is one equivalence class and all other equivalence classes are single points. Then the space $X/\sim$ with the quotient topology is usually denoted X/A and is called the *identification space* of X by A, or with A collapsed to a point. (This notation X/A is common, but sometimes conflicts with other common notation. For example, if X is a topological group and A a subgroup, then X/A usually denotes the quotient group (), which is not the same thing. See also XIII.9.3.6..) We may identify X/A with $(X \setminus A) \cup \{a\}$, where $a = \{A\}$.

In order to obtain a T_1 quotient, A must be closed in X, and the collapsing construction is almost exclusively used in this situation.

More generally, we may take a finite or infinite collection of pairwise disjoint subsets of X and collapse each one to a point. If A and B are disjoint subsets of X (usually closed), write $X/(A,B)$ for the space obtained from X by collapsing A and B to (separate) points. The collapsing can be done consecutively, i.e. by forming $(X/A)/B$ or $(X/B)/A$, or simultaneously. Note that $X/(A,B) \ncong X/(A \cup B)$.

XIII.9.1.2. Proposition. Let X be a topological space and A a closed subset of X, and $f : X \to X/A$ the quotient map. Then f is closed, i.e. the decomposition is usc, and the restriction of f to $X \setminus A$ is a homeomorphism onto the open set $(X/A) \setminus \{a\}$.

Proof: $X \setminus A$ is open in X, $W = (X/A) \setminus \{a\}$ is open in X/A, and f is a bijection of $X \setminus A$ to W. Identifying W with $X \setminus A$, a subset V of W is open in X/A if and only if the corresponding subset of $X \setminus A$ is open in X, i.e. V is relatively open in W if and only if $f^{-1}(V)$ is relatively open in $X \setminus A$. Thus $f|_{X \setminus A}$ is a homeomorphism.

If E is a subset of X, then $f^{-1}(f(E)) = E$ if $E \cap A = \emptyset$, and $f^{-1}(f(E)) = E \cup A$ if $E \cap A \neq \emptyset$. In any case, if E is closed, $f^{-1}(f(E))$ is closed, so f is closed.

All the results can fail if A is not closed (Exercise XIII.9.5.1.).

XIII.9.1.3. Conversely, if $f : X \to Y$ is a quotient map and every point of Y but one has a singleton preimage, then $Y = X/f^{-1}(\{q\})$, where q is the exceptional point. If $\{q\}$ is closed in Y (e.g. if Y is T_1), then f is closed and $f|_{X \setminus f^{-1}(\{q\})}$ is a homeomorphism onto $Y \setminus \{q\}$.

XIII.9.1.4. Examples. (i) Let f be the map from $[0,1]$ to $\mathbb{T} \cong S^1$ from XIII.8.1.3.(ii). We have $S^1 \cong [0,1]/\{0,1\}$, i.e. a circle is a closed interval with endpoints identified.

(ii) More generally, let B^n be the closed unit ball in $\mathbb{R}^n$. The boundary is the $(n-1)$-sphere S^{n-1}. We have that $B^n/S^{n-1} \cong S^n$, as is easily checked (e.g. via XIII.9.1.8.). For another construction of S^n by collapsing, see XIII.9.1.12.(iii).

XIII.9.1.5. Proposition. Let X be a Hausdorff topological space, and $A \subseteq X$. Then X/A is Hausdorff if and only if, for each $p \in X \setminus A$, A and p have disjoint neighborhoods. (This entails that A must be a closed subset of X.) In particular, if X is regular and A is closed, then X/A is Hausdorff.

XIII.9.1.6. An identification space from a regular space need not be regular. For example, if X is regular but not normal, and A and B are disjoint closed sets in X which do not have disjoint neighborhoods, then X/A is not regular. However, we have the following special case of XIII.8.3.9.:

XIII.9.1.7. Proposition. Let X be a normal space and A a closed subset of X. Then X/A is normal.

A particularly important special case is:

XIII.9.1.8. Proposition. Let X be a compact Hausdorff space, and A a closed subset of X. Then X/A is homeomorphic to the one-point compactification (XIII.11.8.3.) of $X \setminus A$.

Proof: Combine XIII.9.1.3. and XIII.11.8.5.. 👍

Cones and Suspensions

We can define a cone over any topological space X by forming a cylinder with base X and collapsing the top edge to a point:

XIII.9.1.9. Definition. Let X be a topological space. The space $CX = ([0,1] \times X)/(\{1\} \times X)$ is called the *cone* over X.

This construction is primarily used when X is a compact Hausdorff space. In this case, CX is homeomorphic to the one-point compactification of $[0,1) \times X$ (XIII.9.1.8.).

XIII.9.1.10. Examples. (i) The cone over $[0,1]$ is (homeomorphic to) a solid triangle.

(ii) The cone over S^1 is an ordinary cone, which is homeomorphic to a closed disk.

(iii) More generally, for any n the cone over S^n is homeomorphic to B^{n+1} (note that $B^{n+1} \setminus \{\mathbf{0}\}$ is homeomorphic to $[0,1) \times S^n$).

(iv) The cone CK over the Cantor set K can be identified with the union of all line segments in $\mathbb{R}^2$ between $\left(\frac{1}{2}, 1\right)$ and points of K, regarded as a subset of the x-axis. The point $\left(\frac{1}{2}, 1\right)$ can be replaced by any point of $\mathbb{R}^2$ not on the x-axis. This is a good example of a space which is path-connected but not locally path connected. (CX is path-connected for any topological space X.)

We can similarly form the "suspension" of X by collapsing each end of the cylinder over X to (separate) points:

XIII.9.1.11. DEFINITION. Let X be a topological space. The *suspension* of X is

$$SX = ([0,1] \times X)/(\{0\} \times X, \{1\} \times X) \ .$$

This is also primarily used in the case where X is a compact Hausdorff space, and gives a "two-point compactification" of $(0,1) \times X \cong \mathbb{R} \times X$ in this case. In general, the suspension of X is formed by collapsing the base of CX to a point.

The suspension of X can be thought of as two copies of CX joined base-to-base ().

XIII.9.1.12. EXAMPLES. (i) The suspension of $[0,1]$ can be pictured as a (solid) diamond in $\mathbb{R}^2$.

(ii) The suspension of S^1 can be pictured as the union of two cones joined along their base circles (). This space is homeomorphic to S^2.

(iii) More generally, the suspension of S^n is homeomorphic to S^{n+1}. This can be seen by first forming $CS^n \cong B^{n+1}$, and then collapsing the base, which is the boundary S^n, to a point (cf. XIII.9.1.4.(ii)).

Wedge Sums

XIII.9.1.13. Let $\{X_i : i \in I\}$ be a collection of topological spaces, and for each i let $*_i$ be a designated point of X_i. The *wedge sum* of $\{(X_i, *_i) : i \in I\}$ is the space X obtained from the separated union of the X_i with all the $*_i$ identified together to one point $*$. (The homeomorphism class of X depends in general on the choice of the $*_i$.) If each X_i is a circle, the space X is called a *wedge of circles* or *bouquet of circles*.

If I is a finite set with n elements and each X_i is a circle, the space X is the bouquet of n circles, resembling a flower with n petals. A bouquet of two circles is homeomorphic to a figure 8.

A wedge sum of infinitely many nontrivial spaces is rarely metrizable. For example, a wedge or bouquet of infinitely many circles is not first countable (XIII.9.5.7., XIII.9.5.8.).

XIII.9.2. Gluing and Attaching

Gluing and attaching are perhaps the most extensively used constructions in topology.

XIII.9.2.1. Let X be a topological space, and A and B disjoint subsets of X (with care, the disjointness can be relaxed). Let $f : A \to B$ be a homeomorphism, or more generally a quotient map. We take the quotient of X obtained by identifying x with $f(x)$ for all $x \in A$. More precisely, we define an equivalence relation $\sim$ on X by taking

$$[x] = \{y \in A : f(y) = f(x)\} \cup \{f(x)\}$$

if $x \in A$, $[x] = \{x\} \cup f^{-1}(\{x\})$ if $x \in B$, and $[x] = \{x\}$ if $x \notin A \cup B$. The space $X/\sim$ is called the *identification space* of X with A identified with B via f, or with A *glued* (or *sewed*) to B via f. Note that if f is a homeomorphism, the equivalence classes just consist of $\{x, f(x)\}$ for $x \in A$ and $\{x\}$ for $x \notin A \cup B$.

We may more generally do more than one gluing, gluing A_1 to B_1 and A_2 to B_2, etc. We need that A_j and B_j are disjoint for each j, but A_j and A_k, B_j and B_k, and A_j and B_k need not be disjoint for $j \neq k$, but the gluings have to be consistent on the overlaps. The gluing can be done consecutively or simultaneously. See e.g. XIII.9.2.3.(v).

XIII.9.2.2. In the gluing construction, to obtain a nice quotient space it is not necessary that A and B be closed (XIII.9.2.3.(ii)); however, the case where they are not closed is rather delicate. The construction is more robust if they are closed.

XIII.9.2.3. EXAMPLES. (i) Let $X = [0,1]$, $A = \{0\}$, $B = \{1\}$, f the obvious homeomorphism. Then the identification space with A glued to B is S^1, as in XIII.8.1.3.(ii) or XIII.9.1.4..

(ii) Let $X = (0,1)$, $A = \left(0, \frac{1}{3}\right)$, $B = \left(\frac{2}{3}, 1\right)$, $f(x) = x + \frac{2}{3}$ for $x \in A$. Then the identification space is homeomorphic to S^1.

The identification space depends crucially on the choice of homeomorphism. If instead we take $f(x) = 1 - x$ for $x \in A$, the identification space is not Hausdorff.

(iii) Let $X = [0,1]^2$, $A = [0,1] \times \{0\}$, $B = [0,1] \times \{1\}$. Define $f : A \to B$ by $f(t,0) = (t,1)$. Then the identification space is a cylinder (Figure XIII.2). Informally, we take a rectangle and glue the ends together with no twist.

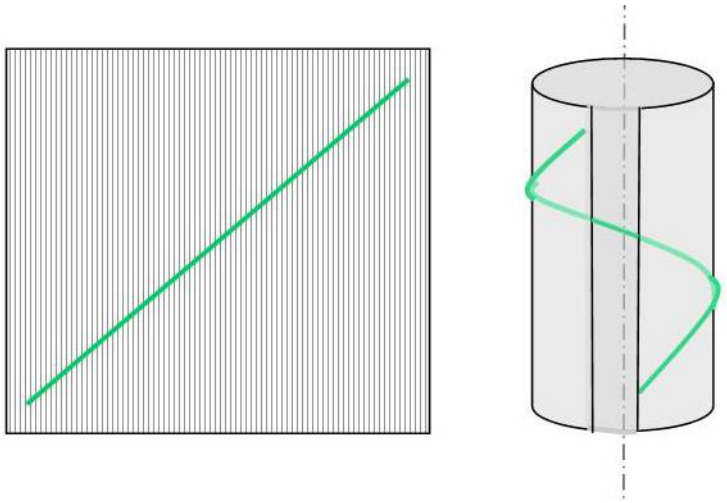


Figure XIII.2: A cylinder.

(iv) Let $X = [0,1]^2$, $A = [0,1] \times \{0\}$, $B = [0,1] \times \{1\}$. Define $f : A \to B$ by $f(t,0) = (1-t,1)$. Then the identification space is a Möbius strip (Figure XIII.3, from https://upload.wikimedia.org/wikipedia/commons/d/d9/M bius_strip.jpg). Informally, we take a rectangle and glue the ends together with a twist.

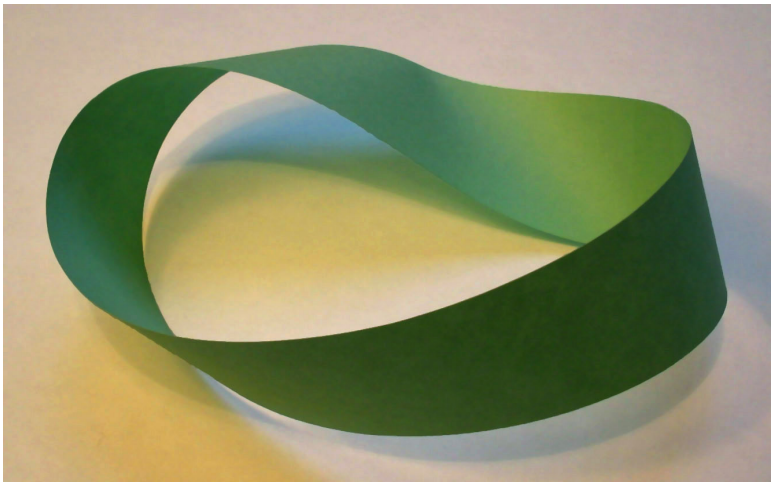


Figure XIII.3: A Möbius strip.

(v) Let $X = [0,1]^2$, $A_1 = [0,1] \times \{0\}$, $B_1 = [0,1] \times \{1\}$, $A_2 = \{0\} \times [0,1]$, $B_2 = \{1\} \times [0,1]$. Define $f_1 : A_1 \to B_1$ by $f(t,0) = (t,1)$ and $f_2 : A_2 \to B_2$ by $f_2(0,t) = (1,t)$. The identification space is homeomorphic to a 2-torus $T^2 \cong S^1 \times S^1$. It can be made from a rectangle by first gluing the top and bottom edges to form a cylinder, and then gluing the end circles (with no twist); see Figure XIII.4, from http://bakingandmath.com/tag/torus/. (See this website, or https://en.wikipedia.org/wiki/Homotopy, for an animation of the topological equivalence of a torus and a coffee cup; the old joke is that a topologist is a mathematician who cannot tell the difference between a doughnut and a coffee cup.)

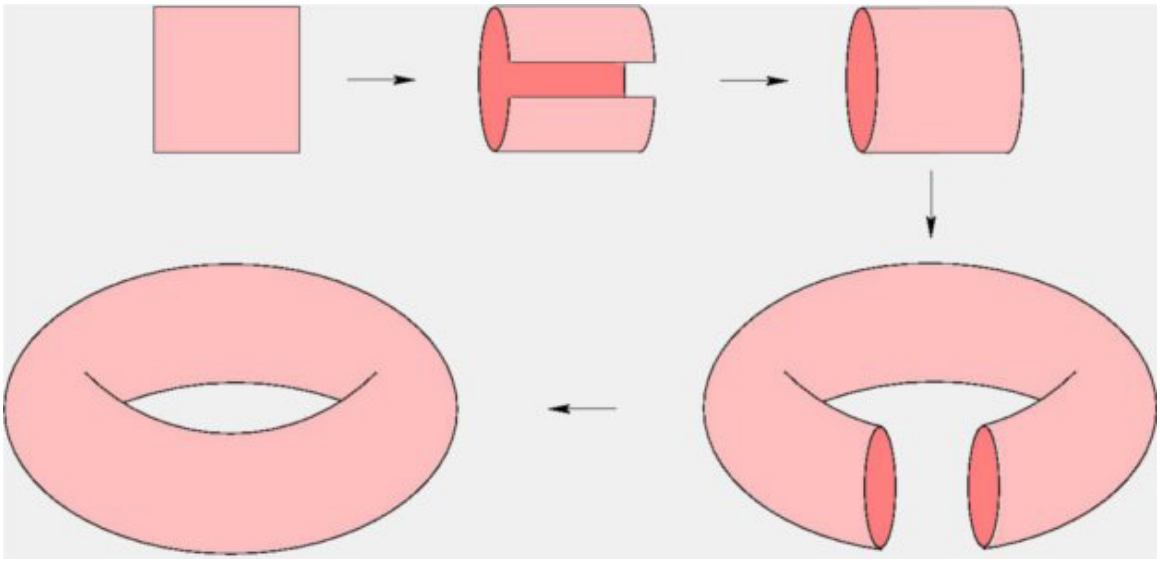


Figure XIII.4: Construction of a torus.

(vi) Let $X = [0,1]^2$, $A_1 = [0,1]\times\{0\}$, $B_1 = [0,1]\times\{1\}$, $A_2 = \{0\}\times[0,1]$, $B_2 = \{1\}\times[0,1]$. Define $f_1 : A_1 \to B_1$ by $f(t,0) = (t,1)$ and $f_2 : A_2 \to B_2$ by $f_2(0,t) = (1,1-t)$. The identification space is called a *Klein bottle* (actually in German it is *Kleinsche Fläche*, "Klein surface," not *Kleinsche Flasche*; the English name may be a mistranslation, although the standard picture indicates that the "bottle" name is not inappropriate). It can be made from a rectangle by first gluing the top and bottom edges to form a cylinder, and then gluing the end circles with an orientation-reversing map; see Figure XIII.5 (from https://en.wikipedia.org/wiki/Klein_bottle) or, for an animated version, https://www.youtube.com/watch?v=E8rifKlq5hc; note that the apparent self-intersection is only an artifact of trying to depict the Klein bottle as a subset of $\mathbb{R}^3$ – the picture is really a higher-dimensional version of a knot diagram in $\mathbb{R}^2$. A Klein bottle cannot be topologically embedded in $\mathbb{R}^3$, but it can be embedded in $\mathbb{R}^4$. A Klein bottle can also be constructed from a Möbius strip by suitably identifying points on the boundary circle (Figure XIII.6, from http://squareone-learning.com/blog/2014/09/around-and-round-and-never-off-the-surface/), or by gluing two Möbius strips together along their boundary circle. There are many nice pictures and videos of the Klein bottle on the internet, e.g. https://www.youtube.com/watch?v=sRTKSzAOBr4.

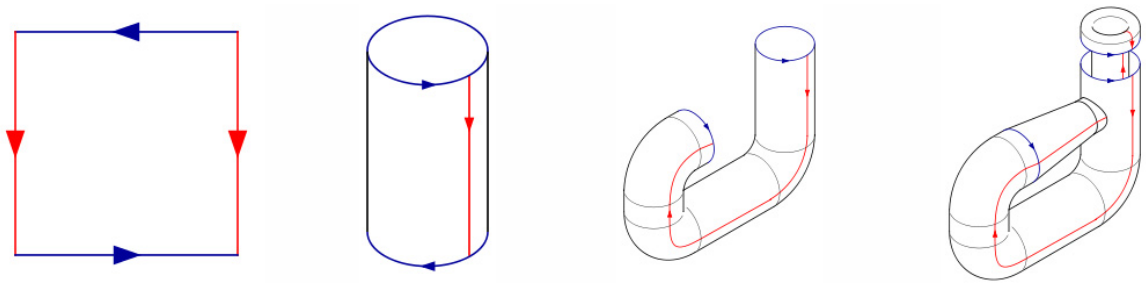


Figure XIII.5: Construction of a Klein bottle.

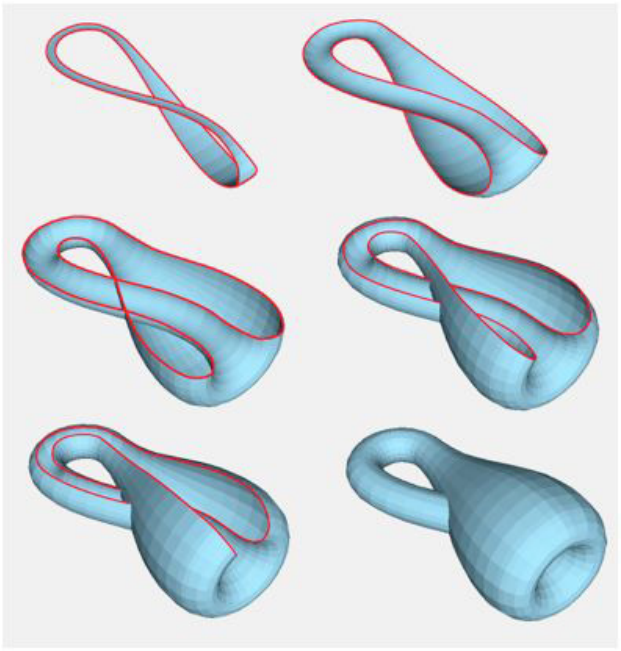


Figure XIII.6: Construction of a Klein bottle from a Möbius strip.

(vii) Let $X = [0,1]^2$, $A_1 = [0,1] \times \{0\}$, $B_1 = [0,1] \times \{1\}$, $A_2 = \{0\} \times [0,1]$, $B_2 = \{1\} \times [0,1]$. Define $f_1 : A_1 \to B_1$ by $f(t,0) = (1-t,1)$ and $f_2 : A_2 \to B_2$ by $f_2(0,t) = (1,1-t)$. The identification space is called the *projective plane*. It can be made from a rectangle by first gluing the top and bottom edges with a twist to form a Möbius strip, and then suitably identifying points on the boundary circle (not in the same way as for a Klein bottle). A better way is to take A as the union of the top and left edges, and B the union of the right and bottom edges (these sets are not quite disjoint), and then gluing A to B. This amounts to gluing the top semicircle of the closed unit disk in $\mathbb{R}^2$ to the bottom semicircle by a 180° rotation, or identifying $\boldsymbol{x}$ with $-\boldsymbol{x}$ for each $\boldsymbol{x}$ of norm 1. See Figure XIII.7, from http://www.math.cornell.edu/~mec/Winter2009/Victor/part3.htm.

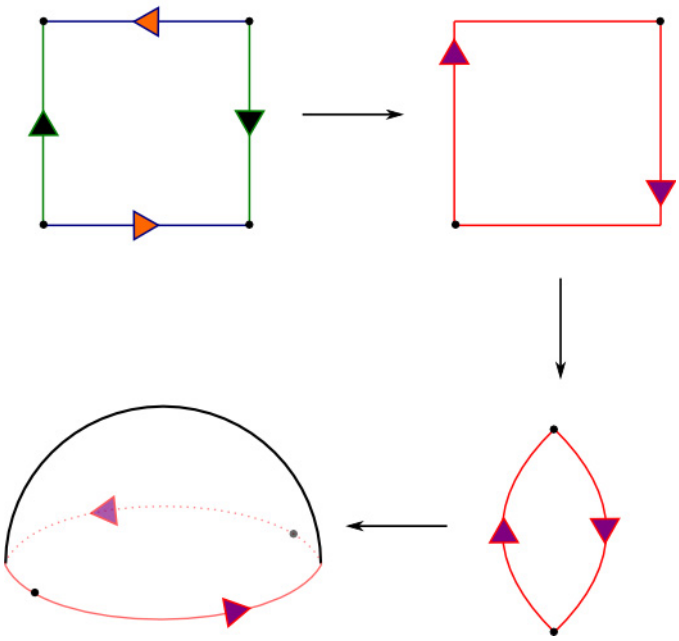


Figure XIII.7: Construction of the projective plane.

A projective plane is harder to picture than a Klein bottle, but there are many ways to do it; see Figure XIII.8, from http://xahlee.info/math/riemann_sphere__real_projective_plane.html for one way. Unlike this picture which has a pinch point, there is a (nonobvious – HILBERT thought it was impossible!) way to immerse the projective plane in $\mathbb{R}^3$ called *Boy's surface* [Boy03]; see https://www.youtube.com/watch?v=9gRx66xKXek for a nice animation. Note again that the apparent self-intersection is only an artifact of trying to depict the projective plane as a subset of $\mathbb{R}^3$. A projective plane cannot be topologically embedded in $\mathbb{R}^3$, but it can be embedded in $\mathbb{R}^4$. There are several other standard ways to construct the projective plane (XIII.9.3.7.); another nice way to visualize it is as *Steiner's Roman surface* (see https://en.wikipedia.org/wiki/Roman_surface#/media/File:Steiner's_Roman_Surface.gif for an animation). The projective plane is denoted $\mathbb{R}P^2$; there are generalizations (real projective spaces) $\mathbb{R}P^n$ for each n (XIII.9.3.7.(iii)), as well as complex projective spaces $\mathbb{C}P^n$ (XIII.9.3.7.(iv)).

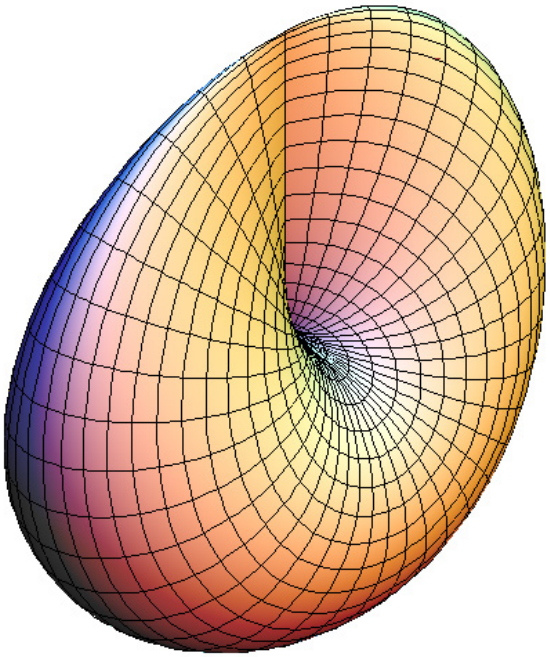


Figure XIII.8: Projective plane as disk with cross-cap.

(viii) Every closed surface (closed 2-manifold) (including S^2!) can be obtained by starting with a regular $(2n)$-gon for some n and identifying pairs of edges in an appropriate way (XIII.9.5.4.).

(ix) A more complicated identification construction important in the theory of dynamical systems is the "flow under a constant function." Let X be a topological space (typically a compact metrizable space), and $\phi : X \to X$ a homeomorphism (typically a *minimal homeomorphism*, i.e. all orbits are dense). From $[0,1] \times X$, form the identification space by identifying $\{0\} \times X$ with $\{1\} \times X$ by identifying $(0,x)$ with $(1,\phi(x))$ for $x \in X$. The identification space is often called the *mapping torus of* ϕ. If ϕ is the identity map, one just obtains $S^1 \times X$, but if ϕ is minimal a (typically complicated) connected space is obtained. The construction in XIII.19.5.18. is a prime example. Another simple example is to take $X = S^1$ and ϕ the antipodal map (this homeomorphism is not minimal!); the construction then yields the Klein bottle. Even a minimal homeomorphism does not necessarily give something exotic: if $X = S^1$ and ϕ is an irrational rotation, the construction just gives a torus. However, in general more than just the identification space is obtained from the construction: there is an obvious canonical action of $\mathbb{R}$ on the identification space ("flow"), and the orbits of this action (if X is infinite and the homeomorphism is minimal) are densely embedded continuous images of $\mathbb{R}$. In the last case of an irrational rotation, one gets copies of $\mathbb{R}$ wrapped around the torus at irrational angle. See XIII.9.5.10.(c) for more about mapping tori.

Attaching

XIII.9.2.4. A variation of gluing is to have two spaces X and Y, subsets A of X and B of Y, and a homeomorphism (or quotient map) $f : A \to B$. This can be put into the previous setting by considering the disjoint union of X and Y. The glued identification space is usually denoted $X \cup_f Y$; we say X is *attached* to Y via f.

XIII.9.2.5. There are two especially important examples of attaching: CW complexes () and surgery. See XIII.26.2.25. for some interesting examples where f is not injective; actually collapsing a subset to a point is an example (where $B = Y$ is a single point).

Handles and Surgery

Surgery is a method of creating new manifolds from known ones. The general setup of surgery in topology is based on the following important fact:

XIII.9.2.6. PROPOSITION. Let M and N be n-manifolds with boundary, and let $f : \partial M \to \partial N$ be a homeomorphism. Then $M \cup_f N$ is an n-manifold without boundary.

XIII.9.2.7. The general idea of surgery is to remove from a (usually closed) n-manifold M an open submanifold and glue in another n-manifold N whose boundary is homeomorphic to the boundary of the removed piece to obtain a new closed n-manifold. The simplest interesting example is:

XIII.9.2.8. Let M be a closed surface, e.g. S^2. Cut two disjoint open disks out of a larger disk in M. What is left is a compact 2-manifold with boundary M_0 whose boundary consists of a disjoint union of two circles, which may be given a consistent orientation. The cylinder $N = [0, 1] \times S^1$ is also a closed manifold with boundary, whose boundary consists of two circles. Glue N to M_0 along the boundary circles, giving them the same orientation. The result is called M *with a handle attached*. See Figure XIII.9, from http://www.uff.br/cdme/pdp/pdp-html/pdp-en.html. It is easy to see that S^2 with a handle attached is homeomorphic to T^2, and T^2 with a handle attached is a two-holed torus. An n-holed torus is obtained by adding n handles to S^1 (Figure XIII.10, from https://upload.wikimedia.org/wikipedia/commons/a/ac/Sphere_with_three_handles.png). It turns out that every closed surface is homeomorphic to either S^2 (if orientable) or $\mathbb{R}P^2$ or a Klein bottle (if nonorientable) with some finite number of handles attached. In the orientable case, the number of handles is called the *genus* and is a topological invariant (). Genus can also be defined in the nonorientable case (XIII.9.2.12.).

If N is attached to S^2 with the end circles given opposite orientations, the surface obtained is the Klein bottle, not the torus.

XIII.9.2.9. A variation is to take closed surfaces M and N, cut out small disks from both of them, and glue the remainders together along the boundary circles (see Figure XIII.11, from https://upload.wikimedia.org/wikipedia/commons/thumb/5/52/Connected_sum.svg/1280px-Connected_sum.svg). The resulting surface is called the *connected sum* of M and N, denoted $M\#N$. The connected sum of an m-holed torus and an n-holed torus is an $(m + n)$-holed torus. Connected sum with a torus is topologically the same thing as adding a handle.

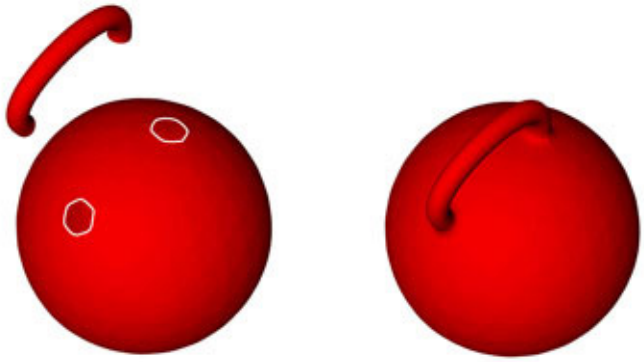


Figure XIII.9: Attaching a handle to a sphere.

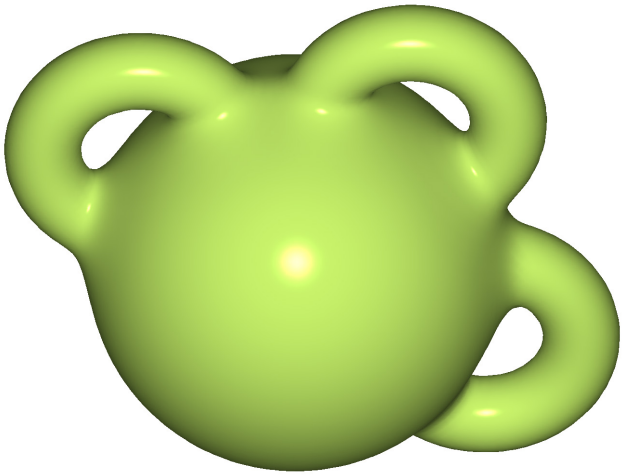


Figure XIII.10: Sphere with three handles.

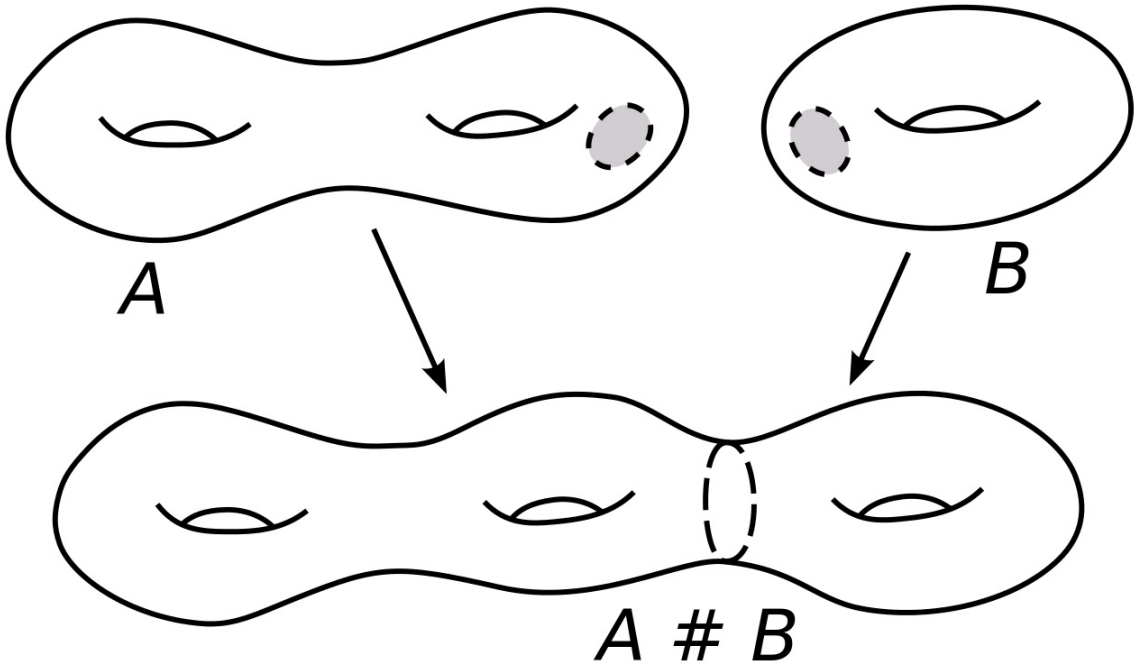


Figure XIII.11: Connected sum.

Connected sum can be done for n-manifolds by cutting out a small ball B^n from both and gluing the remainders along the boundary S^{n-1}'s. In general, the orientation chosen for the identification of the boundaries affects the homeomorphism class of the connected sum ().

XIII.9.2.10. In higher dimensions, there is more flexibility in surgery. A key observation is that for any p and q, the boundary of $S^p \times B^q$ is homeomorphic to $S^p \times S^{q-1}$, as is the boundary of $B^{p+1} \times S^{q-1}$. (We used this for handles with $p = 0$ and $q = 2$). Thus, if M is an n-manifold, we can take any p and q with $p + q = n$, cut out an open submanifold whose closure is homeomorphic to $S^p \times B^q$, and glue in a copy of $B^{p+1} \times S^{q-1}$.

If $n = 3$, the only possibility of this type is to add a "three-dimensional handle," cutting out two 3-balls (a copy of $S^0 \times B^3$) and replacing it by a copy of $B^1 \times S^2 \cong [0, 1] \times S^2$. But if $n = 4$, there is another interesting possibility: cutting out a copy of $S^1 \times B^3$ and replacing it with a copy of $B^2 \times S^2$. See ().

XIII.9.2.11. Actually, even in three dimensions, there is another interesting possibility: the removed manifold with boundary can be glued back in, but with a different homeomorphism on the boundary. (It may be hard to believe this could make a different manifold, but such a self-homeomorphism of the boundary does not generally extend to a self-homeomorphism of the removed manifold.) This can be done also in two dimensions by removing a handle and gluing it back on reversing the orientation of one of the boundary circles, but the three-dimensional case is more interesting since if the removed piece is a solid torus, there are many essentially different ways of gluing it back in (), none of which affect orientability. The removed solid torus may be knotted or unknotted, and this type of surgery is called *surgery on a knot*. More generally, we can do surgery on a link. It is known (Lickorish-Wallace Theorem) that every closed orientable 3-manifold can be obtained from S^3 by surgery on a link ([Wal60], [Lic62]).

XIII.9.2.12. Other types of surgery are possible. For example, the boundary of a Möbius strip is a circle, so if M is a 2-manifold, we may cut a disk out of M and glue in a Möbius strip (called a *cross-cap*). A 2-sphere with a cross-cap added is a projective plane (Figure XIII.8). Adding a cross-cap is the same thing as taking connected sum with a projective plane. Every nonorientable closed surface is obtained by adding a number of cross-caps to a sphere; this number is the genus (). Similar things can be done in higher dimensions.

XIII.9.3. Group Actions on Spaces

XIII.9.3.1. Another important source of quotients arises from group actions. A *group of homeomorphisms* of a topological space X is a set G of homeomorphisms of X such that

(i) G is closed under composition: if $g, h \in G$, then $g \circ h$ (usually written gh) is also in G.

(ii) G is closed under inversion: if $g \in G$, then $g^{-1} \in G$.

(recall that compositions and inverses of homeomorphisms are homeomorphisms). It follows that the identity map on X is in G.

It turns out to be convenient to change point of view slightly:

XIII.9.3.2. DEFINITION. Let G be a group and X a topological space. An *action* of G on X is a homomorphism $\sigma : G \to \text{Homeo}(X)$ from G to the homeomorphism group of X.

We usually write σ_g in place of $\sigma(g)$, and often just write $g \cdot x$ for $\sigma_g(x)$. The homomorphism property of σ is that $\sigma_{gh} = \sigma_g \circ \sigma_h$ (written $\sigma_g\sigma_h$), or $(gh) \cdot x = g \cdot (h \cdot x)$, for $g, h \in G$, $x \in X$; we have $\sigma_{g^{-1}} = \sigma_g^{-1}$ for $g \in G$, and σ_e is the identity homeomorphism if e is the identity of g.

XIII.9.3.3. The group action picture includes the group of homeomorphisms picture: if G is a group of homeomorphisms of X, the inclusion map of G into $\text{Homeo}(X)$ is an action of G on X. Conversely, if σ is an injective homeomorphism from G to $\text{Homeo}(X)$ (we say σ is a *faithful* action of G on X), then G can be identified with the group $\sigma(G)$ of homeomorphisms of X. The group action picture is slightly more general since group actions need not be faithful (in the theory of group representations () it turns out to be important to allow nonfaithful representations). The conceptual importance of the group action picture is that the group G can be considered abstractly and is not *a priori* tied to homeomorphisms of a particular space, and it is useful to let the same abstract group act on many different spaces.

Group actions of groups on sets can be similarly defined with no relation to topology: just consider homomorphisms from a group to the group of bijections (permutations) of the set. This can be regarded as a special case of group actions of groups on topological spaces where the space has the discrete topology.

Actions of topological groups on topological spaces are also considered (). In this section, we will not consider the case of topological groups (or, alternatively, we will give all groups the discrete topology).

Orbits

XIII.9.3.4. DEFINITION. Let σ be an action of a group G on a topological space X, and $x \in X$. The *orbit* of x under G (or σ) is the set

$$[x] = \{g \cdot x : g \in G\} = \{\sigma_g(x) : g \in G\} \ .$$

Of course, $x \in [x]$ since $x = e \cdot x$. The crucial fact about orbits is that they form a partition of the space:

XIII.9.3.5. PROPOSITION. Let σ be an action of a group G on a topological space X. If $x, y \in X$, then either $[x] = [y]$ or $[x]$ and $[y]$ are disjoint. Equivalently, the relation $x \sim y$ if $[x] = [y]$ is an equivalence relation on X.

PROOF: Suppose $z \in [x] \cap [y]$. Then there are $g, h \in G$ with $z = g \cdot x$ and $z = h \cdot y$. Then $y = h^{-1} \cdot z = (h^{-1}g) \cdot x$, and similarly $x = (g^{-1}h) \cdot y$, so $[x] = [y]$.

XIII.9.3.6. Thus, if σ is an action of a group G on a topological space X, we may form the quotient space X/σ of orbits, called the *orbit space* of the action, and give it the quotient topology. The orbit space is also often written X/G, especially if G is given concretely as a group of homeomorphisms of X, although this notation conflicts with the notation for identification spaces (XIII.9.1.1.). 

Many interesting topological spaces can be constructed as orbit spaces of group actions:

XIII.9.3.7. EXAMPLES. (i) Let X be the unit circle in $\mathbb{R}^2$. The set consisting of the identity and reflection across the x-axis is a group of homeomorphisms of X. Each orbit consists of two points except for $[(1,0)]$ and $[(-1,0)]$, which are singletons. The orbit space is homeomorphic to the closed interval $[-1,1]$, and the quotient map can be identified with projection onto the x-axis.

(ii) Let $\mathbb{Z}$ act on $\mathbb{R}$ by translation: $n \cdot x = x + n$. Each orbit contains a unique point in the interval $(0,1]$. But the quotient topology is not the subspace topology on $(0,1]$: for example, we have $\left[\frac{1}{n}\right] \to [0] = [1]$. In fact, the quotient space $\mathbb{R}/\mathbb{Z}$ is homeomorphic to the circle obtained from $[0,1]$ by identifying 0 and 1.

(iii) Let X be the unit sphere S^{n-1} in $\mathbb{R}^n$. The set consisting of the identity and the antipodal map $\boldsymbol{x} \mapsto -\boldsymbol{x}$ is a group of homeomorphisms of S^{n-1}. Every orbit has two points. The quotient space consists of S^{n-1} with antipodal points identified, and is called *real projective* $(n-1)$*-space*, denoted $\mathbb{R}P^{n-1}$. If $n = 3$, we obtain the (real) projective plane (XIII.9.2.3.). If $n = 2$, we obtain the quotient of the circle by the antipodal map, which is a (smaller) circle. Note that this action of $\mathbb{Z}_2$ on the circle is very different from the action in (i).

(iv) We may also define complex projective space. Let $\mathbb{T}$ be the multiplicative group of complex numbers of absolute value 1. Regard S^{2n-1} as the unit sphere in $\mathbb{C}^n$. Then $\mathbb{T}$ acts on S^{2n-1} by $\lambda \cdot \boldsymbol{x} = \lambda \boldsymbol{x}$. The quotient $S^{2n-1}/\mathbb{T}$ is called *complex projective* n*-space*, denoted $\mathbb{C}P^{n-1}$. It is a $(2n-2)$-dimensional manifold.

Projective spaces can also be described in other ways. $\mathbb{R}P^n$ can be described as the set of lines through the origin in $\mathbb{R}^{n+1}$, the quotient of $\mathbb{R}^{n+1} \setminus \{\mathbf{0}\}$ by identifying two points if they lie on the same line through $\mathbf{0}$. Each such line contains exactly two points on the unit sphere, which are identified; this leads to the picture of $\mathbb{R}P^n$ as the closed upper hemisphere of S^n with antipodal points on the "equator" identified, the higher-dimension analog of the picture in XIII.9.2.3.(vii). Similarly, $\mathbb{C}P^n$ is the set of complex "lines" through the origin (one-dimensional complex subspaces) in $\mathbb{C}^{n+1}$.

$\mathbb{R}P^n$ can also be described as $\mathbb{R}^n$ with certain points at infinity added: identify $\boldsymbol{x} \in \mathbb{R}^n$ with the line through the origin and the point $(\boldsymbol{x}, 1) \in \mathbb{R}^{n+1}$ (or, equivalently, the equivalent point of norm 1 on the upper hemisphere of S^n). The points at infinity added to make $\mathbb{R}P^n$ can be regarded as "lines at infinity," which form a copy of $\mathbb{R}P^{n-1}$. There is a similar description for $\mathbb{C}P^n$ as $\mathbb{C}^n$ with "complex lines at infinity" added (a copy of $\mathbb{C}P^{n-1}$). These pictures are the origin of the name "projective space."

(v) The 3-sphere S^3 can be regarded as the unit sphere in $\mathbb{C}^2$:

$$S^3 = \{(z,w) \in \mathbb{C}^2 : |z|^2 + |w|^2 = 1\} .$$

Fix $p > 1$ and let $\lambda = e^{2\pi i/p}$. Then λ is a complex number of absolute value 1, and $\lambda^p = 1$. Define a faithful action of $\mathbb{Z}_p$ (or a nonfaithful action of $\mathbb{Z}$) on S^3 by

$$k \cdot (z,w) = (\lambda^k z, \lambda^k w) .$$

The quotient space $S^3/\mathbb{Z}_p$ is called a *lens space*. If $p = 2$, we obtain $\mathbb{R}P^3$. More generally, if q is an integer relatively prime to p (we may restrict to $0 < q < p$), we can define the action by

$$k \cdot (z,w) = (\lambda^k z, \lambda^{qk} w) .$$

The lens spaces obtained, denoted $L(p,q)$, $L(p;q)$, or $L_{q/p}$, depend rather subtly on p and q: they can be homotopy equivalent without being homeomorphic, and, although orientable, can fail to have orientation-reversing homeomorphisms (). Lens spaces can also be constructed by gluing two solid tori along their boundary (); see [Mun84] for another construction of lens spaces by identifying points on the boundary of B^3, or [Thu97] for more constructions and an explanation of why these are called "lens spaces."

Homogeneous Spaces

XIII.9.3.8. A simple but particularly important example of a group action is when the space X is a topological group and G is a subgroup of X. A topological group acts on itself by (right) translation (), and so does any subgroup, i.e. $g \cdot x = xg^{-1}$ for $g \in G$, $x \in X$ (this formula with the inverse is needed to make $(gh) \cdot x = g \cdot (h \cdot x)$ for $g, h \in G$, $x \in X$). The orbit space X/G, which is the space of left cosets () of G in X, with the quotient topology is called a *homogeneous space*. (There is additional structure here: the group X acts on X/G by left translation. If G is a normal subgroup of X, then X/G is also a topological group.) The quotient space has reasonable properties, e.g. is Hausdorff, only when G is a closed subgroup of X (cf. Exercise XIII.9.5.1.). The simplest case is when G is a closed discrete subgroup.

Example XIII.9.3.7.(ii) is a basic example of this construction. A closely related example is to take $X = \mathbb{R}^n$ and $G = \mathbb{Z}^n$ the additive subgroup of integer points; the quotient is the n-torus $\mathbb{T}^n$, which is a product of n circles. In the case $n = 2$ we obtain the ordinary torus (XIII.9.2.3.(v)); higher-dimensional tori can be similarly constructed by identifying opposite faces of an n-cube.

XIII.9.3.9. The lens spaces $L(p, 1)$ (XIII.9.3.7.(v)) are another example. The topological group $SU(2)$ (also isomorphic to $Sp(1)$ and to the multiplicative group of unit quaternions) is homeomorphic to S^3, and we can take the quotient by the cyclic subgroup generated by $diag(\lambda, \bar{\lambda})$. Other lens spaces do not arise as homogeneous spaces of $SU(2)$.

$SU(2)$ has other interesting finite subgroups, perhaps the most interesting of which is the *binary icosahedral group*, which covers the group of rigid motions of a regular icosahedron centered at the origin (a subgroup of $SO(3)$, which is a quotient of $SU(2)$); the quotient of $SU(2) \cong S^3$ by this group is called the *Poincaré homology sphere*, a closed 3-manifold with the same homology as S^3 but not homeomorphic to S^3 (it is not simply connected; its fundamental group is the binary icosahedral group, a group of order 120 with no nontrivial abelian quotients). The Poincaré homology sphere can also be constructed from a regular dodecahedron by identifying opposite faces with an appropriate twist (). See [Thu97] and [Wee02] for a nice description and for many other examples of 3-manifolds obtained from polyhedra by gluing faces (cf. XIII.9.5.6.).

Stabilizers and Freeness

XIII.9.4. Local Homeomorphisms and Covering Maps

XIII.9.5. Exercises

XIII.9.5.1. (a) Let $X = \mathbb{R}$ and $A = \mathbb{Q}$, and consider the identification space X/A. Show that all the conclusions of XIII.9.1.2. fail, and X/A is not T_1. Is X/A T_0?

(b) Let $\mathbb{R}/\mathbb{Q}$ be the homogeneous space (XIII.9.3.8.) considering $\mathbb{Q}$ as an additive subgroup of $\mathbb{R}$. Show that $\mathbb{R}/\mathbb{Q}$ is the space of XIII.8.1.11., and thus has the indiscrete topology.

Note that, although the notation $\mathbb{R}/\mathbb{Q}$ is used in both (a) and (b), the spaces are quite different.

XIII.9.5.2. Figures XIII.12 and XIII.13 show orientable closed surfaces. By the classification, each is homeomorphic to S^2 with some number of handles. Find the number of handles (genus), and describe a homeomorphism. (There are not only homeomorphisms, but even isotopies; cf. https://www.youtube.com/watch?v=k8Rxep2Mkp8.)

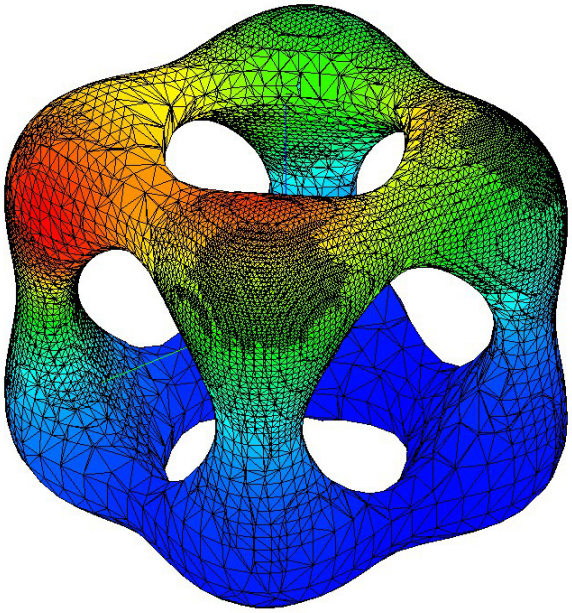


Figure XIII.12: An orientable closed surface.

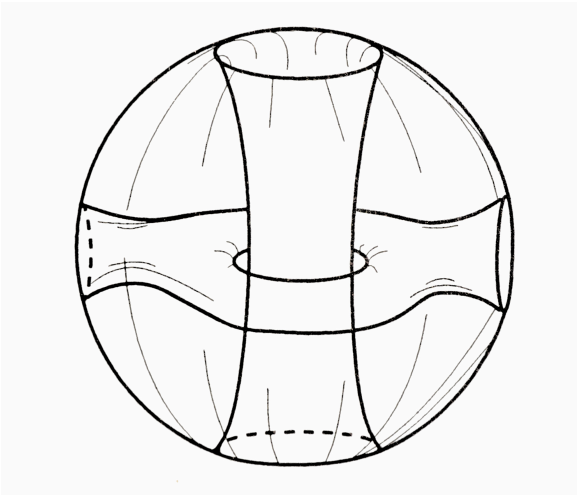


Figure XIII.13: Another orientable closed surface.

XIII.9.5.3. (a) Show that $\mathbb{R}P^2\#\mathbb{R}P^2$ is homeomorphic to the Klein bottle K.

(b) Show that $\mathbb{R}P^2\#K$ is homeomorphic to $\mathbb{R}P^2\#T^2$, and hence $\mathbb{R}P^2\#T^2$ is homeomorphic to $\mathbb{R}P^2\#\mathbb{R}P^2\#\mathbb{R}P^2$.

(c) Show that $\mathbb{R}P^2$ with n handles attached is homeomorphic to the connected sum of $2n+1$ copies of $\mathbb{R}P^2$, and that K with n handles attached is homeomorphic to the connected sum of $2n+2$ copies of $\mathbb{R}P^2$.

XIII.9.5.4. Let P be a polygon with $2n$ sides, which we may take to be a regular polygon. A *gluing scheme* on P consists of a choice of an orientation and a labeling of each edge of P, such that each label names exactly two edges. A gluing scheme defines a gluing of P by gluing each pair of edges with the same label according to the designated orientation. For example, the gluing scheme of Figure XIII.5 gives a Klein bottle when glued.

(a) Show that the gluing scheme of Figure XIII.14 gives a two-holed torus.

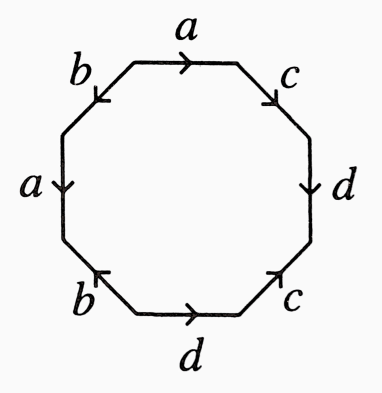


Figure XIII.14: Gluing scheme for a two-holed torus.

(b) Show that the gluing schemes of Figure XIII.15 both give S^2 when glued.

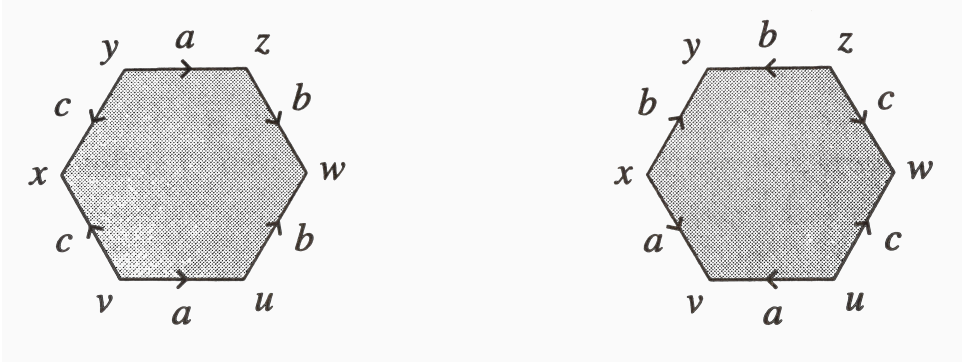


Figure XIII.15: Gluing schemes for a sphere.

(c) Show that in any gluing scheme, two consecutive edges with the same label and opposite orientations can be eliminated. [Fold them into the interior of the polygon.]

(d) Show that if a gluing scheme contains a labeled pair of edges with the same orientation (with respect to the boundary circle), then the glued space contains a Möbius strip. [See Figure XIII.16.] Thus the glued space is not orientable.

(e) Show that if all pairs of glued edges have opposite orientations, a sphere with some number of handles (perhaps none) is obtained from the gluing.

(f) In the situation of (d), show that the glued space consists of adding a cross-cap to a glued space from a polygon with fewer sides. Conclude by induction that the glued space is a connected sum of projective planes, or, equivalently, a sphere with a number of cross-caps added (cf. Exercise XIII.9.5.3.).

(g) Conclude from (e) and (f) that the glued space is always a closed surface which is either a sphere with handles attached if orientable, or a sphere with cross-caps attached (a connected sum of projective planes) if nonorientable. Show that every such surface can be obtained this way (although the polygon and gluing scheme are not unique).

(h) Show that if a connected closed surface can be triangulated (), it can be represented as the gluing of a polygon by a gluing scheme. Since every 2-manifold can be triangulated, every connected closed surface (2-manifold) is of the form of (g).

(i) Show that no two of the surfaces of (g) with different numbers of handles or projective planes are homeomorphic (). Thus (g) gives a complete classification of closed 2-manifolds.

See [Mas77] (or [Mas91]), or [Blo97] (from which the above pictures are taken), for a nice almost complete discussion of the classification of closed 2-manifolds. See [Tho92] for a different approach, and [FW99] or [Wee02] for a modern proof due to J. Conway. (Note that these proofs are all incomplete since they use the nontrivial fact that every surface can be triangulated, first proved by T. Radó in 1925 [**?**], cf. [AS60]. For a short proof of triangulability see [DM68] or [Tho92], and `http://arxiv.org/pdf/1312.3518.pdf` for a different argument not using the Jordan-Schönflies Theorem. See [Ber03, p. 139] for a simple approach to triangulating a compact surface using Riemannian geometry.) For a complete classification of connected noncompact (triangulable) surfaces, see [Ric63].

XIII.9.5.5. (a) Let $X = [0,1]$, $n \in \mathbb{N}$, $A_n = \left\{\frac{k}{n} : 0 \le k \le n\right\}$. Show that X/A_n is a bouquet of n circles. In particular, X/A_2 is a figure 8. Show that X/A_n is the one-point compactification of the disjoint union of n open intervals (copies of $\mathbb{R}$).

(b) Let $X = [0,1]$, $A_\omega = \{0\} \cup \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$. Show that X/A_ω is homeomorphic to the Hawaiian earring (XIII.19.5.8.). Show that X/A_ω is the one-point compactification of the disjoint union of $\aleph_0$ open intervals (copies of $\mathbb{R}$).

(c) Let K be the Cantor set. Show that the quotient $[0,1]/K$ is also homeomorphic to the Hawaiian earring. [The complement is a countable disjoint union of open intervals.] What about the quotient by identifying other compact subsets of $[0,1]$?

XIII.9.5.6. [Thu97] Identify the faces of two solid tetrahedra as in Figure XIII.16 (there is a unique identification consistent with the picture). Note that the six thick edges are all identified together, as are the six thin edges and all eight vertices.

(a) Show that every point of the identification space X has a neighborhood homeomorphic to $\mathbb{R}^3$ except the point v coming from the vertices. Thus, if $Y = X \setminus \{v\}$, then Y is a (noncompact) 3-manifold. If Z is X

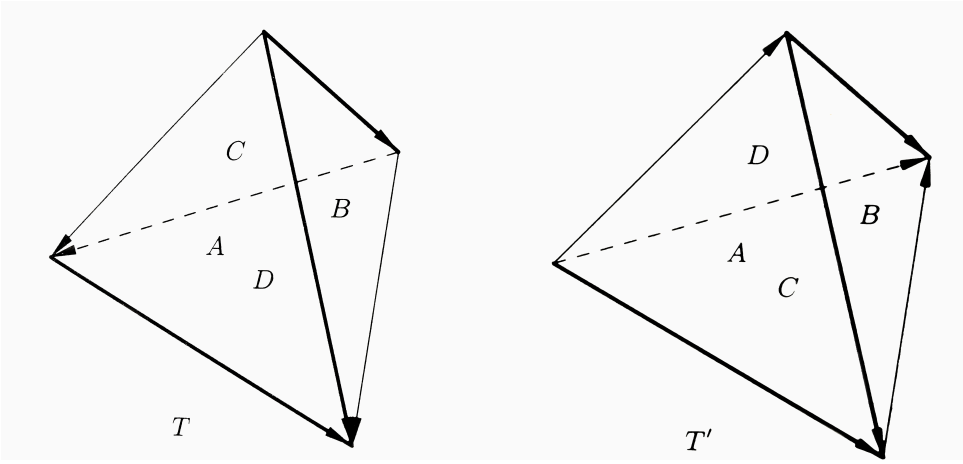


Figure XIII.16: Gluing scheme for two tetrahedra.

with a suitable small open neighborhood of v removed, then $Z \subseteq Y$ is a compact 3-manifold with boundary homeomorphic to a 2-torus, and there is a deformation retraction of Y onto Z.

(b) Show that Y is homeomorphic to the complement in S^3 of a figure 8 knot (), and Z is the complement of a tubular neighborhood of the knot.

(c) [Thu97, p. 129-132] Show similarly that there are two gluing schemes on a pair of octahedra with vertices removed, which respectively give the complement in S^3 of the Whitehead link (XIII.19.6.8.) and the Borromean rings (XIII.19.6.4.). Generalize.

XIII.9.5.7. [Eng89, 1.6.18] Consider the quotient space X obtained from $\mathbb{R}$ by collapsing $\mathbb{Z}$ to a point z (in the notation of XIII.9.1.1., this is $\mathbb{R}/\mathbb{Z}$, but this notation is generally reserved for the quotient group $\mathbb{T}$ as in XIII.9.3.8., quite a different space.)

(a) Show that X is perfectly normal (XIII.10.6.11.). In particular, $\{z\}$ is a G_δ.

(b) Show that z does not have a countable local base. Thus X is not first countable.

(c) Show that X is a Fréchet-Urysohn space (XIII.3.5.2.).

(d) Describe the topological differences, if any, between this space, the Hawaiian earring (XIII.19.5.8.), the space of XIII.19.5.9., an infinite wedge of circles (), and the one-point compactification of a countable separated union of copies of $\mathbb{R}$ (cf. XIII.9.5.8.).

XIII.9.5.8. (a) As a special case of XIII.8.5.1.(a), show that collapsing subsets of a space can be done in two or more steps: if (A_j) and (B_k) respectively are families of pairwise disjoint subsets of a topological space X and each A_j is a union of B_k's, the quotient space of X obtained by collapsing each A_j to a (separate) point can be obtained by first collapsing the B_k to (separate) points and then collapsing the set of points corresponding to each A_j to a point.

(b) As a special case, let X be a countable separated union of copies of $[0, 1]$, indexed by $\mathbb{Z}$. First identify the endpoints of the n'th interval to a single point $*_n$ to obtain a separated union of circles with distinguished

point. Then identify all the $*_n$ to a single point. Show that an infinite wedge of circles is obtained.

(c) On the other hand, first identify the right endpoint of the n'th interval with the left endpoint of the $(n+1)$'st interval for each n. Show that the quotient space is homeomorphic to $\mathbb{R}$ with the identified points corresponding to $\mathbb{Z}$. Then identify these points together to obtain the space from XIII.9.5.7..

(c) Conclude that both the space of XIII.9.5.7. and a (countably) infinite wedge of circles are homeomorphic to the quotient of X obtained by identifying all endpoints of all the intervals together to one point, and, in particular, the first two spaces are homeomorphic.

XIII.9.5.9. Let X and Y be compact Hausdorff spaces, A a closed subset of X, and f a continuous function from A onto a (necessarily closed) subset B of Y.

(a) Use XIII.8.3.13. to show that f is a quotient map from A to B.

(b) Show that the quotient map π from the separated union of X and Y to $X \cup_f Y$ is usc. [Adapt the proof of XIII.9.1.8..]

(c) Conclude that $X \cup_f Y$ is compact and Hausdorff. [Use XIII.8.3.8..]

XIII.9.5.10. Let X and Y be topological spaces, and $f : X \to Y$ a continuous function. Let A be the (closed) subspace $X \times \{1\}$ of $X \times [0,1]$, and define $g : A \to Y$ by $g((x,1)) = f(x)$. The identification space $(X \times [0,1]) \cup_g Y$ is called the *mapping cylinder* of f.

(a) Show that if X and Y are normal and f is a quotient map from X to $f(X)$, then the mapping cylinder of f is also normal.

(b) Show that if X and Y are compact Hausdorff spaces, so is the mapping cylinder of any map from X to Y. [Use XIII.8.3.13..]

(c) The *mapping torus* of $f : X \to Y$ is the space obtained from the mapping cylinder by identifying $(x,0)$ with $(f(x),1)$ for every $x \in X$. Show that the conclusions of (a) and (b) hold also for the mapping torus under the hypotheses there. The flow under a constant function (XIII.9.2.3.(ix)) is the mapping torus of the homeomorphism.

XIII.9.5.11. Let X be a metrizable space, and A a closed subspace of X. Show that the identification space X/A embeds in the cone CX over X. [Let ρ be a metric for X with $diam(X) = 1$ (), and identify $x \in X$ with $(x, 1 - \rho(x, A))$.]

XIII.10. Metrizable Spaces

Although the notion of topological space was designed to abstract the concept of a metric space in order to take specific reference to a metric out of the picture, metrizable spaces (topological spaces whose topology comes from a metric) have some nice properties not shared by more general topological spaces, and it is useful to know which topological spaces are metrizable. In this section we give (partial) answers to this question and pursue some of the special properties of metrizable spaces. Completely metrizable spaces (ones whose topology comes from a complete metric) are especially important and also have intrinsic characterizations.

Note that a metrizable space is *not* the same thing as a metric space: a metrizable space is a topological space for which there *exists* a metric (which is highly non-unique) giving the topology. A metric space can be regarded as a (metrizable) topological space with a specified metric giving its topology.

XIII.10.1. Metrizability

It is a nontrivial problem to determine exactly when a topology comes from a metric, i.e. when a topological space is metrizable. Necessary conditions are that the space be first countable and completely normal (XIII.10.1.2.), but these are not sufficient (). (Note that any subspace of a metrizable space is metrizable).

The first observation is that a metric space (X, ρ) is Hausdorff: if $x, y \in X$ $x \neq y$, and $\epsilon = \frac{\rho(x,y)}{3}$, then $B_\epsilon(x)$ and $B_\epsilon(y)$ are disjoint neighborhoods of x and y. We also have the following strengthening of T_4:

XIII.10.1.1. Proposition. Let X be a metrizable space, $A, B \subseteq X$ with $\bar{A} \cap B = A \cap \bar{B} = \emptyset$ (such A and B are said to be *separated*). Then A and B have disjoint open neighborhoods in X.

Proof: Let ρ be a metric on X defining the given topology. The conditions on A and B are equivalent to: $\rho(x, B) > 0$ for all $x \in A$ and $\rho(x, A) > 0$ for all $x \in B$. Set

$$U = \{x \in X : \rho(x, A) < \rho(x, B)\}$$

$$V = \{x \in X : \rho(x, B) < \rho(x, A)\}$$

and note that $A \subseteq U$, $B \subseteq V$. An easy application of the triangle inequality shows that U and V are open, and they are obviously disjoint.

XIII.10.1.2. A T_1 topological space satisfying the conclusion of this proposition is called *completely normal*. A completely normal space is obviously normal. A topological space X is completely normal if and only if every subspace of X is normal (XIII.10.6.11.), and thus a completely normal space is sometimes called *hereditarily normal*. This property is also sometimes called T_5.

The first metrization theorem, and still probably the most useful, was due to Urysohn in [Ury25b]:

XIII.10.1.3. Theorem. [Urysohn Metrization Theorem] Every second countable regular topological space is metrizable.

Proof: A second countable regular space is Lindelöf (XIII.1.3.5.), hence normal (XIII.7.6.12.).

Let X be a second countable regular (hence normal) space, and let $\mathcal{U}$ be a countable base for the topology of X. For each pair (U, V) of sets in $\mathcal{U}$ with $\bar{V} \subseteq U$, let f_{UV} be a continuous function from X to $[0, 1]$ which is 1 on $\bar{V}$ and 0 on U^c; such a function exists by Urysohn's Lemma (XIII.7.6.13.). Let $\mathcal{C}$ be the set of all the f_{UV}. Then $\mathcal{C}$ is countable. If E is a closed set in X and $x \in X \setminus E$, then E^c is an open neighborhood of x, so there is a $U \in \mathcal{U}$ with $x \in U \subseteq E^c$. There is an open neighborhood W of x with $\bar{W} \subseteq U$ (XIII.7.5.1.), and there is a $V \in \mathcal{U}$ with $x \in V \subseteq W$. Then $\bar{V} \subseteq U$, and f_{UV} satisfies $f_{UV}(x) = 1$ and $f_{UV} = 0$ on E. Thus by the proof of XIII.7.8.4., X embeds into the cube $[0, 1]^{\mathcal{C}}$. Since $\mathcal{C}$ is countable, this cube is metrizable (). 👍

Note that the proof shows slightly more than the statement of the theorem: any second countable regular space embeds in the Hilbert cube ().

XIII.10.1.4. There are many other metrization theorems, some of which are quite recent, and there are now various known necessary and sufficient conditions for metrizability. We will state only one of these results here since too many additional definitions are required for the others, and the results are of limited use in analysis. See [SS95] or [Pea75] for a survey.

XIII.10.1.5. Theorem. [Nagata-Smirnov Theorem] [] A topological space $(X, \mathcal{T})$ is metrizable if and only if it is regular and has a sequence $(\mathcal{U}_n)$ of locally finite open covers (XIII.11.10.1.) whose union forms a base for $\mathcal{T}$ (such an X is *σ-locally finite*).

XIII.10.1.6. Corollary. [Dug78, p. 207] Let $(X, \mathcal{T})$ be a topological space. Then X is metrizable if and only if it is regular and has a sequence $(\mathcal{U}_n)$ of open covers such that if $\mathcal{V}$ is any finite open cover of X, $\mathcal{U}_n$ refines $\mathcal{V}$ for some n.

Metrization of Products and Separated Unions

It is fairly obvious that a finite product of metrizable spaces is metrizable (). It is not true that an arbitrary product of metrizable spaces is metrizable; an uncountable product of nontrivial Hausdorff spaces is not even first countable (XIII.6.3.2.). But a *countable* product of metrizable spaces is metrizable. And an arbitrary separated union of metrizable spaces is metrizable.

The technical device giving a simple proof of both facts is:

XIII.10.1.7. Proposition. Let ρ be a metric on a set X. Then there is a uniformly equivalent metric $\tilde{\rho}$ on X with $\tilde{\rho}(x, y) \leq 1$ for all $x, y \in X$. If ρ is complete, then $\tilde{\rho}$ is complete.

Proof: There are two standard ways to construct such a metric. One way to do it is to set

$$\tilde{\rho}(x, y) = \frac{\rho(x, y)}{1 + \rho(x, y)} \ .$$

The triangle inequality needs proof for this metric, which is nontrivial but not difficult (see below). There is, however, a different low-tech way to define $\tilde{\rho}$ which is perhaps less elegant but does the trick more simply: let

$$\tilde{\rho}(x,y) = \min(\rho(x,y), 1) \ .$$

Proving the triangle inequality for this version of $\tilde{\rho}$ is just a trivial case-by-case verification. For either definition, we have $\tilde{\rho}(x,y) \leq \rho(x,y)$ for all $x, y \in X$ and $\rho(x,y) \leq 2\tilde{\rho}(x,y)$ if $\rho(x,y) \leq 1$, from which it follows easily that ρ and $\tilde{\rho}$ are uniformly equivalent. Since uniformly equivalent metrics have the same Cauchy sequences, the completeness statement follows.

For the reader's edification, we give a proof of the triangle inequality for the metric

$$\tilde{\rho}(x,y) = \frac{\rho(x,y)}{1+\rho(x,y)} \ .$$

The proof is not difficult, but requires a little trick. Let $x, y, z \in X$. We may assume $x \neq z$. Then

$$\tilde{\rho}(x,z) = \frac{\rho(x,z)}{1+\rho(x,z)} = \frac{1}{\frac{1}{\rho(x,z)}+1} \leq \frac{1}{\frac{1}{\rho(x,y)+\rho(y,z)}+1}$$

$$= \frac{\rho(x,y)+\rho(y,z)}{1+\rho(x,y)+\rho(y,z)} = \frac{\rho(x,y)}{1+\rho(x,y)+\rho(y,z)} + \frac{\rho(y,z)}{1+\rho(x,y)+\rho(y,z)}$$

$$\leq \frac{\rho(x,y)}{1+\rho(x,y)} + \frac{\rho(y,z)}{1+\rho(y,z)} = \tilde{\rho}(x,y) + \tilde{\rho}(y,z) \ .$$



XIII.10.1.8. THEOREM. A countable product of metrizable spaces is metrizable in the product topology. A countable product of completely metrizable spaces is completely metrizable.

PROOF: Let (X_k) be a sequence of metrizable spaces, and for each k let ρ_k be a metric on X_k giving the topology on X_k, with the property that $\rho_k(x,y) \leq 1$ for all $x, y \in X_k$, with ρ_k complete if X_k is completely metrizable (XIII.10.1.7.).

On $X = \prod_{k\in\mathbb{N}} X_k$, define

$$\rho(x,y) = \sum_{k=1}^{\infty} 2^{-k}\rho_k(x_k, y_k)$$

for $x = (x_1, x_2, \dots)$ and $y = (y_1, y_2, \dots)$ in X. Note that the infinite series converges absolutely for any x, y. It is simple and routine to verify that ρ is a metric on X. So we need to show that ρ defines the product topology. It suffices to show that, for fixed $x = (x_1, x_2, \dots) \in X$, every ρ-ball around x contains an open rectangle around x, and conversely.

Let $\epsilon > 0$. Fix n with $2^{-n+1} < \epsilon$. Then

$$B^{\rho_1}_{\epsilon/2}(x_1) \times B^{\rho_2}_{\epsilon/2}(x_2) \times \cdots \times B^{\rho_n}_{\epsilon/2}(x_n) \times \prod_{k=n+1}^{\infty} X_k$$

is an open rectangle in X containing x and contained in $B^\rho_\epsilon(x)$: if $y = (y_1, y_2, \dots) \in X$ with $\rho_k(x_k, y_k) < \frac{\epsilon}{2}$ for $1 \le k \le n$, then

$$\sum_{k=1}^{\infty} 2^{-k}\rho_k(x_k, y_k) = \sum_{k=1}^{n} 2^{-k}\rho_k(x_k, y_k) + \sum_{k=n+1}^{\infty} 2^{-k}\rho_k(x_k, y_k) < \frac{\epsilon}{2}\sum_{k=1}^{n} 2^{-k} + \sum_{k=n+1}^{\infty} 2^{-k} < \frac{\epsilon}{2} + 2^{-n} < \epsilon \ .$$

Conversely, if

$$V = U_1 \times U_2 \times \cdots U_m \times \prod_{k=n+1}^{\infty} X_k$$

is an open rectangle containing x, for $1 \le k \le m$ there is an ϵ_k such that $B^{\rho_k}_{\epsilon_k}(x_k) \subseteq U_k$. Set

$$\epsilon = \min(\epsilon_1/2, \epsilon_2/4, \dots, 2^{-m}\epsilon_m) \ .$$

If $y = (y_1, y_2, \dots) \in X$ with $\rho(x, y) < \epsilon$, then, for $1 \le k \le m$

$$\rho_k(x_k, y_k) < 2^k\rho(x, y) < 2^k\epsilon \le \epsilon_k$$

so $y \in V$, i.e. $B^\rho_\epsilon(x) \subseteq V$.

If $(x^{(n)}) = ((x^{(n)}_k))$ is a Cauchy sequence in (X, ρ), the last inequality shows that $(x^{(n)}_k)$ is a Cauchy sequence in (X_k, ρ_k) for each k. If each ρ_k is complete, then these sequences all converge, and hence $(x^{(n)})$ also converges, so ρ is complete. 👍

This proof requires the Countable AC (but only to choose the ρ_k).

XIII.10.1.9. Proposition. An arbitrary separated union () of metrizable spaces is metrizable. An arbitrary separated union of completely metrizable spaces is completely metrizable.

Proof: Let $\{X_i : i \in I\}$ be a collection of metrizable spaces. Fix a metric ρ_i on X_i for each i giving the topology on X_i, with the property that $\rho_i(x, y) \le 1$ for all $x, y \in X_i$, with ρ_i complete if X_i is completely metrizable (XIII.10.1.7.). Let X be the separated union of the X_i. For $x, y \in X$, set $\rho(x, y) = \rho_i(x, y)$ if x and y are both in the same X_i, and $\rho(x, y) = 1$ otherwise. It is easily checked that ρ is a metric giving the topology of X, and that ρ is complete if each ρ_i is. 👍

This proof uses the AC; the case of I countable requires the Countable AC.

Metrizability of Compact Hausdorff Spaces

XIII.10.1.10. Theorem. Let X be a compact Hausdorff space. The following are equivalent:

(i) X is metrizable.

(ii) X is second countable.

(iii) X embeds in the Hilbert cube.

(iv) $C_{\mathbb{R}}(X)$ is separable (for the uniform norm).

PROOF: (i) ⇒ (ii): Let d be a metric on X. (X, d) is totally bounded, hence separable, and a separable metric space is second countable.

(ii) ⇒ (iii): Let $\{U_n : n \in \mathbb{N}\}$ be a countable base for the topology. For each n, m with $\bar{U}_n \subseteq U_m$, let $f_{n,m}$ be a continuous function from X to $[0, 1]$ with $f = 1$ on $\bar{U}_n$ and $f = 0$ on U_m^c (there is such an $f_{n,m}$ by Urysohn's Lemma since X is normal). Set $I = \{(n, m) : \bar{U}_n \subseteq U_m\} \subset \mathbb{N}^2$ and $Y = [0, 1]^I$ with the product topology. Then Y is homeomorphic to the Hilbert cube (unless I is finite). Define $\phi : X \to Y$ by $\phi(x)_{(n,m)} = f_{n,m}(x)$. Then ϕ is continuous by the coordinatewise criterion since each $f_{n,m}$ is continuous. If $x_1, x_2 \in X$, $x_1 \neq x_2$, let V be a neighborhood of x_1 not containing x_2, U_m a basic neighborhood of x_1 contained in V, and U_n a basic neighborhood of x_1 with $\bar{U}_n \subseteq U_m$ (there is such a U_n since X is regular). Then $f_{n,m}(x_1) = 1$ and $f_{n,m}(x_2) = 0$, so $\phi(x_1) \neq \phi(x_2)$ and thus ϕ is one-to-one. To show ϕ is a homeomorphism onto its image, suppose (x_i) is a net in X and $\phi(x_i) \to \phi(x)$ for some $x \in X$. If $x_i \not\to x$, then there is a neighborhood V of x which does not contain any x_i for a cofinal set of i. As in the previous part of the proof, there are n, m such that $x \in U_n \subseteq \bar{U}_n \subseteq U_m \subseteq V$; then $f_{n,m}(x) = 1$ but $f_{n,m}(x_i) = 0$ for i in a cofinal set, contradicting that $\phi(x_i) \to \phi(x)$. Thus $x_i \to x$.

(iii) ⇒ (i) since the Hilbert cube is metrizable.

(iii) ⇒ (iv): note first that $C_{\mathbb{R}}(\mathbb{H})$ is separable: polynomials in the coordinates with rational coefficients are dense by the Stone-Weierstrass Theorem. $C_{\mathbb{R}}(X)$ is a quotient of $C_{\mathbb{R}}(\mathbb{H})$ by restriction of functions (this map is surjective by the Tietze Extension Theorem), so $C_{\mathbb{R}}(X)$ is also separable.

(iv) ⇒ (iii): if $C_{\mathbb{R}}(X)$ is separable, the set $\mathcal{G}$ of continuous functions from X to $[0, 1]$ is also separable in the topology of uniform convergence. Let $\mathcal{F}$ be a countable dense subset of $\mathcal{G}$, and let $Y = [0, 1]^{\mathcal{F}}$. Then Y is homeomorphic to the Hilbert cube. Define $\phi : X \to Y$ by $\phi(x)_f = f(x)$. If $x \in X$ and V is any neighborhood of x, there is a $g \in \mathcal{G}$ with $g(x) = 1$ and $g = 0$ on V^c since X is completely regular, and g can be uniformly approximated arbitrarily closely by an $f \in \mathcal{F}$. The proof that ϕ is a homeomorphism onto its image is then almost identical to the proof of (ii) ⇒ (iii). 👍

XIII.10.1.11. COROLLARY. Let X and Y be compact Hausdorff spaces. If X is metrizable and there is a continuous function from X onto Y, then Y is also metrizable.

PROOF: Let $\phi : X \to Y$ be surjective. Then $f \mapsto f \circ \phi$ embeds $C_{\mathbb{R}}(Y)$ as a subspace of $C_{\mathbb{R}}(X)$. Since $C_{\mathbb{R}}(X)$ is separable and metrizable, so is $C_{\mathbb{R}}(Y)$. 👍

XIII.10.1.12. Note that the normed algebra $C(X)$ of complex-valued continuous functions on X is separable if and only if $C_{\mathbb{R}}(X)$ is separable. Thus, if X is a compact Hausdorff space, $C(X)$ is separable if and only if X is metrizable.

XIII.10.1.13. Compactness is essential in this result: it fails in general even if X is locally compact and second countable, and Y is perfectly normal. There is a countable perfectly normal space Y which is not first countable (), hence not metrizable, and Y is a continuous image of $\mathbb{N}$ with the discrete topology. There is even a counterexample if f is a closed map (XIII.9.5.7.). The problem seems to be that a single point can be a G_δ even if the point does not have a countable neighborhood base, which cannot happen in a compact space (XIII.11.4.6.).

XIII.10.2. Continuity and Oscillation

We can generalize the notion of oscillation discussed in ().

XIII.10.2.1. DEFINITION. Let X be a set, (Y, σ) a metric space, and $f : X \to Y$ a function.
(i) If $\emptyset \neq A \subseteq X$, define the *oscillation of f on A* to be

$$\omega(f, A) = \sup sigma(f(x_1), f(x_2)| : x_1, x_2 \in A\} \in [0, +\infty] \ .$$

(ii) If X is a topological space and $x \in X$, define the *oscillation of f at x* to be

$$\omega(f, x) = \inf\{\omega(f, U) : U \text{ a neighborhood of } x\} \in [0, +\infty] \ .$$

The next characterization is obvious.

XIII.10.2.2. PROPOSITION. Let X be a topological space, (Y, σ) a metric space, and $f : X \to Y$. Then f is continuous at $x \in X$ if and only if $\omega(f, x) = 0$.

XIII.10.2.3. The oscillation of a function depends on the metric on Y. However, if the metric is replaced by an equivalent metric, whether the oscillation of a function on a set or at a point is zero or not does not change. Thus in XIII.10.2.2., the requirement that Y be a metric space can be relaxed to just the requirement that Y be metrizable. This makes sense since continuity of f depends only on the topology of Y, not on the specific metric.

The next observation has important consequences:

XIII.10.2.4. COROLLARY. Let X be a topological space, (Y, σ) a metric space, and $f : X \to Y$. Then the set of points of X at which f is continuous is a G_δ.

PROOF: It is easy to check that $\{x \in X : \omega(f, x) < \epsilon\}$ is open in X for any $\epsilon > 0$. 👍

If (in the case where X is a topological space) the function f is only defined on a subset A of X, we can extend the definition of oscillation at a point as follows:

XIII.10.2.5. DEFINITION. Let X be a topological space, (Y, σ) a metric space, $A \subseteq X$, and $f : A \to Y$ a function. Let $x \in \bar{A}$. Then the *oscillation* of f at x is

$$\omega(f, x) = \inf\{\omega(f, U \cap A) : U \text{ a neighborhood of } x\} \in [0, +\infty] \ .$$

Note that this definition only makes sense if $x \in \bar{A}$, since otherwise there will be a neighborhood U of x with $U \cap A = \emptyset$.

Corollary XIII.10.2.4. extends to this situation:

XIII.10.2.6. COROLLARY. Let X be a topological space, (Y, σ) a metric space, $A \subseteq X$, and $f : A \to Y$. Then the set of points x of $\bar{A}$ at which $\omega(f, x) = 0$ is a G_δ in $\bar{A}$ for the relative topology..

XIII.10.3. Completely Metrizable Spaces

Once metrizable spaces are characterized, there is a clean characterization of topological spaces which are homeomorphic to complete metric spaces (such spaces are called *completely metrizable*).

XIII.10.3.1. THEOREM. Let X be a topological space. The following are equivalent:

(i) X is completely metrizable.

(ii) X is metrizable, and whenever X is embedded in a completely metrizable space, it is a G_δ.

(iii) X is metrizable, and whenever X is embedded in a metrizable space, it is a G_δ.

XIII.10.3.2. It is trivial that (iii) $\Rightarrow$ (ii), and the converse is nearly trivial too: if X is embedded in a metrizable space Y, it is naturally embedded in the completion $\bar{Y}$. If (ii) holds, X is a G_δ in $\bar{Y}$, i.e. $X = \cap_n U_n$ for open sets U_n in $\bar{Y}$; but then $X = \cap_n (Y \cap U_n)$, so X is a G_δ in Y in the relative topology.

A topological space satisfying (ii) (hence (iii)) is called an *absolute* G_δ.

To prove the equivalence of (i) with (ii)–(iii), it is convenient to separate the two directions. Begin with (ii) $\Rightarrow$ (i):

XIII.10.3.3. PROPOSITION. Let X be a completely metrizable space, and E a G_δ in X. Then the relative topology on E is completely metrizable.

PROOF: Fix a complete metric ρ on X. Suppose E is a G_δ in X. We could simply write down an explicit formula for a complete metric on E giving the relative topology, but the formula is somewhat complicated and deserves some motivation.

Let U be an open subset of X, and for $x \in U$ define $g(x) = \frac{1}{\rho(x, U^c)}$. Then g is a continuous function from U to $\mathbb{R}$, and the graph

$$\Gamma = \{(x, g(x)) : x \in U\} \subseteq X \times \mathbb{R}$$

is a closed subset of $X \times \mathbb{R}$ which is homeomorphic to U via projection onto the first coordinate. $X \times \mathbb{R}$ is a complete metric space under the metric

$$\sigma((x, s), (y, t)) = \rho(x, y) + \min(|s - t|, 1) \ .$$

Hence the restriction of σ to Γ transfers to a complete metric on U, also denoted σ:

$$\sigma(x, y) = \rho(x, y) + \min(|g(x) - g(y)|, 1) \ .$$

Now write $E = \cap_n U_n$ with U_n open, and let g_n and σ_n be the corresponding function and metric for U_n. For $x, y \in E$, set

$$\tau(x,y) = \sum_{n=1}^{\infty} 2^{-n}\sigma_n(x,y) = \rho(x,y) + \sum_{n=1}^{\infty} 2^{-n}\min(|g_n(x) - g_n(y)|, 1) \ .$$

Then τ is a metric on E, and it is easily shown (Exercise ()) that τ induces the relative topology on E. If (x_k) is a τ-Cauchy sequence, then (x_k) is a ρ-Cauchy sequence, hence ρ-converges to some $x \in X$. Since (x_k) is also a σ_n-Cauchy sequence for each n, it σ_n-converges to some $x^{(n)} \in U_n$. But the σ_n-topology on U_n is the relative topology, so $x_k \to x^{(n)}$ in X and hence $x^{(n)} = x$, i.e. $x \in U_n$ for all n, so $x \in E$ and x_k τ-converges to x; thus (E, τ) is complete. 👍

For the converse (i) ⇒ (ii), we need a result of independent interest:

XIII.10.3.4. Proposition. Let X be a topological space and Y a completely metrizable space, and A a dense subspace of X. If $f : A \to Y$ is continuous, then f extends to a continuous function f^* from a G_δ subset A^* of X containing A to Y. If X is metrizable, the requirement that A be dense in X can be eliminated.

Proof: We first reduce the last statement to the first. If X is metrizable, then the closed subset $\bar{A}$ is a G_δ in X, and thus any G_δ in $\bar{A}$ (in the relative topology) is a G_δ in X. Replacing X by $\bar{A}$, we may thus assume A is dense in X.
Set $A^* = \{x \in X : \omega(f, x) = 0\}$; then A^* is a G_δ in X containing A. If $x \in A^*$, choose a sequence $(V_x^{(n)})$ of neighborhoods of x with $diam(f(A \cap V_x^{(n)})) < \frac{1}{n}$, and $x_n \in A \cap V_x^{(n)}$. Then $(f(x_n))$ is a τ-Cauchy sequence in Y, hence converges to an element of Y we call $f^*(x)$. The element $f^*(x)$ is independent of the choice of the x_n and even the $V_x^{(n)}$ (any neighborhood of x has nonempty intersection with A since A is dense in X). Then $f : A^* \to Y$ is continuous, and if $x \in A$, we have $f^*(x) = f(x)$. 👍

We now prove (i) ⇒ (ii) in XIII.10.3.1..
Proof: Suppose X is a completely metrizable subspace of a completely metrizable space Y (the metrics may be different!). Then $\bar{X}$ is a closed subspace of Y, hence a G_δ, so any G_δ in $\bar{X}$ is a G_δ in Y, i.e. it suffices to show that X is a G_δ in $\bar{X}$, so we may assume X is dense in Y. If g is the inclusion map from X to Y, let f be the "inverse" which is the identity map from the subspace X of Y to the completely metrizable space X. Then f extends to a continuous function f^* from a G_δ X^* in Y to X. We have that $g \circ f^*(x) = x$ for all x in the dense subspace X of X^*, hence $g \circ f^*(x) = x$ for all $x \in X^*$. But the range of g is X, so $X^* = X$. This completes the proof of XIII.10.3.1.. 👍

Note that the AC is used in the proof of XIII.10.3.4. and hence in XIII.10.3.1.. However, if (Y, τ) is σ-complete (), a simple modification of the proof along the lines of () removes any dependence on Choice.

XIII.10.3.5. Corollary. Let X and Y be metrizable spaces, with X completely metrizable, and $A \subseteq X$, $B \subseteq Y$. If A and B are homeomorphic and A is a G_δ in X, then B is a G_δ in Y.

Proof: A, hence B, is completely metrizable. 👍

Extension of Homeomorphisms

Here is another interesting consequence of XIII.10.3.4.:

XIII.10.3.6. THEOREM. Let A be a subset of a completely metrizable space X, and let f be a homeomorphism from A onto a subset B of a completely metrizable space Y. Then f extends to a homeomorphism from a G_δ subset of X containing A onto a G_δ subset of Y containing B.

PROOF: Replacing Y by $\bar{B}$, which is completely metrizable and a G_δ in Y, we may assume B is dense in Y. By XIII.10.3.4., f extends to a continuous function f^* from a G_δ A^* in X containing A to Y. By XIII.10.3.3., A^* is itself a completely metrizable space, and any G_δ in A^* is a G_δ in X. So, replacing X by A^*, we may assume that f extends to a continuous function f^* from X to Y. The continuous function $g = f^{-1}$ from B to A extends to a continuous function g^* from a G_δ B^* in Y to X. We have that $(f^* \circ g^*)(y) = y$ for all $y \in B$, hence for all $y \in B^*$ since B is dense in B^*. Thus f^* is a homeomorphism from $D = g^*(B^*)$ onto $f^*(D) = B^*$. Since D is homeomorphic to B^*, it is a G_δ in X by XIII.10.3.5.. 👍

We have the following remarkable and far-reaching analog (not exactly an extension) of XIII.10.3.5.:

XIII.10.3.7. COROLLARY. Let X and Y be completely metrizable spaces, and $A \subseteq X$, $B \subseteq Y$. If A and B are homeomorphic and A is a Borel set in X, then B is a Borel set in Y.

PROOF: Let $f : A \to B$ be a homeomorphism. Then f extends to a homeomorphism f^* from a G_δ A^* in X onto a G_δ B^* in Y. Since A^* is a Borel set in X, the (relative) Borel subsets of A^* are precisely the Borel sets in X contained in A^* (XVI.2.1.8.), and similarly for B^*. Since a homeomorphism sends Borel sets to Borel sets, the result follows. 👍

The following special case was proved by LUZIN using a different argument ():

XIII.10.3.8. COROLLARY. Let A be a Borel subset of $\mathbb{R}$. If B is a subset of $\mathbb{R}$ homeomorphic to A, then B is also a Borel set.

XIII.10.4. A Universal Space

We describe a "universal" separable complete metric space:

XIII.10.4.1. DEFINITION. Let $\mathcal{N} = \mathbb{N}^{\mathbb{N}}$ with the product topology, taking the discrete topology on $\mathbb{N}$.

$\mathcal{N}$ consists of all sequences of natural numbers. If (a_n) is a sequence in $\mathcal{N}$, say $a_n = (a_{n1}, a_{n2}, \cdots)$, then $a_n \to b = (b_1, b_2, \cdots)$ if and only if, for each k, there is an n_k such that $a_{nk} = b_k$ for all $n \geq n_k$.

Recall that set theorists typically write ω for $\mathbb{N}$ (actually $\mathbb{N} \cup \{0\}$), and in set theory $\mathcal{N}$ is often denoted ${}^{\omega}\omega$.

A notation we will use repeatedly: if $s = (s_1, s_2, \dots) \in \mathcal{N}$, write $s|n = (s_1, \dots, s_n) \in \mathbb{N}^n$. These are called *initial segments* of s. Take $s|0$ to be the "empty string."

XIII.10.4.2. PROPOSITION. $\mathcal{N}$ is separable.

PROOF: The set of sequences which are eventually 1 is countable and dense in $\mathcal{N}$. 👍

$\mathbb{N}$ is a complete metric space under the discrete metric $\delta(a,b) = 1$ if $a \neq b$ (). The topology on $\mathcal{N}$ also comes from a complete metric:

XIII.10.4.3. PROPOSITION. The function $\rho(a,b) = \sum_{k=1}^{\infty} 2^{-k}\delta(a_k,b_k)$ for $a = (a_k)$ and $b = (b_k)$ in $\mathcal{N}$ is a complete metric on $\mathcal{N}$ giving the product topology.

This is a special case of ().

XIII.10.4.4. PROPOSITION. (i) $\mathcal{N}$ is homeomorphic to $\mathbb{N} \times \mathcal{N}$, and to $\{1,\dots,n\} \times \mathcal{N}$ for any $n \in \mathbb{N}$. $\mathcal{N}$ is a disjoint union of a finite or countable sequence of closed subsets, each homeomorphic to $\mathcal{N}$.
(ii) $\mathcal{N}$ is homeomorphic to $\mathcal{N}^n$ for any $n \in \mathbb{N}$, and to $\mathcal{N}^{\mathbb{N}}$.

PROOF: (i): The map $(n,(a_1,a_2,\dots)) \mapsto (n,a_1,a_2,\dots)$ is a homeomorphism from $\mathbb{N} \times \mathcal{N}$ onto $\mathcal{N}$. If $n \in \mathbb{N}$, we have

$$\{1,\dots,n\} \times \mathcal{N} \cong \{1,\dots,n\} \times \mathbb{N} \times \mathcal{N} \cong \mathbb{N} \times \mathcal{N} \cong \mathcal{N}$$

taking a one-one correspondence between $\{1,\dots,n\} \times \mathbb{N}$ and $\mathbb{N}$. To obtain an explicit sequence of disjoint closed subsets homeomorphic to $\mathcal{N}$ partitioning $\mathcal{N}$, for each $n \in \mathbb{N}$ let $\mathcal{N}_n = \{a = (a_k) : a_1 = n\}$.
(ii): If $n \in \mathbb{N}$, then $\mathcal{N}^n \cong \mathbb{N}^{\mathbb{N}\times\{1,\dots,n\}}$, and similarly $\mathcal{N}^{\mathbb{N}} \cong \mathbb{N}^{\mathbb{N}\times\mathbb{N}}$. Fix one-one correspondences between $\mathbb{N} \times \{1,\dots,n\}$ and $\mathbb{N}$, and between $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$. 👍

XIII.10.4.5. REMARK. Note that this proposition does not depend on any form of the Axiom of Choice, since explicit one-one correspondences can be specified (cf. ().)

It follows from XIII.10.4.2. and XIII.10.4.3. that $\mathcal{N}$ is second countable. An explicit countable base for the topology for $\mathcal{N}$ can be given:

XIII.10.4.6. PROPOSITION. For each $m \in \mathbb{N}$ and each $(k_1,\dots,k_m,n_1,\dots,n_m) \in \mathbb{N}^{2m}$ with $k_1 < k_2 < \cdots k_m$, set

$$U_{(k_1,\dots,k_m,n_1,\dots,n_m)} = \{a = (a_k) : a_{k_j} = n_j \text{ for } 1 \leq j \leq m\} .$$

Then the $U_{(k_1,\dots,k_m,n_1,\dots,n_m)}$ form a base for the topology of $\mathcal{N}$.

The sets $U_{(k_1,\dots,k_m,n_1,\dots,n_m)}$ can be explicitly arranged in a sequence without using any form of the Axiom of Choice (Exercise ()).

The space $\mathcal{N}$ is universal among separable complete metric spaces in the following sense:

XIII.10.4.7. THEOREM. Let (X, σ) be a nonempty separable complete metric space. Then there is a locally contractive (hence continuous) map from $\mathcal{N}$ onto X.

PROOF: By induction on n, choose nonempty closed sets $F_{s_1,\dots,s_n}$ for each finite sequence $(s_1, \dots, s_n)$ in $\mathbb{N}$ by choosing a sequence F_k of nonempty closed sets (not necessarily distinct) of diameter ≤ 1 with $\cup_k F_k = X$, and if sets for sequences of length n have been chosen, choose nonempty closed sets $F_{s_1,\dots,s_n,k}$ for each $(s_1, \dots, s_n, k)$ of diameter $\leq 2^{-n}$ such that $\cup_k F_{s_1,\dots,s_n,k} = F_{s_1,\dots,s_n}$ for each $(s_1, \dots, s_n)$. (Such a collection of subsets is called a *Luzin scheme* in X.)

The $F_{s_1,\dots,s_n}$ can be chosen without any form of choice: fix a dense sequence (x_k) in X, let F_k be the closed ball of radius 2^{-1} around x_k, and at the $n+1$'st stage let $F_{s_1,\dots,s_n,k}$ be the intersection of $F_{s_1,\dots,s_n}$ with the closed ball of radius 2^{-n-1} around x_k if this intersection has nonempty interior, and otherwise the intersection of $F_{s_1,\dots,s_n}$ with the closed ball of radius 2^{-n-1} around x_j for the first j for which the intersection has nonempty interior. It is easily checked that these sets have the desired properties.

If $s = (s_1, s_2, \dots) \in \mathcal{N}$, there is a unique point in $\cap_n F_{s_1,\dots,s_n}$ by completeness; call this point $f(s)$. Then f is a map from $\mathcal{N}$ onto X. If $s, t \in \mathcal{N}$ with $\rho(s,t) < 1$, choose n so that $2^{-n} \leq \rho(s,t) < 2^{-n+1}$; then $s_j = t_j$ for $1 \leq j \leq n$, so $f(s), f(t) \in F_{s_1,\dots,s_n}$, $\sigma(f(s), f(t)) \leq 2^{-n}$, and f is locally contractive. 👍

Such a map can be explicitly described in some cases:

XIII.10.4.8. EXAMPLES. (i) Define $\phi : \mathbb{N} \to \{0,1\}$ by $\phi(1) = 1$ and $\phi(n) = 0$ for $n > 1$. Applying ϕ coordinatewise, we obtain a map from $\mathcal{N}$ onto $\{0,1\}^{\mathbb{N}}$ which is a contraction. Since there is a uniform homeomorphism between $\{0,1\}^{\mathbb{N}}$ and the Cantor set K, we obtain a Lipschitz function from $\mathcal{N}$ onto K. Also, there is a contractive mapping from $\{0,1\}^{\mathbb{N}}$ onto $[0,1]$ (), hence a contractive mapping from $\mathcal{N}$ onto $[0,1]$. Writing $\mathcal{N}$ as a disjoint union of closed sets $\mathcal{N}_n$ as in (), and taking the corresponding Lipschitz map from $\mathcal{N}_n$ onto $[n, n+1]$, we obtain a locally Lipschitz map from $\mathcal{N}$ onto $[0, +\infty)$. By taking a one-one correspondence between $\mathbb{N}$ and $\mathbb{Z}$, we obtain a similar map from $\mathcal{N}$ onto $\mathbb{R}$.

(ii) There is a homeomorphism from $\mathcal{N}$ onto the set $\mathbb{J}$ of irrational numbers given by continued fractions (): the sequence $(n_1, n_2, \dots)$ corresponds to the number with continued fraction $(n_1, n_2, \dots)$, i.e.

$$(n_1, n_2, \dots) \mapsto n_1 + \frac{1}{n_2 + \frac{1}{n_3 + \cdots}}$$

(this actually maps to the irrational numbers greater than 1; use a homeomorphism between $\mathcal{N}$ and $\mathcal{N} \times \{0,1\}$ to get a map onto all of $\mathbb{J}$).

XIII.10.4.9. There is an important extension of this result. Recall that if X is a topological space and $Y \subseteq X$, a *retraction* of X onto Y is a continuous function $f : X \to Y$ such that $f(y) = y$ for all $y \in Y$. Y is called a *retract* of X if there is a retraction from X onto Y. If X is Hausdorff, then a retract of X is necessarily a closed subset of X. It is relatively rare to have a retraction onto a general closed subset of a topological space; however, $\mathcal{N}$ has the following remarkable property:

XIII.10.4.10. THEOREM. Let Y be a nonempty closed subset of $\mathcal{N}$. Then there is a contractive retraction from $\mathcal{N}$ onto Y.

The proof is a slight refinement of the proof of XIII.10.4.7., and is left to the reader (Exercise ().) Alternatively, an easy proof can be given using the language of trees.

Trees

Trees can be described in various ways. For our purposes, we will take the following definition:

XIII.10.4.11. DEFINITION. A *tree* is a directed graph T for which:

(i) Every vertex (node) has a rank in $\mathbb{N} \cup \{0\}$.

(ii) Every edge goes from a node of rank n to a node of rank $n+1$ for some n.

(iii) There is a unique node of rank 0, called the *root*.

(iv) For every node x, there is a unique path (necessarily of length equal to the rank of x) from the root to x. In particular, every node of rank $n \geq 1$ has a unique predecessor of rank $n-1$.

A tree T is *pruned* if every node has at least one successor.

A *subtree* of a tree T is a subgraph which is itself a tree. A subtree is a *pruned subtree* if it is itself a pruned tree. (We also regard the empty subgraph as a subtree, which is a pruned subtree.)

XIII.10.4.12. EXAMPLE. Take the nodes of rank n to be $\{s|n : s \in \mathcal{N}\}$, and take an edge from $s|n$ to $s|(n+1)$ for each $s \in \mathcal{N}$ and $n \in \mathbb{N} \cup \{0\}$. This is a pruned tree, called the *tree over* $\mathbb{N}$, denoted $T_{\mathbb{N}}$. The root is the empty string, and $\mathcal{N}$ can be naturally identified with the infinite paths in $T_{\mathbb{N}}$.

The proof of the next proposition is a simple exercise which is left to the reader.

XIII.10.4.13. PROPOSITION. When $\mathcal{N}$ is identified with the set of infinite paths in $T_{\mathbb{N}}$, there is a natural one-one correspondence between the closed subsets of $\mathcal{N}$ and the pruned subtrees of $T_{\mathbb{N}}$, where the subset of $\mathcal{N}$ corresponding to a pruned subtree S is the set of infinite paths in S.

XIII.10.4.14. *Proof of Theorem XIII.10.4.10.:* In light of XIII.10.4.13., we need only show that if S is a nonempty pruned subtree of $T_{\mathbb{N}}$, there is a retraction ϕ of $T_{\mathbb{N}}$ onto S which preserves rank and edges. Define ϕ by induction on rank as follows. Let $\phi(s|0) = s|0$. Suppose $\phi(s|n)$ has been defined for each $s \in \mathcal{N}$. For each node $s|(n+1)$ (there are countably many distinct nodes of this type), if $s|(n+1) \in S$, set $\phi(s|(n+1)) = s|(n+1)$, and otherwise let $\phi(s|(n+1))$ be any successor of $\phi(s|n)$ in S (to make it definite and avoid using AC, let it be the successor node in S with smallest possible $(n+1)$'st coordinate). It is easily checked that this ϕ has the desired properties. □

XIII.10.5. Polish Spaces

XIII.10.5.1. Definition. A topological space is a *Polish space* if it is homeomorphic to a separable complete metric space.

Note that a Polish space is a topological space, not a metric space: it does not come with a specified metric, and even if it is given as a metric space, the given metric may not be complete – there need only be an equivalent complete metric (cf. XIII.10.5.3.). Any topological space homeomorphic to a Polish space is a Polish space.

XIII.10.5.2. Examples. (i) $\mathbb{R}^n$, or more generally any separable Banach space regarded as a topological space. An open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$, hence a Polish space.

(ii) Any (nonempty) closed subset of a Polish space; in particular, any closed subset of $\mathbb{R}^n$. (See XIII.10.5.3. for a generalization.)

(iii) Any second countable compact Hausdorff space ().

(iv) The space $\mathcal{N}$ (XIII.10.4.1.). More generally, a finite or countable Cartesian product of Polish spaces is Polish with the product topology (XIII.10.1.8.).

(v) A countable separated union of Polish spaces is a Polish space (XIII.10.1.9.).

A Polish space is second countable, Hausdorff, and normal, and is a Baire space (satisfies the conclusion of the Baire Category Theorem).

XIII.10.5.3. Theorem. A subset of a Polish space is a Polish space if and only if it is a G_δ. Thus a topological space is a Polish space if and only if it is a separable absolute G_δ.

Proof: This is a special case of XIII.10.3.1., with the observation that every subset of a separable metrizable space is separable.

Various Polish Topologies on a Set

We begin with an interesting fact which has a nice application when we discuss Borel sets (XVI.2.9.3.).

XIII.10.5.4. Theorem. Let X be a set. Let $(\mathcal{T}_n)$ be a sequence of Polish topologies on X. Suppose there is a Hausdorff topology $\mathcal{T}$ on X which is weaker than all the $\mathcal{T}_n$. If $\mathcal{T}_\infty$ is the weakest topology on X stronger than all the $\mathcal{T}_n$, then $\mathcal{T}_\infty$ is a Polish topology.

Proof: Let $Y = \prod_n (X_n, \mathcal{T}_n)$, where each $X_n = X$. Then Y is a Polish space with the product topology. For $x \in X$, set $\phi(x) = (x, x, x, \dots)$. If $y = (x_1, x_2, \dots) \in Y \setminus \phi(X)$, then there are m, n with $x_m \neq x_n$. Let U and V be disjoint $\mathcal{T}$-neighborhoods of x_m and x_n respectively. Set $W_m = U$, $W_n = V$, and $W_k = X$ for $k \neq m, n$, and $W = \prod W_k$. Then W is an open neighborhood of y in Y disjoint from $\phi(X)$. Thus $\phi(X)$ is closed in Y, and so is Polish in the relative topology, which when transfered back to X by ϕ^{-1} is easily seen to be $\mathcal{T}_\infty$.

We also have the following result of Kuratowski [**?**], which is a sharpening of XIII.10.4.7.. The proof is rather involved, although fairly elementary, and we omit it; see, for example, [**?**, 2.4.1].

XIII.10.5.5. THEOREM. Let X be an uncountable Polish space which is dense in itself. Then there is an injective continuous function from $\mathcal{N}$ onto X such that the image of every open set is an F_σ.

This theorem says that if $(X, \mathcal{T})$ is a perfect Polish space, there is a Polish topology $\mathcal{S}$ on X stronger than $\mathcal{T}$ such that $(X, \mathcal{S})$ is homeomorphic to $\mathcal{N}$ and every open set in the $\mathcal{S}$-topology is an F_σ in the $\mathcal{T}$-topology.

Since by the Cantor-Bendixson Theorem (XIII.2.5.4.) every Polish space is a disjoint union of a countable open subset and a closed subset which is dense-in-itself, we obtain:

XIII.10.5.6. COROLLARY. Every uncountable Polish space has cardinality $2^{\aleph_0}$.

How Many Polish Spaces Are There?

XIII.10.5.7. THEOREM. There are exactly $2^{\aleph_0}$ Polish spaces up to homeomorphism.

PROOF: Let $\mathcal{P}$ be the set of homeomorphism classes of infinite Polish spaces (there are countably many finite Polish spaces, which are just finite discrete spaces). There are $2^{\aleph_0}$ pairwise nonhomeomorphic compact subsets of $\mathbb{R}^2$ (XIII.19.5.1.), each of which is a Polish space, so $card(\mathcal{P}) \geq 2^{\aleph_0}$. On the other hand, if X is an infinite Polish space, ρ is a complete metric on X, and D is a countable dense set in X, then (X, ρ) is the completion of $(D, \rho|_D)$. If D is identified with $\mathbb{N}$, the metric $\rho|_D$ can be transferred to $\mathbb{N}$; thus X is homeomorphic to the completion of $\mathbb{N}$ with respect to some metric. So if $\mathcal{M}$ is the set of all metrics on $\mathbb{N}$, $card(\mathcal{M}) = 2^{\aleph_0}$ and there is a function from $\mathcal{M}$ onto $\mathcal{P}$. Therefore $card(\mathcal{P}) \leq 2^{\aleph_0}$. 👍

XIII.10.5.8. The above proof uses the AC (how?). The full AC is not needed: it suffices to assume that $\mathbb{R}$ can be well ordered, i.e. that $2^{\aleph_0}$ is an aleph. For the best that can be proved in ZF, see Exercise XIII.10.6.10..

XIII.10.6. Exercises

XIII.10.6.1. Arrange the tuples occurring as indexes in () into an explicit sequence, as follows:

(a) The set of pairs $((n_1, \dots, n_m), (k_1, \dots, k_m))$ ($m = 0$ is allowed, giving the "empty pair"), with the n_j strictly increasing in $\mathbb{N}$ and the $k_j \in \mathbb{N}$, is in natural one-one correspondence with the set $\mathcal{C}$ of sequences in $\mathbb{N} \cup \{0\}$ which are eventually 0: the pair $((n_1, \dots, n_m), (k_1, \dots, k_m))$ corresponds to the sequence with k_j in the n_j'th place and zeroes elsewhere.

(b) The set $\mathcal{C}$ is in one-one correspondence with $\mathbb{N}$:

$$(t_1, t_2, \dots) \longleftrightarrow \prod_j p_j^{t_j}$$

where p_j is the j'th prime.

XIII.10.6.2. In the proof of XIII.10.5.3., show that τ gives the relative topology on E. [Hint: Suppose $(x_m) \subseteq E$ ρ-converges to $x \in E$. Given $\epsilon > 0$, fix n so that $2^{-n} < \epsilon/3$, and find M so that $\rho(x_m, x) < \epsilon/3$ and $|g_k(x_m) - g_k(x)| < \epsilon/3$ for all $m \geq M$ and $1 \leq k \leq n$.]

XIII.10.6.3. (a) Show that if K is a compact subset of $\mathcal{N}$, then for each $n \in \mathbb{N}$ there is a $k_n \in \mathbb{N}$ such that $s_n \leq k_n$ whenever $s = (s_1, s_2, \dots) \in K$. [Consider the image of K under the n'th coordinate map $\pi_n : \mathcal{N} \to \mathbb{N}$.]

(b) Show that if $K \subseteq \mathcal{N}$ is compact, then $\mathcal{N} \setminus K$ is dense in $\mathcal{N}$.

(c) Conclude from the Baire Category Theorem that $\mathcal{N}$ is not σ-compact.

XIII.10.6.4. If (X, ρ) is a separable metric space, show that X is homeomorphic to a subspace of the Hilbert cube $[0,1]^{\mathbb{N}}$, as follows.

(a) Let $\sigma = min(\rho, 1)$ be a $[0,1]$-valued metric equivalent to ρ, and let (y_n) be a dense sequence in X. Define $f_n : X \to [0,1]$ by $f_n(x) = \sigma(x, y_n)$. Define $f : X \to [0,1]^{\mathbb{N}}$ by $f(x) = (f_1(x), f_2(x), \dots)$. Show that f is continuous.

(b) Show that f is a homeomorphism from X onto $f(X)$. [If $f(x_n) \to f(x)$, for $\epsilon > 0$ choose m so that $\sigma(x, y_m) < \epsilon/2$. Since $f_m(x_n) = \sigma(x_n, y_m) \to f_m(x) = \sigma(x, y_m) < \epsilon/2$, there is an N such that $f_m(x_n) < \epsilon/2$ for $n \geq N$, so $\sigma(x_n, x) < \epsilon$ for $n \geq N$.]

XIII.10.6.5. Show that every metrizable space is homeomorphic to a closed subspace of a product (uncountable in general) of completely metrizable spaces. [If X is metrizable with metric ρ, let $\bar{X}$ be the completion of X with respect to ρ, and consider

$$\prod_{y \in \bar{X} \setminus X} (\bar{X} \setminus \{y\}) \ .]$$

XIII.10.6.6. [Wil04, 23.2] Give a direct proof of XIII.10.1.11. as follows. Let X and Y be compact Hausdorff spaces, and $f : X \to Y$ a surjective continuous function. Suppose X is second countable. Let $\mathcal{B}$ be a countable base for the topology of X which is closed under finite unions. Set

$$\mathcal{D} = \{Y \setminus f(X \setminus B) : B \in \mathcal{B}\} \ .$$

Show that $\mathcal{D}$ is a countable base for the topology of Y.

XIII.10.6.7. **The Hausdorff Metric.** Let (X, ρ) be a metric space, and let $\mathcal{K}(X)$ be the set of nonempty compact subsets of X. For $x \in X$ and $F \in \mathcal{K}(X)$, set

$$\rho(x, F) = \inf_{y \in F} \rho(x, y) = \min_{y \in F} \rho(x, y) \ .$$

For $E, F \in \mathcal{K}(X)$, set

$$\rho_1(E, F) = \sup_{x \in E} \rho(x, F) = \max_{x \in E} \rho(x, F)$$

and

$$\rho_h(E, F) = \max(\rho_1(E, F), \rho_1(F, E)) \ .$$

This ρ_h is called the *Hausdorff metric* (or *Hausdorff distance*) on $\mathcal{K}(X)$. (It can be defined more generally on the space of closed bounded subsets of X, but not all of the same properties hold in this space. Such spaces of subsets are often called *hyperspaces*.)

(a) Show that ρ_h is a metric on $\mathcal{K}(X)$.

(b) Show that if ρ is complete, then ρ_h is also complete. [It suffices to show that every ρ_h-Cauchy sequence has a convergent subsequence. Let (E_n) be a sequence in $\mathcal{K}(X)$ with $\rho_h(E_n, E_{n+1}) < 2^{-n}$. If

$$D = \{x \in X : \exists x_n \in E_n \text{ with } x_n \to x \text{ and } \rho(x_n, x_{n+1}) < 2^{-n}\}$$

then $E = \bar{D}$ is nonempty and compact (totally bounded), and $E_n \to E$.]

(c) Show that if (X, ρ) is separable, then $(\mathcal{K}(X), \rho_h)$ is also separable. [Consider finite subsets of a countable dense set in X.]

(d) Show that if (X, ρ) is totally bounded, so is $(\mathcal{K}(X), \rho_h)$. [If $\mathcal{D}$ is a δ-cover of X, then $\mathcal{P}(\mathcal{D})$ is a δ-cover of $\mathcal{K}(X)$.] Conclude from (b) and () that if (X, ρ) is compact, so is $(\mathcal{K}(X), \rho_h)$.

(e) The map $x \mapsto \{x\}$ is an isometry from (X, ρ) onto a closed subset of $(\mathcal{K}(X), \rho_h)$. Deduce the converses of the statements in (b), (c), and (d).

(f) If $E, E_1, E_2, \cdots \in \mathcal{K}(X)$ and $E_n \to E$, then $\cap_n E_n \subseteq E \subseteq \overline{\cup_n E_n}$. If Y is a closed subset of X, then $\mathcal{K}(Y)$ is a closed subset of $\mathcal{K}(X)$. If Y is compact and $\mathcal{K}(X)_Y = \{E \in \mathcal{K}(X) : Y \subseteq E\}$, then $\mathcal{K}(X)_Y$ is also a closed subset of $\mathcal{K}(X)$. In particular, if $p \in X$, then

$$\mathcal{K}(X)_p = \mathcal{K}(X)_{\{p\}} = \{E \in \mathcal{K}(X) : p \in E\}$$

is a closed subset of $\mathcal{K}(X)$.

(g) Let $X = \{\frac{1}{n} : n \in \mathbb{N}\} \cup \{0\} \subseteq [0, 1]$ with the metric $\rho(x, y) = |x - y|$. Show that $\mathcal{K}(X)_0$ is homeomorphic to the Cantor set. [A compact subset of X containing 0 is of the form $E_S = \{\frac{1}{n} : n \in S\} \cup \{0\}$ for a subset S of $\mathbb{N}$. Send E_S to the usual sequence in $\{0, 1\}^{\mathbb{N}}$ associated to S as in ().] The complement of $\mathcal{K}(X)_0$ in $\mathcal{K}(X)$ is countable. Describe the topological space $\mathcal{K}(X)$.

(h) $\mathcal{K}([0, 1])$ is topologically infinite-dimensional. [If $n \in \mathbb{N}$ and $x_1, \ldots, x_n$ are distinct numbers in $(0, 1)$, then there is a neighborhood U of $(x_1, \ldots, x_n)$ in $\mathbb{R}^n$ such that $(y_1, \ldots, y_n) \mapsto \{y_1, \ldots, y_n\}$ is a homeomorphism from U into $\mathcal{K}([0, 1])$.] In fact, $\mathcal{K}([0, 1])$ is homeomorphic to the Hilbert cube [?], as is $\mathcal{K}(X)$ for any compact, connected, locally connected metric space X [CS78].

(i) Show that $(\mathcal{K}(X), \rho_h)$ is connected [*resp.* totally disconnected] if and only if (X, ρ) is connected [*resp.* totally disconnected].

XIII.10.6.8. The Hausdorff Metric Continued. [Mic51] Use the notation of Exercise XIII.10.6.7.. Let (X, ρ) be a metric space.

(a) Show that if $E, E_1, E_2, \cdots \in \mathcal{K}(X)$, then $E_n \to E$ if and only if, for every finite set $\{U_1, \ldots, U_n\}$ of open subsets of X with $E \subseteq \cup_k U_k$ and $E \cap U_k \neq \emptyset$ for all k, we have, for all sufficiently large n, $E_n \subseteq \cup_k U_k$ and $E_n \cap U_k \neq \emptyset$ for all k. (A topology can be defined on $\mathcal{K}(X)$ in this way for any topological space X; this topology is called the *finite topology* or *Vietoris topology* on $\mathcal{K}(X)$.)

(b) Let (Y, σ) be another metric space, and $f : X \to Y$ a continuous function. Define $f_* : \mathcal{K}(X) \to \mathcal{K}(Y)$ by setting $f_*(E) = f(E)$ for $E \in \mathcal{K}(X)$ (recall () that $f(E)$ is compact). Show that $f_* : (\mathcal{K}(X), \rho_h) \to (\mathcal{K}(Y), \sigma_h)$ is continuous.

(c) Show that if σ is another metric on X which is equivalent to ρ, then σ_h is equivalent to ρ_h. So the topology on $\mathcal{K}(X)$ depends only on the topology of X, not on the choice of metric.

(d) Show that if $E_n \to E$ in $\mathcal{K}(X)$, then $F = E \cup [\cup_{n=1}^{\infty} E_n]$ is compact in X. [If $\mathcal{U}$ is an open cover of F, let $\{U_1, \ldots, U_m\}$ be a finite subcover of E. Then $E_n \subseteq \cup_{k=1}^{m} U_k$ for all but finitely many n.] More generally, if $\mathcal{E}$ is a compact subset of $\mathcal{K}(X)$, then $\cup_{E \in \mathcal{E}} E$ is a compact subset of X [Mic51, 2.5.1].

XIII.10.6.9. (a) Show that there is a subset of $\mathcal{N}$ homeomorphic to the Cantor set. [Consider all sequences of 1's and 2's.]

(b) Since $\mathcal{N}$ is homeomorphic to $\mathbb{J}$ (XIII.10.4.8.(ii)), there is a subset of $\mathbb{J}$ homeomorphic to the Cantor set. Since this set is compact, it is closed and bounded in $\mathbb{R}$.

(c) If $(X, \mathcal{T})$ is a perfect Polish space, use the function $f : \mathcal{N} \to X$ of XIII.10.5.5. to show that X contains a subset homeomorphic to the Cantor set. [The image of a Cantor set in $\mathcal{N}$ is compact in the relative topology from $\mathcal{N}$ and Hausdorff in the relative $\mathcal{T}$-topology, so these topologies coincide.]

(d) Give a direct constructive proof that every perfect completely metrizable space contains a copy of the Cantor set ().

XIII.10.6.10. Let $\mathcal{P}$ be the set of homeomorphism classes of Polish spaces, and $\kappa = card(\mathcal{P})$. Show that the following weak version of XIII.10.5.7. is the best which can be proved in ZF [recall (II.9.9.9.) that it cannot be proved in ZF that $\aleph_1$ and $2^{\aleph_0}$ are comparable.]

THEOREM.

(i) $\kappa \geq \aleph_1 + 2^{\aleph_0}$.

(ii) $\kappa \precsim 2^{\aleph_0}$ (II.9.4.2.).

(iii) $\kappa < 2^{2^{\aleph_0}}$.

The arguments are the ones from the proof of XIII.10.5.7.. For (i), note that by XIII.19.5.4. there are also $\aleph_1$ pairwise nonhomeomorphic compact subsets of $\mathbb{R}$, none of which are homeomorphic to the ones of XIII.19.5.1.. For (iii), use II.9.9.7. and II.5.3.3..

XIII.10.6.11. Prove the following theorem:

THEOREM. Let X be a (necessarily T_1) topological space. The following are equivalent:

(i) X is completely normal (XIII.10.1.2.).

(ii) Every open subspace of X is normal.

(iii) Every subspace of X is normal.

(iv) Every subspace of X is completely normal.

(a) Show that (i) $\Rightarrow$ (iv).

(b) For (ii) $\Rightarrow$ (i), suppose $A, B \subseteq X$ with $A \cap \bar{B} = \bar{A} \cap B = \emptyset$ and $W = X \setminus (\bar{A} \cap \bar{B})$ normal. Show A and B have disjoint neighborhoods in W, hence in X.

(c) A topological space X is *perfectly normal* if it is normal and every closed set in X is a G_δ (equivalently, every closed set in X is a zero set (), or that every open set in X is an F_σ). Use the theorem and XIII.7.10.8.(b) to show that every perfectly normal space is completely normal. Show that every metrizable space is perfectly normal. There are completely normal spaces which are not perfectly normal, e.g. the one-point compactification of an uncountable discrete space. A space which is perfectly normal is sometimes called T_6.

(d) Since there are completely regular spaces, even locally compact Hausdorff spaces, which are not normal (), not every compact Hausdorff space is completely normal.

XIII.10.6.12. Show that a T_1 topological space X is completely normal if and only if, whenever Y_1 and Y_2 are closed subspaces and $Y = Y_1 \cup Y_2$, there are closed subspaces X_1 and X_2 of X with $X_1 \cup X_2 = X$ and $X_j \cap Y = Y_j$ for $j = 1, 2$. [Set $A = Y_1 \setminus Y_2$ and $B = Y_2 \setminus Y_1$; then A and B are separated. If U and V are disjoint open neighborhoods, set $X_1 = V^c$ and $X_2 = U^c$. Conversely, if A and B are separated, set $Y_1 = \bar{A}$ and $Y_2 = \bar{B}$.]

Note that this condition for Y_1 and Y_2 disjoint is exactly normality.

XIII.10.6.13. **Another Universal Polish Space.** The space $\mathcal{N}$ of XIII.10.4.1. is a universal Polish space in the sense of XIII.10.4.7.. There is a "dual" universal Polish space, the product of a countable number of copies of $\mathbb{R}$ with the product topology from the usual topology, denoted $\mathbb{R}^{\mathbb{N}}$.

(a) Show that $\mathbb{R}^{\mathbb{N}}$ is a Polish space (cf. XIII.10.5.2.(iv)).

(b) Every Polish space is homeomorphic to a closed subset of $\mathbb{R}^{\mathbb{N}}$ ().

(c) $\mathbb{R}^{\mathbb{N}}$ is a topological vector space under coordinatewise operations (), and in fact a Fréchet space.

(d) Every separable infinite-dimensional Banach space (or, more generally, Fréchet space) is [nonlinearly!] homeomorphic to $\mathbb{R}^{\mathbb{N}}$ ([Kad66], [And66], [AB68]; cf. [BP75]). As a result, topologists often call $\mathbb{R}^{\mathbb{N}}$ "Hilbert space" and denote it ℓ^2.

(e) Show that $\mathbb{R}^{\mathbb{N}}$ is not σ-compact (cf. XIII.10.6.3.). Conclude that no infinite-dimensional Banach space is σ-compact.

(f) If A is a σ-compact subset of $\mathbb{R}^{\mathbb{N}}$, then $\mathbb{R}^{\mathbb{N}} \setminus A$ is homeomorphic to $\mathbb{R}^{\mathbb{N}}$ (cf. e.g. [vM89, 6.5.9]). In particular, if B is a separable infinite-dimensional Banach space and p is a point of B, then $B \setminus \{p\}$ is homeomorphic to B.

(g) Let (I_n) be a sequence of nondegenerate intervals in $\mathbb{R}$. If infinitely many I_n are noncompact, then $\prod_n I_n$ is homeomorphic to $\mathbb{R}^{\mathbb{N}}$ [vM89, 6.5.10]. In particular, $[0,1)^{\mathbb{N}}$, $[0,1) \times \mathbb{R}^{\mathbb{N}}$, $[0,1] \times \mathbb{R}^{\mathbb{N}}$, and $\mathbb{H} \times \mathbb{R}^{\mathbb{N}}$ (where $\mathbb{H}$ is the Hilbert cube) are all homeomorphic to $\mathbb{R}^{\mathbb{N}}$.

XIII.11. Compactness

"Something like half of all mathematics consists of showing that some space is compact."

J. & B. Hubbard[1]

Compactness is one of the most important notions in topology, and especially for its applications in analysis. Compact sets behave roughly like finite sets. Compactness is used in applications to insure boundedness, the existence of limits, and the achievement of extrema, among other things.

Use of the Axiom of Choice is pervasive in the theory of compactness, and we will freely employ it; we will usually point out which results depend on it and which do not. Use of the Axiom of Choice can be almost entirely avoided if one sticks only to second countable topological spaces.

XIII.11.1. Basic Definitions

XIII.11.1.1. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. An *open cover* (*$\mathcal{T}$-open cover*) is a set $\mathcal{U} = \{U_i : i \in I\}$ of $\mathcal{T}$-open sets with $\cup_{i\in I} U_i = X$. The open cover $\mathcal{U}$ is *finite* if I is finite. A subset $\mathcal{V}$ of an open cover $\mathcal{U}$ is a *subcover* if it is itself an open cover.
If $Y \subseteq X$ and $\mathcal{U} = \{U_i : i \in I\} \subseteq \mathcal{T}$, then $\mathcal{U}$ is an open cover of Y in X if $Y \subseteq \cup_{i\in I} U_i$. Subcovers of Y in X are defined similarly.

XIII.11.1.2. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. $(X, \mathcal{T})$ is *compact* if every open cover of X has a finite subcover.
If $Y \subseteq X$, then Y is a *compact subspace* of X if every open cover of Y in X has a finite subcover in X.

XIII.11.1.3. **Caution:** In some older references, particularly from Eastern Europe, "compact" is used to mean "countably compact" (XIII.11.7.1.). A space which is compact according to the modern definition XIII.11.1.2. is called "bicompact" in these references. Also, in some references, especially French ones and [Eng89], the term "compact" is reserved for Hausdorff spaces; a space which is compact in the sense of XIII.11.1.2. but not necessarily Hausdorff is called "quasi-compact." In some references like [Pea75], "bicompact" means "compact and Hausdorff." Check the definitions in any reference. For brevity, we will call a compact Hausdorff space a *compactum*, plural *compacta* (some authors require a compactum to be metrizable, but we will not).

The first observation is that compactness of a subspace depends only on the relative topology of the subspace, not on the way it is embedded:

XIII.11.1.4. PROPOSITION. If Y is a subspace of a topological space $(X, \mathcal{T})$, then Y is a compact subspace of X if and only if $(Y, \mathcal{T}_Y)$ is compact.

PROOF: This is obvious from the fact that every relatively open subset of Y is of the form $U \cap Y$, where $U \in \mathcal{T}$. (The AC can be avoided in this argument as in ().)

The next result, an immediate corollary of XIII.11.2.2. and the Bolzano-Weierstrass theorem, gives the most important basic examples of compact spaces, and begins to give a picture of the meaning of compactness.

[1] [HH99, p. 89]. The comment is an intentional exaggeration, but not far wrong in spirit.

XIII.11.1.5. THEOREM. A subset of $\mathbb{R}^n$ is compact if and only if it is closed and bounded.

Here is a far-reaching version (not quite a generalization) of XIII.11.1.5.. Because of this, compact sets are often thought of as sets which are "absolutely closed" (although this is only really accurate in Hausdorff spaces). See XIII.11.4.1. and XIII.11.4.2. for generalizations of the second statement.

XIII.11.1.6. PROPOSITION. A closed subspace of a compact space is compact. A compact subspace of a Hausdorff space is closed.

PROOF: Let $(X, \mathcal{T})$ be a compact space, and Y a closed subset of X. Let $\mathcal{U}$ be an open cover of Y in X. Then $\mathcal{V} = \mathcal{U} \cup \{X \setminus Y\}$ is an open cover of X. Let $\{U_{i_1}, \ldots, U_{i_n}, X \setminus Y\}$ be a finite subcover of $\mathcal{V}$; then $\{U_{i_1}, \ldots, U_{i_n}\}$ is a finite subcover of the cover $\mathcal{U}$ of Y.

If $(X, \mathcal{T})$ is a Hausdorff space and Y is a compact subspace of X, and $x \in X \setminus Y$, for each $y \in Y$ let U_y and V_y be disjoint open neighborhoods (in X) of x and y respectively. Then $\{V_y : y \in Y\}$ is an open cover of Y in X. Let $\{V_{y_1}, \ldots, V_{y_n}\}$ be a finite subcover. Then $U = U_{y_1} \cap \cdots \cap U_{y_n}$ is an open neighborhood of x disjoint from Y. Thus $X \setminus Y$ is open. 👍

Note that the Axiom of Choice is used in the second half of the proof. The result is true, however, without any Choice assumption (Exercise XIII.11.14.2.).

A compact subspace of a general topological space, even a T_1-space, is not necessarily closed (Exercise XIII.11.14.1..)

A continuous image of a closed set is not necessarily closed. But one of the important properties of compactness is that it is preserved under continuous images:

XIII.11.1.7. PROPOSITION. Let $(X, \mathcal{T})$ and $(Z, \mathcal{S})$ be topological spaces, and $f : X \to Z$ a continuous function. If Y is a compact subset of X, then $f(Y)$ is a compact subset of Z.

PROOF: Let $\mathcal{V} = \{V_i : i \in I\}$ be an open cover of $f(Y)$ in Z. For each i, let $U_i = f^{-1}(V_i)$. Then $\mathcal{U} = \{U_i : i \in I\}$ is an open cover of Y in X. If $\{U_{i_1}, \ldots, U_{i_n}\}$ is a finite subcover of Y, then $\{V_{i_1}, \ldots, V_{i_n}\}$ is a finite subcover of $f(Y)$. 👍

An important corollary generalizes the Extreme Value Theorem of calculus:

XIII.11.1.8. COROLLARY. Let X be a topological space, $f : X \to \mathbb{R}$ a continuous function, and E a compact subset of X. Then f is bounded on E, and attains its maximum and minimum values on E.

PROOF: By XIII.11.1.7., $f(E)$ is compact in $\mathbb{R}$, and therefore closed and bounded by XIII.11.1.5.. 👍

There is a useful rephrasing of the definition of compactness, using the Finite Intersection Property. This characterization of compactness is strongly reminiscent of the Nested Intervals Property of $\mathbb{R}$.

XIII.11.1.9. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. A (nonempty) collection $\mathcal{F} = \{F_i : i \in I\}$ of closed subsets of X has the *Finite Intersection Property* if, whenever $F_{i_1}, \ldots, F_{i_n}$ is a finite collection of sets in $\mathcal{F}$, we have $F_{i_1} \cap \cdots \cap F_{i_n} \neq \emptyset$.

XIII.11.1.10. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space. Then $(X, \mathcal{T})$ is compact if and only if, whenever $\mathcal{F} = \{F_i : i \in I\}$ is a nonempty collection of closed sets in X with the Finite Intersection Property, we have $\cap_{i \in I} F_i \neq \emptyset$.

PROOF: Suppose $(X, \mathcal{T})$ is compact, and let $\mathcal{F} = \{F_i : i \in I\}$ be a collection of closed sets in X with $\cap_{i \in I} F_i = \emptyset$. Let $U_i = X \setminus F_i$ for each i. Then $\mathcal{U} = \{U_i : i \in I\}$ is an open cover of X. Let $\{U_{i_1}, \ldots, U_{i_n}\}$ be a finite subcover. Then $F_{i_1} \cap \cdots \cap F_{i_n} = \emptyset$, so $\mathcal{F}$ does not have the Finite Intersection Property.

Conversely, suppose every collection of closed sets in X with the Finite Intersection Property has a nonempty intersection, and let $\mathcal{U} = \{U_i : i \in I\}$ be an open cover of X. Set $F_i = X \setminus U_i$ for each i. Then $\mathcal{F} = \{F_i : i \in I\}$ is a collection of closed subsets of X with $\cap_{i \in I} F_i = \emptyset$, so $\mathcal{F}$ does not have the Finite Intersection Property, i.e. there is a finite set $\{F_{i_1}, \ldots, F_{i_n}\}$ with $F_{i_1} \cap \cdots \cap F_{i_n} = \emptyset$. But then $\{U_{i_1}, \ldots, U_{i_n}\}$ is a finite subcover of $\mathcal{U}$. 👍

Here is a useful variation:

XIII.11.1.11. PROPOSITION. Let X be a topological space, $\{E_i : i \in I\}$ a collection of closed subsets of X, with at least one compact. If U is an open subset of X containing $\cap_{i \in I} E_i$, then there is a finite subset F of I such that $\cap_{i \in F} E_i \subseteq U$.

PROOF: Fix an i_0 with E_{i_0} compact. For each i, let $K_i = E_i \cap E_{i_0} \cap U^c$. Then the K_i are closed subsets of E_{i_0}, and they have empty intersection, so they do not have the Finite Intersection Property, i.e. there are $i_1, \ldots, i_n$ such that $K_{i_1} \cap \cdots \cap K_{i_n} = \emptyset$. Then $E_{i_0} \cap E_{i_1} \cap \cdots \cap E_{i_n} \subseteq U$. 👍

XIII.11.1.12. COROLLARY. Let X be a topological space. Let (F_n) be a decreasing sequence of closed compact subsets of X ("closed" is redundant if X is Hausdorff) with $F = \cap_n F_n$. If U is any open set in X with $F \subseteq U$, then $F_n \subseteq U$ for all sufficiently large n.

Here is an interesting and important special case:

XIII.11.1.13. COROLLARY. Let X be a topological space, $p \in X$. Let (V_n) be a decreasing sequence of closed compact neighborhoods of p ("closed" is redundant if X is Hausdorff) with $\cap_n V_n = \{p\}$. Then the V_n form a local neighborhood base at p for the topology.

XIII.11.2. Compactness and Convergence

The next theorem gives the main equivalent ways compactness is characterized:

XIII.11.2.1. THEOREM. Let $(X, \mathcal{T})$ be a topological space. The following are equivalent:

(i) $(X, \mathcal{T})$ is compact.

(ii) Every nonempty collection of closed subsets of X with the Finite Intersection Property has nonempty intersection.

(iii) Every net in X has a cluster point.

(iv) Every net in X has a convergent subnet.

(v) Every filter on X has a cluster point.

PROOF: (i) $\iff$ (ii) is XIII.11.1.10., (iii) $\iff$ (iv) is (), and (iii) $\iff$ (v) is ().
(ii) $\Rightarrow$ (iii): Let $(x_i)_{i\in I}$ be a net in x. For each i, let F_i be the closure of $\{x_j : j \geq i\}$, and let $\mathcal{F} = \{F_i : i \in I\}$. If $i_1, \ldots, i_n \in I$, then since I is directed there is an $i_0 \geq i_k$ for $1 \leq k \leq n$. Thus $x_{i_0} \in F_{i_1} \cap \cdots \cap F_{i_n}$, and hence $F_{i_1} \cap \cdots \cap F_{i_n} \neq \emptyset$ and so $\mathcal{F}$ has the Finite Intersection Property. By (ii), there is an $x \in \cap_{i\in I} F_i$, and x is a cluster point of the net (x_i).
(iii) $\Rightarrow$ (ii): Let $\mathcal{F}$ be a nonempty collection of closed sets with the Finite Intersection Property. We may assume that $\mathcal{F}$ is closed under finite intersections. Set

$$I = \{(x, F) \in X \times \mathcal{F} : x \in F\} .$$

Put a preorder on I by setting $(x_1, F_1) \leq (x_2, F_2)$ if $F_2 \subseteq F_1$. Then I is directed: for any $(x_1, F_1), (x_2, F_2)$ we have

$$(x_1, F_1), (x_2, F_2) \leq (x_3, F_1 \cap F_2)$$

where x_3 is any element of $F_1 \cap F_2$. Now, for $i = (x, F) \in I$, define $x_i = x$. Then (x_i) is a net in X. By (iii) this net has a cluster point y. If U is any neighborhood of y, and $F_0 \in \mathcal{F}$, and $x_0 \in F_0$, then there is an $i = (x, F)$ with $(x_0, F_0) \leq i$ (i.e. $F \subseteq F_0$) and $x_i = x \in U$. Since $x \in F \subseteq F_0$, we have that $U \cap F_0 \neq \emptyset$. Fixing F_0 and letting U vary, we conclude that y is a limit point of F_0, hence $y \in F_0$ since F_0 is closed. This is true for all $F_0 \in \mathcal{F}$, so $y \in \cap_{F\in\mathcal{F}} F$ and $\mathcal{F}$ has nonempty intersection. 👍

Note that no form of Choice is needed for this result.

XIII.11.2.2. COROLLARY. Let X be a second countable topological space. Then X is compact if and only if every sequence in X has a convergent subsequence (a topological space with this property is called *sequentially compact*).

This result is false in general for spaces which are just first countable, even for normal spaces: the space ω_1 of ordinals less than the first uncountable ordinal, with the order topology (), is a normal Hausdorff space which is first countable and sequentially compact, but not compact. A first countable topological space which is compact is sequentially compact, but $\beta\mathbb{N}$ (XIII.11.9.16.), which is not first countable, is compact but not sequentially compact. See XIII.11.7.3..

There is a variation of XIII.11.2.1.(v) which is useful. One direction of this result uses the Ultrafilter Property (), which is equivalent to the Boolean Prime Ideal Theorem ().

XIII.11.2.3. PROPOSITION. (BPI) Let $(X, \mathcal{T})$ be a topological space. Then $(X, \mathcal{T})$ is compact if and only if every ultrafilter on X converges.

PROOF: If X is compact, then by XIII.11.2.1. every ultrafilter on X has a cluster point, hence converges (). Conversely, suppose every ultrafilter on X converges. Let $\mathcal{F}$ be a filter on X. Then by the Ultrafilter Property () there is an ultrafilter $\mathcal{G}$ on X containing $\mathcal{F}$, which converges by assumption. The limit of $\mathcal{G}$ is a cluster point for $\mathcal{F}$, so X is compact by XIII.11.2.1.. 👍

XIII.11.3. Compact Metrizable Spaces

There is a very clean characterization of compact metric spaces, using the idea of total boundedness:

XIII.11.3.1. DEFINITION. Let ρ be a metric on a set X. Then ρ is *totally bounded* if for every $\epsilon > 0$, a finite number of open ρ-balls of radius ϵ cover X.

In other words, ρ is totally bounded if for every $\epsilon > 0$, there is a finite set of points $\{x_1, \ldots, x_n\}$ such that every point of X is within ϵ of at least one of the x_k.

A totally bounded metric is bounded: taking $\epsilon = 1$, if $\{x_1, \ldots, x_n\}$ are the centers of balls of radius 1 covering X, and M is the maximum of the distances between the x_k's, then any two points of X are distance less than $M + 2$ apart. But a bounded metric is not necessarily totally bounded: for example, the discrete metric () on an infinite set is bounded but not totally bounded.

XIII.11.3.2. PROPOSITION. A totally bounded metric space is separable (and hence second countable by ()).

PROOF: For each $m \in \mathbb{N}$ let $S_m = \{x_{m,1}, \ldots, x_{m,n_m}\}$ be a finite set of points of X such that the open balls of radius $\frac{1}{m}$ centered at the $x_{m,k}$ cover X. If $S = \cup_m S_m$, then S is a countable dense subset of X. 👍

XIII.11.3.3. PROPOSITION. Let (X, ρ) be a totally bounded metric space. Then every sequence in X has a Cauchy subsequence.

PROOF: For each m, choose a finite set $\{y_{k,m}\}$ in X such that every element of X is within $\frac{1}{2m}$ of one of the y_{km}.

Let (x_n) be a sequence in X. Since there are only finitely many $y_{k,1}$, there are infinitely many x_k within $\frac{1}{2}$ of some single $y_{k,1}$. Thus there is a subsequence $(x_{n,1})$ of (x_n) all of whose terms lie within $\frac{1}{2}$ of this y_{k1}, hence within 1 of each other. Successively let $(x_{n,m+1})$ be a subsequence of $(x_{n,m})$, all of whose terms lie within $\frac{1}{2(m+1)}$ of some $y_{k,m+1}$ and hence within $\frac{1}{m}$ of each other. Then the diagonal sequence $(x_{n,n})$ is a subsequence of (x_n) which is a Cauchy sequence. 👍

The Countable AC () is needed in this proof to choose the $y_{k,m}$, and DC () is needed to choose the subsequences.

The main characterization theorem is:

XIII.11.3.4. THEOREM. Let (X, ρ) be a metric space. Then X is compact (in the metric topology) if and only if ρ is complete and totally bounded.

PROOF: Suppose X is compact. Since X is first countable, every sequence in X has a convergent subsequence. In particular, if (x_n) is a Cauchy sequence, it must have a convergent subsequence, hence must converge (). Thus (X, ρ) is complete. If ρ is not totally bounded, there is an $\epsilon > 0$ such that no finite set of open balls of radius ϵ covers X, so a sequence (x_n) can be chosen so that $\rho(x_n, x_m) \geq \epsilon$ for all n and m. But such a sequence can have no convergent subsequence, a contradiction. Thus ρ is totally bounded.

Conversely, suppose ρ is complete and totally bounded. Suppose $\mathcal{U} = \{U_i : i \in I\}$ is an open cover of X. Since X is second countable (XIII.11.3.2.), it is Lindelöf (), so $\mathcal{U}$ has a countable subcover $\{U_1, U_2, \dots\}$. Suppose this cover does not have a finite subcover. For each n, let $x_n \in X \setminus \cup_{k=1}^{n} U_k$. By XIII.11.3.3., (x_n) has a Cauchy subsequence, which must converge to some x since (X, ρ) is complete. We have $x \in U_n$ for some n, but U_n is open and $x_k \notin U_n$ for all $k > n$, so no subsequence of (x_n) can converge to x, a contradiction. Thus there is a finite subcover, which is a subcover of $\mathcal{U}$. So X is compact. 👍

This result requires some Choice (DC suffices), cf. XIII.15.12.9..

Combining the previous results, we get:

XIII.11.3.5. COROLLARY. Let X be a compactum. Then X is metrizable if and only if it is second countable.

PROOF: If X is metrizable, it is second countable by XIII.11.3.4. and XIII.11.3.2.. Conversely, since a compactum is regular, if it is second countable it is metrizable by the Urysohn Metrization Theorem (). 👍

XIII.11.3.6. There are compacta which are both separable and first countable, but not second countable and hence not metrizable (e.g. XIII.11.14.4.).

Lebesgue Numbers of an Open Cover

The Lebesgue Property and Lebesgue numbers of an open cover are particularly relevant for compact metric spaces, but can be defined in general:

XIII.11.3.7. DEFINITION. Let $\mathcal{U}$ be an open cover of a metric space (X, ρ). $\mathcal{U}$ has the *Lebesgue Property* if there is a $\delta > 0$ such that, for every $x \in X$, there is a $U \in \mathcal{U}$ with $B_\delta(x) \subseteq U$. Such a δ is a *Lebesgue number* of the open cover $\mathcal{U}$.

Any positive number smaller than a Lebesgue number for $\mathcal{U}$ is a Lebesgue number for $\mathcal{U}$. The supremum of all Lebesgue numbers for $\mathcal{U}$ can be ∞, and if finite may or may not be a Lebesgue number for $\mathcal{U}$; thus $\mathcal{U}$ need not have a largest Lebesgue number. If $\mathcal{V}$ is a refinement of $\mathcal{U}$ (in particular, if it is a subcover of $\mathcal{U}$), any Lebesgue number of $\mathcal{V}$ is a Lebesgue number of $\mathcal{U}$. The Lebesgue Property and Lebesgue numbers depend on the metric and are not topological properties.

XIII.11.3.8. EXAMPLES. (i) In $\mathbb{R}$ with the usual metric, the open cover $\mathcal{U} = \{(-n, n) : n \in \mathbb{N}\}$ has the Lebesgue Property, and any positive number is a Lebesgue number for $\mathcal{U}$.

(ii) In $\mathbb{R}$ with the usual metric, the open cover $\mathcal{U}$ consisting of all open intervals of length less than 2 has the Lebesgue Property, and any positive number less than 1 is a Lebesgue number for $\mathcal{U}$. The supremum of the Lebesgue numbers is 1, but 1 is not a Lebesgue number for $\mathcal{U}$.

(iii) In $(0, 1)$ with the usual metric, the open cover $\left\{\left(\frac{1}{n}, 1\right) : n \in \mathbb{N}\right\}$ does not have the Lebesgue Property.

(iv) Even a finite open cover need not have the Lebesgue Property. Let $X = (0, 1)$ with the usual metric. Fix two disjoint decreasing sequences (x_n) and (y_n) in $(0, 1)$ converging to 0, and set $U = (0, 1) \setminus \{x_n : n \in \mathbb{N}\}$ and $V = (0, 1) \setminus \{y_n : n \in \mathbb{N}\}$. Then the open cover $\{U, V\}$ does not have the Lebesgue Property.

XIII.11.3.9. PROPOSITION. Let (X, ρ) be a metric space of bounded diameter d. If $\mathcal{U}$ is an open cover of X by proper subsets, and $\mathcal{U}$ has the Lebesgue property, then any Lebesgue number for $\mathcal{U}$ is $\leq d$. If $\mathcal{U}$ is a finite open cover of X by proper subsets, and $\mathcal{U}$ has the Lebesgue Property, then the supremum of all Lebesgue numbers for $\mathcal{U}$ is a Lebesgue number for $\mathcal{U}$, i.e. $\mathcal{U}$ has a largest Lebesgue number.

PROOF: For any $x \in X$ and $U \in \mathcal{U}$, there is a $y \in X$ with $y \notin U$. Since $\rho(x, y) \leq d$, any Lebesgue number of $\mathcal{U}$ must be $\leq d$. If $\mathcal{U} = \{U_1, \ldots, U_n\}$ is a finite open cover by proper subsets, and δ is the supremum of the Lebesgue numbers of $\mathcal{U}$ ($\delta < \infty$ by the first part of the proof), fix $x \in X$. Then there must be a k such that $B_{\delta - 1/m}(x) \subseteq U_k$ for infinitely many m; hence $B_\delta(x) \subseteq U_k$. So δ is a Lebesgue number for $\mathcal{U}$. 👍

The main result is:

XIII.11.3.10. PROPOSITION. Let (X, ρ) be a compact metric space. Then any open cover of X has the Lebesgue Property.

PROOF: Let $\mathcal{U}$ be an open cover of X. Suppose $\mathcal{U}$ does not have the Lebesgue property. Then for each n there are points $x_n, y_n \in X$ with $\rho(x_n, y_n) < \frac{1}{n}$, but for which there is no $U \in \mathcal{U}$ containing both x_n and y_n. Passing to a subsequence if necessary, we may assume (x_n) converges to a point $x \in X$. Then also $y_n \to x$. There is a $U \in \mathcal{U}$ with $x \in U$. Then x_n and y_n are both in U if n is sufficiently large, a contradiction. 👍

See XIII.11.14.3. for an alternate proof which gives an explicit Lebesgue number.

Here is a sample application:

XIII.11.3.11. COROLLARY. Let (X, ρ) and (Y, σ) be metric spaces, with X compact, and $F : X \to Y$ a continuous function. Then f is uniformly continuous.

PROOF: Let $\epsilon > 0$. For each $y \in Y$, let $U_y = f^{-1}(B_{\epsilon/2}(y))$. Then $\{U_y : y \in Y\}$ is an open cover of X. If $\delta > 0$ is the Lebesgue number of this open cover, and $x_1, x_2 \in X$ with $\rho(x_1, x_2) < \delta$, there is a y with $B_\delta(x_1) \subseteq U_y$; thus $f(x_1)$ and $f(x_2)$ are both in $B_{\epsilon/2}(y)$ and $\sigma(f(x_1), f(x_2)) < \epsilon$. 👍

XIII.11.4. Separation Properties

The results of this subsection give important senses in which compact sets behave like finite sets (see XIII.11.10.19. for another example). The first result is a strengthening of the second half of XIII.11.1.6., with essentially the same proof:

XIII.11.4.1. PROPOSITION. Let X be a Hausdorff space, E a compact subset of X, and y a point of $X \setminus E$. Then E and y have disjoint open neighborhoods.

PROOF: This is another example of an elementary result in topology whose standard proof uses an unnecessary application of the AC (cf. XIII.1.4.2.). We give a proof due to A. KARAGILA which avoids the AC, and is no more difficult than the standard one.

Let $\mathcal{U}$ be the collection of all open sets U in X which have the property that there is an open neighborhood of y disjoint from U. Then $\mathcal{U}$ covers E since X is Hausdorff. Let $\{U_1, \ldots, U_n\}$ be a finite subcover of E. For each k, let V_k be an open neighborhood of y disjoint from U_k. Then $U = \cup_{k=1}^n U_k$ and $V = \cap_{k=1}^n V_k$ are disjoint neighborhoods of E and y. 👍

There is a nice generalization:

XIII.11.4.2. PROPOSITION. Let X be a Hausdorff space, and E and F disjoint compact subsets of X. Then E and F have disjoint open neighborhoods.

PROOF: Let $\mathcal{U}$ be the collection of all open sets U in X such that there is an open neighborhood of F disjoint from U. Then $\mathcal{U}$ covers E by XIII.11.4.1.. Let $\{U_1, \ldots, U_n\}$ be a finite subcover of E. For each k, let V_k be an open neighborhood of F disjoint from U_k. Then $U = \cup_{k=1}^n U_k$ and $V = \cap_{k=1}^n V_k$ are disjoint neighborhoods of E and F. 👍

XIII.11.4.3. COROLLARY. A compactum is normal.

In an almost identical manner (Exercise ()), we can show:

XIII.11.4.4. PROPOSITION. Let X be a regular space, and E and F disjoint subsets of X, with E compact and F closed. Then E and F have disjoint open neighborhoods.

In fact, we have (no Choice needed):

XIII.11.4.5. THEOREM. A compactum is fully normal (XIII.7.7.5.).

PROOF: Let X be a compactum, and $\mathcal{U}$ an open cover of X. Then $\mathcal{U}$ has a finite subcover $\mathcal{U}_0$. The finite cover $\mathcal{U}_0$ has a closed shrinking since X is normal (XIII.11.4.3., XIII.7.6.8.). Thus XIII.7.7.8. applies to give a barycentric refinement of $\mathcal{U}_0$ and hence of $\mathcal{U}$. 👍

Note the following consequence of XIII.11.1.12.:

XIII.11.4.6. PROPOSITION. Let X be a compactum, and p a point of X. The following are equivalent:

(i) $\{p\}$ is a G_δ, and whenever (F_n) is a decreasing sequence of closed neighborhoods of p with $\cap_n F_n = \{p\}$, then $\{F_n : n \in \mathbb{N}\}$ is a local base at p for the topology of X.

(ii) There is a countable local base at p for the topology of X.

(iii) $\{p\}$ is a G_δ.

PROOF: (i) $\Rightarrow$ (ii): Let (U_n) be a sequence of open sets in X with $\cap_n U_n = \{p\}$. Set $V_0 = X$ and inductively let V_n be an open neighborhood of p with $\bar{V}_n \subseteq U_n \cap V_{n-1}$ (using regularity of X). Then $(\bar{V}_n)$ is a decreasing sequence of closed neighborhoods of p with intersection $\{p\}$, so they form a local base at p (XIII.11.1.13.). (ii) $\Rightarrow$ (iii) is trivial (since X is Hausdorff, the intersection of any local base at p is $\{p\}$), and (iii) $\Rightarrow$ (i) is a special case of XIII.11.1.12.. 👍

XIII.11.5. Automatic Continuity and Comparison of Topologies

Propositions XIII.11.1.6. and XIII.11.1.7. combine in a useful consequence:

XIII.11.5.1. PROPOSITION. Let X and Y be topological spaces, with X compact and Y Hausdorff. If $f : X \to Y$ is continuous, then f is a closed mapping.

PROOF: Let E be a closed subset of X. Then E is compact by XIII.11.1.6., so $f(E)$ is compact by XIII.11.1.7. and hence closed by XIII.11.1.6.. 👍

XIII.11.5.2. COROLLARY. Let X and Y be topological spaces, with X compact and Y Hausdorff. If $f : X \to Y$ is continuous and injective, then $f(X)$ is closed in Y and f is a homeomorphism from X onto $f(X)$. In particular, X must also be Hausdorff.

PROOF: A bijective continuous closed mapping is automatically a homeomorphism (i.e. the inverse map is continuous). 👍

XIII.11.5.3. Let X be a set. Any topology on X weaker than a compact topology is compact, and any topology stronger than a Hausdorff topology is Hausdorff. Thus if $\mathcal{T}$ is a compact Hausdorff topology on X, then $\mathcal{T}$ is delicately balanced: no strictly stronger topology is compact, and no strictly weaker topology is Hausdorff [consider the identity map and apply XIII.11.5.2.].

XIII.11.6. Products of Compact Spaces and Tikhonov's Theorem

Perhaps the most important theorem of general topology is Tikhonov's Theorem (see XIII.11.9.21. for some historical comments):

XIII.11.6.1. THEOREM. [TIKHONOV'S THEOREM] Let $\{(X_i, \mathcal{T}_i) : i \in I\}$ be a collection of topological spaces, $X = \prod_{i \in I} X_i$, and $\mathcal{T}$ the product topology on X. If each $(X_i, \mathcal{T}_i)$ is compact, then $(X, \mathcal{T})$ is compact.

This theorem is one of the two theorems of topology, along with the Baire Category Theorem, which are used most extensively in analysis.

Tikhonov's Theorem can, among other things, be interpreted as giving compelling evidence that the product topology is the "right" topology on the Cartesian product.

XIII.11.6.2. Tikhonov's Theorem depends on the Axiom of Choice; in fact, it is equivalent to the Axiom of Choice (cf. Exercise XIII.11.14.6.). Important and frequently-used special cases, however, need only weaker versions of AC or no choice at all. We will prove some important versions under weaker choice assumptions.

There are several known proofs of Tikhonov's Theorem; they are for the most part variations of the arguments we use in the various special cases. For a more complete discussion of these proofs and their history, see [Kel75], [Mun75], ...

Finite Products

The simplest case is that of finite products, where no version of AC is needed:

XIII.11.6.3. THEOREM. Let $(X_1, \mathcal{T}_1), \ldots, (X_n, \mathcal{T}_n)$ be a finite number of topological spaces, and $\mathcal{T}$ the product topology on $X = X_1 \times \cdots \times X_n$. If each $(X_k, \mathcal{T}_k)$ is compact, then $(X, \mathcal{T})$ is compact.

PROOF: The argument is a slight generalization of the one in (). We prove the theorem in the case $n = 2$; this case can be applied successively, or the arguments in the proof repeated finitely many times, to give the general result for finite products. We will also change notation slightly for clarity: we begin with two compact spaces $(X, \mathcal{T})$ and $(Y, \mathcal{S})$, and show that $X \times Y$ is compact in the product topology.

Suppose $(z_i)_{i \in I}$ is a net in $X \times Y$, with $z_i = (x_i, y_i)$. Then (x_i) is a net in X, and since X is compact there is a convergent subnet $(x_{\alpha(j)})_{j \in J}$, where $\alpha : J \to I$, with limit $x \in X$. The net $(y_{\alpha(j)})$ is a net in Y, and Y is compact, so has a convergent subnet $(y_{\alpha(\beta(k))})_{k \in K}$, where $\beta : K \to J$, with limit $y \in Y$. The subnet $(x_{\alpha(\beta(k))})_{k \in K}$ of $(x_{\alpha(j)})$ also converges to x, and hence the subnet $(z_{\alpha(\beta(k))})$ of (z_i) converges to $z = (x, y)$ by ().

XIII.11.6.4. This proof can be modified to show that a countable product of compact spaces is compact, but there is a logical complication and one additional step needed. The additional step is to construct a "diagonal" subnet after a sequence of subnets has been chosen; see Exercise () for details. The logical complication is that an infinite sequence of subnets must be chosen, each depending on the previous one; for this, the Axiom of Dependent Choice is needed. Using another argument (), one can get away with the weaker Countable Axiom of Choice for this result (in fact it is equivalent to Countable AC).

The General Tikhonov Theorem

Next we prove the full Tikhonov Theorem (XIII.11.6.1.). This proof is due to P. CHERNOFF [Che92]; cf. [Fol99,]. At one point we use an argument similar to the one in the proof of XIII.11.6.3.. This proof uses the full Axiom of Choice (in the form of Zorn's Lemma), as it must.

XIII.11.6.5. PROOF: Let $(x_\alpha)_{\alpha\in A}$ be a net in $X = \prod_{i\in I} X_i$. We need to show that (x_α) has a cluster point. If $J_1 \subseteq J_2 \subseteq I$, write π_{J_1} for the projection from $\prod_{i\in J_2} X_i$ onto $\prod_{i\in J_1} X_i$ (the notation avoids mention of J_2 to avoid unnecessary complication). Consider the set of all pairs (J, y), where J is a nonempty subset of I and $y \in \prod_{i\in J} X_i$ is a cluster point for the net $(\pi_J(x_\alpha))_{\alpha\in A}$. The collection $\mathcal{S}$ of such pairs is nonempty since there is such a y whenever J is a singleton by the assumption that each X_i is compact. Put a partial ordering on $\mathcal{S}$ by setting $(J_1, y_1) \leq (J_2, y_2)$ if $J_1 \subseteq J_2$ and $y_1 = \pi_{J_1}(y_2)$.

Suppose $\mathcal{C} = \{(J_k, y_k)\}$ is a chain in $\mathcal{S}$, and set $J_0 = \cup_k J_k$. For each $i \in J_0$, choose k with $i \in J_k$ and let $z_i = \pi_{\{i\}}(y_k)$. If $i \in J_{k_1} \cap J_{k_2}$, then either $J_{k_1} \subseteq J_{k_2}$ or vice versa, and if $J_{k_1} \subseteq J_{k_2}$ we have $\pi_{\{i\}}(y_1) = \pi_{\{i\}}(\pi_{J_1}(y_2)) = \pi_{\{i\}}(y_2)$, so z_i is well defined independent of the choice of k. Set $y = (z_i) \in \prod_{i\in J_0} X_i$. For each $i \in J_0$ we have that z_i is a cluster point of the net $(\pi_{\{i\}}(x_\alpha))_{\alpha\in A}$. If U is a neighborhood of y_0 in $\prod_{i\in J_0} X_i$, then there are indices $i_1, \ldots, i_n \in J_0$ and neighborhoods U_r of z_{i_r} in X_{i_r} such that

$$\prod_{r=1}^{n} U_r \quad \times \prod_{i\in J\setminus\{i_1,\ldots,i_n\}} X_i \quad \subseteq U \ .$$

Since $\mathcal{C}$ is totally ordered, there is a k such that $i_1, \ldots, i_n \in J_k$; for any α_0 there is an $\alpha \geq \alpha_0$ with

$$\pi_{J_k}(x_\alpha) \in \prod_{r=1}^{n} U_r \quad \times \prod_{i\in J_k\setminus\{i_1,\ldots,i_n\}} X_i$$

and hence $\pi_{J_0}(x_\alpha) \in U$. So we have that y_0 is a cluster point of the net $(\pi_{J_0}(x_\alpha))_{\alpha\in A}$ by (). Thus (J_0, y_0) is an upper bound for $\mathcal{C}$.

By Zorn's Lemma, $\mathcal{S}$ has a maximal element (J, y). We need to show that $J = I$, for then y will be the desired cluster point of (x_α). Suppose $J \neq I$, and let $i \in I\setminus J$, $J' = J\cup\{i\}$. There is a subnet $(\pi_J(x_{\alpha(\beta)}))_{\beta\in B}$ of $(\pi_J(x_\alpha))_{\alpha\in A}$ which converges to y by (). The net $\pi_{\{i\}}(x_{\alpha(\beta)}))_{\beta\in B}$ has a subnet $(\pi_{\{i\}}(x_{\alpha(\beta(\gamma))}))_{\gamma\in C}$ which converges to some $z \in X_i$ since X_i is compact. The subnet $\pi_J(x_{\alpha(\beta(\gamma))}))_{\gamma\in C}$ of $(\pi_J(x_{\alpha(\beta)}))_{\beta\in B}$ converges to y, so the subnet $(\pi_{J'}(x_{\alpha(\beta(\gamma))}))_{\gamma\in C}$ of $(\pi_{J'}(x_\alpha))_{\alpha\in A}$ converges to (y, z), and thus $(J', (y, z))$ is an element of $\mathcal{S}$ strictly larger than (J, y), contradicting maximality. Thus $J = I$ and the theorem is proved. 👍

Tikhonov's Theorem for Hausdorff Spaces

Tikhonov's Theorem for Hausdorff spaces does not require the full Axiom of Choice, only the Ultrafilter Property, which is equivalent to the Boolean Prime Ideal Theorem (II.6.3.11.).

The proof we give was developed by H. CARTAN for BOURBAKI. CARTAN'S proof had an additional step (using the full Axiom of Choice) and gave a proof of the full Tikhonov Theorem, but it simplifies nicely to give an efficient proof in the Hausdorff case. The proof is deceptively quick to write, but it depends on earlier results about convergence and projection of ultrafilters.

XIII.11.6.6. THEOREM. [TIKHONOV'S THEOREM FOR HAUSDORFF SPACES] Let $\{(X_i, \mathcal{T}_i) : i \in I\}$ be a collection of Hausdorff topological spaces, $X = \prod_{i\in I} X_i$, and $\mathcal{T}$ the product topology on X. If each $(X_i, \mathcal{T}_i)$ is compact, then $(X, \mathcal{T})$ is compact.

PROOF: By XIII.11.2.3. (which depends on BPI), it suffices to show that if $\mathcal{F}$ is an ultrafilter on X, then $\mathcal{F}$ converges. For each $i \in I$, the ultrafilter $\mathcal{F}_i = \pi_{\{i\}}(\mathcal{F})$ () converges to a unique point $x_i \in X_i$ since X_i is compact and Hausdorff. Then $\mathcal{F}$ converges to $x = (x_i) \in X$ by (). 👍

XIII.11.6.7. If the X_i are not Hausdorff, the same proof works except that the limit point x_i is not necessarily unique. Thus an application of the Axiom of Choice is necessary to choose an x_i for each i. (If I is countable, the Countable AC plus the BPI suffices; cf. XIII.11.6.4..)

XIII.11.6.8. Under a fairly drastic set-theoretic assumption, there is an explicit metrizable product of compact Hausdorff spaces (two-point discrete spaces) which is not compact (XIII.15.12.9.).

Tikhonov's Theorem for the Hilbert Cube

It turns out that it can be proved that the Hilbert cube $[0,1]^{\mathbb{N}}$ is compact without any form of choice. We give an argument here from [Wag93].

XIII.11.6.9. THEOREM. The Hilbert cube $[0,1]^{\mathbb{N}}$ is compact.

PROOF: Suppose $\mathcal{U}$ is an open cover of $[0,1]^{\mathbb{N}}$ with no finite subcover. We will obtain a contradiction by repeatedly subdividing coordinatewise. By passing to a refinement, we may assume that the sets in $\mathcal{U}$ are rectangular open sets of the form $U_1 \times \cdots \times U_n \times [0,1] \times [0,1] \times \cdots$, where the U_k are open sets in $[0,1]$ (this can be done without using any form of choice by simply replacing each $U \in \mathcal{U}$ with the collection of all rectangular open sets contained in U).

Divide the space into $[0,1/2] \times [0,1] \times [0,1] \times \cdots$ and $[1/2,1] \times [0,1] \times [0,1] \times \cdots$. At least one of these sets does not have a finite subcover of $\mathcal{U}$; choose the first one which does not, and call it F_1. Further subdivide into eight pieces, ordered lexicographically:

$$[0,1/4] \times [0,1/2] \times [0,1] \times \cdots, [0,1/4] \times [1/2,1] \times [0,1] \times \cdots, [1/4,1/2] \times [0,1/2] \times [0,1] \times \cdots, \text{ etc.}$$

Let F_2 be the first of these sets in this order for which there is no finite subcover of $\mathcal{U}$. Continue inductively to generate a decreasing sequence of closed sets (F_k) which are products of subintervals. For each n the n'th coordinate intervals $I_{k,n}$ of the F_k have an intersection consisting of a single point x_n, and the point $x = (x_n) \in [0,1]^{\mathbb{N}}$ is in the intersection of all the F_k (in fact, this intersection is a singleton). There is a $U \in \mathcal{U}$ with $x \in U$. Since U is a rectangular open set determined by finitely many coordinates, we have that $F_k \subseteq U$ for some sufficiently large k, contradicting the choice of F_k as a set with no finite subcover from $\mathcal{U}$. 👍

XIII.11.6.10. It is shown more generally in [**?**] by a somewhat similar argument that compactness of a product $\prod_{i \in I} X_i$ of compact spaces can be proved without any form of choice if the index set can be well ordered and there is a choice function for the family of all, or sufficiently many, closed sets in the X_i. (Of course, if the AC is assumed, this proof then gives the full Tikhonov Theorem.)

XIII.11.6.11. If, for each n, X_n is a closed subset of the Hilbert cube, then without any form of choice it follows that $X = \prod_{n=1}^{\infty} X_n$ also embeds as a closed subset of the Hilbert cube, and thus from XIII.11.6.9. it follows that X and each X_n is compact. But it does not seem to be possible to prove that a countable product of compact metrizable spaces is compact without some form of choice, even though such a space embeds in the Hilbert cube (by making some choices): a specific such embedding must be chosen for each n before one can conclude by the above argument that the product embeds, requiring the Countable AC. If one tries to prove directly that the product space is totally bounded with respect to the product metric () (proving

completeness is no problem), choices have to be made also using the Countable AC. Thus this form of the theorem seems to require Countable AC; if so, it is logically no easier than the version for countable products of arbitrary compact spaces (XIII.11.6.4.). It would also be in contrast with the corresponding results for arbitrary products: no more choice is needed to prove that arbitrary products of compact metrizable spaces (indeed, arbitrary products of compact Hausdorff spaces) are compact than to prove that arbitrary powers of $[0,1]$ are compact (each is equivalent to BPI; cf. ()).

XIII.11.7. Variations on Compactness

Countable compactness, sequential compactness, quasicompactness, bicompactness, σ-compactness, local compactness paracompactness, pseudocompactness, metacompactness, Lindelof spaces, precompactness (relative compactness), ...

XIII.11.7.1. DEFINITION. A topological space X is *countably compact* if every countable open cover of X has a finite subcover.

Some useful alternate characterizations of countably compact spaces are given by the next result:

XIII.11.7.2. THEOREM. Let X be a topological space. The following are equivalent:

(i) X is countably compact.

(ii) If (F_n) is a sequence of closed sets in X with the Finite Intersection Property, then $\cap_{n=1}^{\infty} F_n$ is nonempty.

(iii) Every sequence in X has a cluster point.

PROOF: (i) $\iff$ (ii) is proved by taking complements exactly as in the proof of XIII.11.1.10. with the word "countable" added in appropriate places.

(iii) $\Rightarrow$ (i): Suppose $\{U_1, U_2, \dots\}$ is a countable open cover of X with no finite subcover. For each n, let $x_n \in X \setminus \cup_{k=1}^{n} U_k$. If $x \in X$, then $x \in U_m$ for some m; but $x_n \notin U_m$ for $n \geq m$, so x is not a cluster point of (x_n). Thus (x_n) is a sequence in X with no cluster point.

(i) $\Rightarrow$ (iii): Suppose (x_n) is a sequence in X with no cluster point. For each n, let F_n be the closure of $\{x_k : k > n\}$, and let $U_n = F_n^c$. Then $\mathcal{U} = \{U_n : n \in \mathbb{N}\}$ is an open cover of X since any point of X has a neighborhood containing only finitely many of the x_n. Since the F_n are decreasing, the U_n are increasing. If there is a finite subcover of $\mathcal{U}$, then $U_n = X$ for some n. This is a contradiction since F_n is nonempty. Thus $\mathcal{U}$ has no finite subcover. 👍

XIII.11.7.3. COROLLARY. A sequentially compact topological space (XIII.11.2.2.) is countably compact. A first countable topological space is countably compact if and only if it is sequentially compact.

A countably compact space, even a compactum, which is not first countable need not be sequentially compact (XIII.11.9.16.(v)).

XIII.11.7.4. PROPOSITION. A countably compact Lindelöf space is compact. A countably compact metrizable space is compact.

PROOF: Let X be countably compact. If X is Lindelöf, then any open cover of X has a countable subcover, which has a finite subcover. If X is metrizable, fix a metric ρ on X. Any Cauchy sequence in X has a cluster point, hence converges by (); thus (X, ρ) is complete. If X is not totally bounded, then for some $\epsilon > 0$ a sequence (x_n) exists in X with $\rho(x_n, x_m) \geq \epsilon$ for $n \neq m$; this sequence cannot have a cluster point, a contradiction.

XIII.11.7.5. EXAMPLE. Let ω_1 be the first uncountable ordinal, with the order topology. Every sequence in ω_1 has an upper bound, hence is contained in a compact initial segment. Thus ω_1 is sequentially compact, hence countably compact (it is also first countable). But ω_1 is not compact: the "identity net" in ω_1 has no cluster point.

As a variation, the "long line" (XIII.20.5.5.) is sequentially compact but not compact.

σ-Compact Spaces

XIII.11.7.6. DEFINITION. A topological space X is *σ-compact* if X is a countable union of compact subsets of X. More generally, a subset E of X is σ-compact if E is a union of a countable number of compact subsets of X.

Some authors require a σ-compact space to be locally compact, but we will not.

XIII.11.7.7. EXAMPLES. (i) Any compact space is σ-compact.

(ii) $\mathbb{R}^n$ is σ-compact [the closed ball around 0 of radius k is compact for any k].

(iii) A manifold (XIII.20.1.7.) is σ-compact if and only if it has only countably many connected components, i.e. satisfies the equivalent conditions of XIII.20.1.4..

(iv) Any countable topological space is σ-compact. A space with the discrete topology is σ-compact if and only if it is countable.

The next result is an immediate corollary of XIII.11.1.7.:

XIII.11.7.8. PROPOSITION. Let X and Y be topological spaces, and $f : X \to Y$ a continuous function. If E is a σ-compact subset of X, then $f(E)$ is a σ-compact subset of Y.

Recall () that an F_σ in a topological space is a countable union of closed subsets.

XIII.11.7.9. PROPOSITION. Let X and Y be topological spaces, with X σ-compact and Y Hausdorff. Then

(i) Every F_σ in X is σ-compact.

(ii) Every σ-compact subset of Y is an F_σ.

PROOF: (i): Suppose $X = \cup_n K_n$, where K_n is compact. If F is a closed subset of X, then $F = \cup_n (F \cap K_n)$, and by XIII.11.1.6. $F \cap K_n$ is compact. Thus every closed set in X is σ-compact, and hence so is every F_σ. (ii): If $E = \cup_n E_n \subset Y$ with E_n compact, then E_n is closed by XIII.11.1.6., so E is an F_σ. 👍

XIII.11.7.10. COROLLARY. Let X be a σ-compact Hausdorff space. Then a subset of X is σ-compact if and only if it is an F_σ.

It follows that the set of irrational numbers in $\mathbb{R}$ is not σ-compact since it is not an F_σ ().

Combining these results, we get an interesting consequence. Although a continuous image of a closed set is not closed in general, we have:

XIII.11.7.11. PROPOSITION. Let X and Y be topological spaces, with X σ-compact and Y Hausdorff, and let $f : X \to Y$ be a continuous function. If E is an F_σ in X, then $f(E)$ is an F_σ in Y.

PROOF: E is σ-compact by XIII.11.7.9.(i), so $f(E)$ is also σ-compact by XIII.11.7.8., and hence by XIII.11.7.9.(ii) we conclude that $f(E)$ is an F_σ in Y. 👍

In particular, under the hypotheses, the image of any closed subset of X is an F_σ in Y. The hypotheses hold if X and Y are Euclidean spaces, or manifolds with X separable. The corresponding statements for G_δ's are false in general (XIII.11.14.18.). The result fails in general if X is not σ-compact: the image of a closed set need not even be a Borel set, even if X is locally compact and Y is compact [for a simple example, let $X = Y = [0, 1]$ with Y having the ordinary topology and X the discrete topology, and $f = id$]. Even if X and Y are Polish spaces, the image of a closed set in X need not be a Borel set in Y ().

XIII.11.7.12. PROPOSITION. A σ-compact topological space is Lindelöf. Thus a σ-compact regular space is normal.

PROOF: Let $X = \cup_n K_n$, with K_n compact. If $\mathcal{U}$ is an open cover of X, for each n let $\mathcal{V}_n$ be a finite subset of $\mathcal{U}$ which covers K_n. Then $\mathcal{V} = \cup_n \mathcal{V}_n$ is a countable subset of $\mathcal{U}$ which covers X. (This argument uses the Countable AC.) The last statement follows from (). 👍

XIII.11.8. Locally Compact Spaces

Many important spaces such as Euclidean spaces are not compact, but are "locally compact":

XIII.11.8.1. DEFINITION. A Hausdorff space X is *locally compact* if every point of X has a neighborhood (not necessarily open) which is compact.

We will deal with the non-Hausdorff case in XIII.11.8.20..

XIII.11.8.2. EXAMPLES. (i) Any compactum is locally compact. (The whole space is a compact neighborhood of any point.)

(ii) $\mathbb{R}^n$ is locally compact since every closed ball is compact. More generally, any open set in $\mathbb{R}^n$ is locally compact in the relative topology for the same reason.

(iii) Any space with the discrete topology is locally compact.

(iv) Any closed subspace of a locally compact Hausdorff space is locally compact in the relative topology: if Y is a closed subset of a locally compact Hausdorff space X, $p \in Y$, and K is a compact neighborhood of p in X, then $K \cap Y$ is a compact neighborhood of p in Y in the relative topology. In particular, any closed subset of $\mathbb{R}^n$ is locally compact in the relative topology.

(v) Any open subset of a compactum is locally compact (in the relative topology): if U is an open subset of a compactum X, and $p \in U$, then by XIII.11.4.2. p and $X \setminus U$ have disjoint neighborhoods V and W in X. $\bar{V}$ is a compact neighborhood of p in X which is contained in U, i.e. a compact neighborhood of p in U.

Actually, any open subset of a locally compact Hausdorff space is locally compact (XIII.11.8.8.); in fact, any locally compact Hausdorff space is homeomorphic to an open set in a compactum (XIII.11.8.4.).

(vi) Spaces such as $\mathbb{Q}$, $\mathcal{N} = \mathbb{N}^{\mathbb{N}}$ (), and $\mathbb{R}^{\mathbb{N}}$ () are not locally compact; in fact, no point in any of these spaces has a compact neighborhood.

The One-Point Compactification

XIII.11.8.3. If X is a locally compact Hausdorff space, a single point can be added to X to make a compactum, called the *one-point compactification* of X and denoted $X^\dagger$. Specifically, take a point not in X which we call ∞, and let $X^\dagger = X \cup \{\infty\}$. Put a topology on $X^\dagger$ by taking all open subsets of X to be open in $X^\dagger$, along with all subsets consisting of ∞ along with the complement of a compact subset of X (including the complement of $\emptyset$, giving the whole space $X^\dagger$). It is simple and routine to verify that this is a topology (Exercise XIII.11.14.14.), and that X is an open subset of $X^\dagger$. See XIII.11.14.14. for a generalization of the one-point compactification construction.

XIII.11.8.4. PROPOSITION. $X^\dagger$ is a compactum with this topology.

PROOF: It is easy to see that $X^\dagger$ is compact. If $\{U_i\}$ is an open cover, there is a set U_{i_0} containing ∞. The complement of U_{i_0} is a compact subset of X, so the open cover of $X^\dagger \setminus U_{i_0}$ by the remaining U_i has a finite subcover $\{U_{i_1}, \ldots, U_{i_n}\}$. Then $\{U_{i_0}, U_{i_1}, \ldots, U_{i_n}\}$ is a finite subcover of $X^\dagger$.

To see that $X^\dagger$ is Hausdorff, it suffices to show that if $p \in X$, then p and ∞ have disjoint neighborhoods. Since X is locally compact, there is an open neighborhood U of p in X whose closure in X is compact (e.g. the interior of a compact neighborhood). Thus U and $\{\infty\} \cup (X \setminus \bar{U})$ are disjoint neighborhoods of p and ∞. 👍

The one-point compactification of X is essentially unique. Thus it can often be realized or recognized in a different form:

XIII.11.8.5. PROPOSITION. Let X be a locally compact Hausdorff space, Y a compactum, and $\phi : X \to Y$ a homeomorphism from X onto a (necessarily open) subspace $\phi(X)$ of Y with $Y \setminus \phi(X)$ a single point q. Then ϕ extends to a homeomorphism $\bar{\phi}$ from $X^\dagger$ onto Y with $\bar{\phi}(\infty) = q$.

PROOF: The map $\bar{\phi}$ is obviously completely determined; it is only necessary to show it is a homeomorphism. If V is an open set in Y, and V does not contain q, then since ϕ is continuous, $\bar{\phi}^{-1}(V) = \phi^{-1}(V)$ is open in X, hence in $X^\dagger$. And if $q \in V$, then V^c is a compact subset of $\phi(X)$, hence $\phi^{-1}(V^c)$ is compact since ϕ^{-1} is continuous, so $\bar{\phi}^{-1}(V) = X^\dagger \setminus \phi^{-1}(V^c)$ is open in $X^\dagger$. Thus $\bar{\phi}$ is continuous. A continuous bijection from a compact space to a Hausdorff space is a homeomorphism by (), so $\bar{\phi}$ is a homeomorphism. 👍

XIII.11.8.6. EXAMPLES. (i) Let Y be any compactum, and $q \in Y$. Then $X = Y \setminus \{q\}$ is a locally compact Hausdorff space with $X^\dagger$ homeomorphic to Y. By the one-point compactification construction, every locally compact Hausdorff space arises this way.

(ii) An n-sphere with a point removed is homeomorphic to $\mathbb{R}^n$, e.g. by stereographic projection (). Thus $(\mathbb{R}^n)^\dagger$ is homeomorphic to S^n. In particular, $\mathbb{R}^\dagger$ is homeomorphic to a circle.

XIII.11.8.7. If X is already compact, then $\emptyset$ is the complement of a compact subset of X, so $\{\infty\}$ is open in $X^\dagger$, i.e. ∞ is an isolated point in $X^\dagger$, and $X^\dagger$ is the separated union of X and a one-point discrete space. Conversely, if ∞ is isolated in $X^\dagger$, then X is a closed subset of $X^\dagger$, hence compact. Thus X is dense in $X^\dagger$ if and only if X is noncompact.

XIII.11.8.8. It follows that any open subset of a locally compact Hausdorff space is locally compact: if U is an open subset of a locally compact Hausdorff space X, then U is an open subset of the compactum $X^\dagger$ and thus is locally compact in the relative topology by XIII.11.8.2.(v).

We also get the following corollaries of the one-point compactification construction:

XIII.11.8.9. PROPOSITION. A locally compact Hausdorff space is completely regular ().

PROOF: $X^\dagger$ is normal by XIII.11.4.3., hence completely regular, and complete regularity passes to subspaces. 👍

However, a locally compact Hausdorff space need not be normal (XIII.11.14.7., XIII.20.5.8.).

XIII.11.8.10. PROPOSITION. Let X be a locally compact Hausdorff space. If $p \in X$ and V is a neighborhood of p, then there is an open neighborhood U of p such that $\bar{U}$ is compact and contained in V.

PROOF: There is a compact neighborhood K of p in X; the interior W of K is an open neighborhood of p whose closure is compact. Replacing V by $V \cap W$, we may assume $\bar{V}$ is compact. Since X is regular, $\{p\}$ and V^c have disjoint neighborhoods U and Z. Then $\bar{U} \subseteq V$, and $\bar{U}$ is compact since it is a closed subset of $\bar{V}$. 👍

XIII.11.8.11. PROPOSITION. Let X be a locally compact Hausdorff space. The following are equivalent:

(i) The point ∞ has a countable local base in $X^\dagger$.

(ii) $\{\infty\}$ is a G_δ in $X^\dagger$.

(iii) X is σ-compact.

(iv) X is Lindelöf.

PROOF: It is obvious from the construction of $X^\dagger$ that (ii) $\iff$ (iii), and it follows from XIII.11.4.6. that (i) $\iff$ (ii). (iii) $\Rightarrow$ (iv) by XIII.11.7.12.. For (iv) $\Rightarrow$ (iii), each $x \in X$ has an open neighborhood V_x with $\bar{V}_x$ compact. $\{V_x : x \in X\}$ has a countable subcover $\{V_{x_n} : n \in \mathbb{N}\}$, and $X = \cup_n \bar{V}_{x_n}$. 👍

XIII.11.8.12. PROPOSITION. Let X be a locally compact, σ-compact Hausdorff space. Then there is a sequence (U_n) of open sets in X with $\bar{U}_n$ compact and contained in U_{n+1} for each n, and $X = \cup_{n=1}^\infty U_n$.

PROOF: Let (K_n) be a sequence of compact subsets of X with $X = \cup K_n$. Let U_1 and V_1 be disjoint neighborhoods of K_1 and $\{\infty\}$ in $X^\dagger$; then U_1 is open and the closure $\bar{U}_1$ of U_1 in $X^\dagger$ is compact and contained in X. Inductively take U_{n+1} and V_{n+1} to be disjoint open neighborhoods of $\bar{U}_n \cup K_{n+1}$ and $\{\infty\}$ in $X^\dagger$. 👍

XIII.11.8.13. THEOREM. Let X be a locally compact Hausdorff space. The following are equivalent:

(i) X is metrizable and σ-compact.

(ii) There is a countable base for the topology of X consisting of compact sets.

(iii) X is second countable.

(iv) X is a Polish space ().

(v) $X^\dagger$ is metrizable.

There are various other ways to phrase these conditions which are equivalent by general considerations, e.g. "X is separable and metrizable" (). There are metrizable locally compact spaces which do not satisfy the conditions, e.g. an uncountable discrete space.

PROOF: (i) $\Rightarrow$ (v): Let (U_n) be an increasing sequence of open sets in X as in XIII.11.8.12.. Then $\bar{U}_n$ is compact and metrizable for each n, hence second countable (). Thus U_n is second countable. There is also a countable local base at ∞ in $X^\dagger$ by XIII.11.8.11.. Putting together a countable base for the topology of U_n for each n and a countable local base at ∞, we obtain a countable base for the topology of $X^\dagger$; hence $X^\dagger$ is a second countable compactum, so it is metrizable ().

(v) $\Rightarrow$ (iv): A compact metrizable space is a Polish space (); so $X^\dagger$ is Polish. Thus the open subset X of $X^\dagger$ is also a Polish space ().

(iv) ⇒ (iii): A separable metric space is second countable ().

(iii) ⇒ (ii): It suffices to show that if $\{U_n\}$ is a countable base, then the collection

$$\{\bar{U}_n : n \in \mathbb{N},\ \bar{U}_n \text{ compact}\}$$

is also a base. Let $p \in X$ and V a neighborhood of p in X. There is an open set U containing p such that $\bar{U}$ is compact and contained in V (XIII.11.8.10.). Take U_n so that $p \in U_n \subseteq U$. Then $\bar{U}_n \subseteq V$, and $\bar{U}_n$ is compact since it is a closed subset of $\bar{U}$.

(ii) ⇒ (i): X is regular (XIII.11.8.9.), hence metrizable by the Urysohn Metrization Theorem (). (ii) trivially implies that X is σ-compact.

XIII.11.8.14. There is no universally accepted standard notation for the one-point compactification of a space X. I have chosen a notation $X^\dagger$ that is not widely used, and an explanation is in order. The most common notations are X^*, X^+, X_∞; other notations sometimes used are X_*, X_+, X^∞, $\tilde{X}$, and perhaps others.

There are two reasons I have chosen not to use X^*. One is that topologists also often use this notation to mean something else, e.g. the corona space $\beta X \setminus X$ (). The other reason comes from my background in Operator Algebras, where * is normally used to mean something else (); in fact, it is used to mean too many different things in Operator Algebras!) In addition, there is a natural relationship between adding a point at infinity in topology and adding a unit in algebra: a commutative C*-algebra is an algebra of the form $C_0(X)$, the continuous functions vanishing at infinity on a locally compact Hausdorff space X (), and the unitization of this algebra is precisely the continuous functions on the one-point compactification of X. Thus it makes sense to use the same notation for one-point compactification as for adding a unit to a C*-algebra. A common notation in algebra is to write A^+ for the unitization of the algebra A, and thus X^+ would be logical notation for the one-point compactification, giving that $C_0(X)^+ = C(X^+)$. However, a C*-algebra A also has a positive cone which is usually denoted A_+ or sometimes A^+, and the two notations are too easily confused. So in [Bla06] I settled on the notation $A^\dagger$ for the unitization of a C*-algebra, and thus $X^\dagger$ for the one-point compactification of a space (so that $C_0(X)^\dagger = C(X^\dagger)$), as a variation of A^+ which is easily distinguishable but similar enough to be suggestive, and I have kept that notation here. Some readers may not be happy with my choice, but it is my book ...

Urysohn's Lemma for Locally Compact Spaces

Although a locally compact space need not be normal (XIII.20.5.8.), there is a version of Urysohn's Lemma. We first make an observation and a definition:

XIII.11.8.15. Proposition. Let X be a locally compact Hausdorff space, and A a compact subset of X, and V a neighborhood of A in X. Then there is an open neighborhood U of A contained in V with $\bar{U}$ compact.

Proof: This is a routine compactness argument from XIII.11.8.10.. For each $p \in U$, let U_p be an open neighborhood of p contained in V with $\bar{U}_p$ compact. The open cover $\{U_p : p \in A\}$ of A has a finite subcover $\{U_{p_1}, \ldots, U_{p_n}\}$. Set $U = U_{p_1} \cup \cdots \cup U_{p_n}$.

XIII.11.8.16. DEFINITION. Let X be a topological space. A continuous function $f : X \to \mathbb{C}$ has *compact support* if the closure of the support $S_f = \{x \in X : f(x) \neq 0\}$ of f is compact. Denote by $C_c(X)$ the set of all continuous functions of compact support from X to $\mathbb{C}$.

XIII.11.8.17. It is easily verified that $C_c(X)$ is a subalgebra of $C(X)$. But in general it need not contain any nonzero function. However, if X is locally compact, $C_c(X)$ is large enough to separate points and closed sets:

XIII.11.8.18. THEOREM. [URYSOHN'S LEMMA, LOCALLY COMPACT CASE] Let X be a locally compact Hausdorff space. If A and B are disjoint closed subsets of X with A compact, then there is a continuous function $f : X \to [0, 1]$ of compact support with $f(x) = 1$ for all $x \in A$ and $f(x) = 0$ for all $x \in B$.

PROOF: The set $V = B^c$ is a neighborhood of A. Let U be an open neighborhood of A in X contained in V with $\bar{U}$ compact. Regarding X, and hence A, B, U, as subsets of $X^\dagger$, let $C = X^\dagger \setminus U$. Then A and C are disjoint closed subsets of the compact, hence normal, space $X^\dagger$. If $g : X \to [0, 1]$ with $g(x) = 1$ for all $x \in A$ and $g(x) = 0$ for all $x \in C$, then the restriction of g to X is the desired f. 👍

XIII.11.8.19. In particular, if X is a locally compact Hausdorff space, B a closed subset of X, and $p \in X$, $p \notin B$, then (applying the theorem with $A = \{p\}$) there is a continuous function $f : X \to [0, 1]$ of compact support with $f(p) = 1$ and $f(x) = 0$ for all $x \in B$.

The Non-Hausdorff Case

XIII.11.8.20. There is no natural definition of local compactness for non-Hausdorff spaces. Each of the following conditions, which are equivalent for Hausdorff spaces, is a reasonable potential definition for a topological space X:

(i) Every point of X has a neighborhood (not necessarily open) which is compact.

(ii) For any $p \in X$ and any open neighborhood V of X, there is a compact neighborhood U (not necessarily open) of p contained in V.

(iii) Every point of X has an open neighborhood whose closure is compact.

(iii′) For any $p \in X$ and any open neighborhood V of X, there is an open neighborhood U of X contained in V whose closure is compact.

(iv) For any $p \in X$ and any open neighborhood V of X, there is an open neighborhood U of X whose closure is compact and contained in V.

In general, (iv) $\Rightarrow$ (iii) $\Longleftrightarrow$ (iii′) $\Rightarrow$ (i) and (iv) $\Rightarrow$ (ii) $\Rightarrow$ (i). In Hausdorff spaces, (i) $\Rightarrow$ (iv) (XIII.11.8.10.), so all the conditions are equivalent. None of the other implications hold in general [SS95, Examples 9, 18, 35].

XIII.11.9. Compactifications

It is often very convenient to embed a given topological space into a compactum. The first instance of this is to add $\pm\infty$ to the real line to obtain the extended real line, which is topologically and order-theoretically the same as the compact space $[0,1]$. There are some systematic ways to compactify a topological space, which we discuss in this section.

XIII.11.9.1. DEFINITION. Let X be a Hausdorff topological space. A *compactification* of X is a compactum Y and a homeomorphism ϕ of X onto a dense subset of Y.

Some authors consider compactifications where X and/or Y is not Hausdorff (cf. XIII.11.9.20.), but we will only consider the Hausdorff case. By (), a (Hausdorff) space has a compactification if and only if it is completely regular.

XIII.11.9.2. In some references, a pair (Y,ϕ) is called a compactification of X if Y is a compactum and ϕ is simply an injective continuous function from X onto a dense subset of Y. Instances of this are the standard maps of $\mathbb{R}$ to a figure 8, a figure 6, a torus via an irrational angle wraparound (), or a solenoid (). Another instance would be a map from $\mathbb{N}$ with the discrete topology onto a dense subset of $[0,1]$. An important and natural example is the Bohr compactification of $\mathbb{R}$ (XXIII.5.2.2.). We will not consider these to be compactifications, but they are what we call compact representations (XIII.11.9.9.).

XIII.11.9.3. The density of X in a compactification is a natural condition, but is not restrictive: if ϕ is a homeomorphism of X with a subspace of a compactum Z, then the closure of $\phi(X)$ in Z is a compactification of X.

XIII.11.9.4. There are several special instances of compactifications already considered:

(i) If (X,ρ) is a totally bounded metric space, then the completion of (X,ρ) is a compactification of X.

(ii) If X is a totally ordered space with the order topology, then the order-completion of X with endpoints added if necessary, with the order topology, is a compactification of X ().

(iii) If X is a locally compact, noncompact Hausdorff space, then the one-point compactification $X^\dagger$ (XIII.11.8.3.) is a compactification of X. (If X is already compact, then $X^\dagger$ is not a compactification of X since X is not dense in $X^\dagger$; in fact, the only compactification of a compact space X is X itself. The fact that the one-point compactification is not a compactification in this case is an unfortunate inconsistency in standard terminology.)

(iv) There are more complicated compactifications of $\mathbb{R}$ than the extended real line and the one-point compactification (circle). For example, there is a homeomorphism of $\mathbb{R}$ onto the graph of $y=\sin\left(\frac{1}{x}\right)$ for $0<x<\frac{1}{\pi}$, so the topologist's sine curve (XIII.19.5.5.) is a compactification of $\mathbb{R}$.

(v) If X is completely regular, and ϕ is a homeomorphism of X onto a subspace of a cube (XIII.7.8.4.), then the closure of $\phi(X)$ is a compactification of X.

The compactification of (v) is actually universal: every compactification can be done (up to homeomorphism) in this way. For if Y is any compactification of X, then Y can be embedded in a cube as in XIII.7.8.4.. The construction of XIII.11.9.10. using the restrictions to X of the coordinate functions of this embedding yields Y.

XIII.11.9.5. Since a compact metrizable space is second countable, a space with a metrizable compactification is second countable. Conversely, by the construction of XIII.7.8.4. a second countable completely regular space can be embedded in the Hilbert cube, which is metrizable, and hence has a metrizable compactification.

XIII.11.9.6. Although the homeomorphism ϕ is often understood, it is logically cleaner to regard it as a part of the compactification. Two compactifications (Y, ϕ) and (Z, ψ) of a space X are *equivalent* if there is a homeomorphism $\omega : Y \to Z$ with $\psi = \omega \circ \phi$. If there is just a continuous function $\omega : Y \to Z$ with $\psi = \omega \circ \phi$, we say (Y, ϕ) *dominates* (Z, ψ). Such an ω is unique since it is completely determined on the dense subset $\phi(X)$ as $\psi \circ \phi^{-1}$, and it is necessarily surjective since $\omega(Y)$ is a compact subset of Z containing the dense subset $\psi(X)$.

Dominance gives a partial ordering on the equivalence classes of compactifications of X, setting $(Y, \phi) \geq (Z, \psi)$ if (Y, ϕ) dominates (Z, ψ). If X is locally compact and noncompact, then the one-point compactification of X is the smallest compactification of X under this ordering (); conversely, it is not hard to show (cf. ()) that if X has a minimal compactification, it must be a one-point compactification and X must thus be locally compact. There is always a largest compactification of X if X is completely regular, the Stone-Čech compactification βX ().

XIII.11.9.7. More generally, if X_1 and X_2 are (completely regular) spaces with compactifications (Y, ϕ) and (Z, ψ) respectively, and $f : X_1 \to X_2$ is a continuous function, an *extension* of f to (Y, Z) is a continuous function $\tilde{f} : Y \to Z$ with $\tilde{f} \circ \phi = \psi \circ f$. Such an extension, if it exists, is unique.

XIII.11.9.8. A compactification (Y, ϕ) of X naturally defines a uniform structure () on X by restricting the unique uniform structure on Y (). (However, not every uniform structure on X arises in this manner in general.) A continuous function $f : X_1 \to X_2$ extends to compactifications (Y, ϕ) and (Z, ψ) in the sense of XIII.11.9.7. if and only f is uniformly continuous for these uniform structures.

XIII.11.9.9. More generally, we define a *compact representation* of a (completely regular) space X to be a pair (Y, ϕ), where Y is a compactum and ϕ is a continuous function from X onto a dense subspace of Y. Thus a compactification is a special kind of compact representation, as are the "compactifications" of XIII.11.9.2.. Another important example of a compact representation is a bounded continuous function from X to $\mathbb{R}$, e.g. a continuous function from X to $[0, 1]$. Equivalence, domination, and extension can be defined in this more general context. Any compact representation which dominates a compactification is itself a compactification.

The Stone-Čech Compactification

There is a largest (in the sense of XIII.11.9.6.) compactification of any completely regular space X, called the *Stone-Čech compactification* of X and denoted βX. This compactification is attributed to M. Stone and E. Čech (independently), but see XIII.11.9.21..

The Stone-Čech compactification of a space X can be constructed in a number of ways (Stone and Čech used rather different constructions). One construction is via Boolean algebras and/or ultrafilters (cf. XIII.11.9.14.). A very slick construction is as the spectrum of the C*-algebra of bounded complex-valued continuous functions on X (XX.16.2.21.), but this approach is probably not the best for beginning students or topologists. We will give two versions of a purely topological construction.

The first version is based on the next result, which is proved by a pullback construction:

XIII.11.9.10. Proposition. Let X be a completely regular space, and $\{(Y_i, \phi_i) : i \in I\}$ a set of compact representations (XIII.11.9.9.) of X. Then there is a compact representation (Y, ϕ), unique up to equivalence, which is the least upper bound of the (Y_i, ϕ_i).

Proof: It is obvious that the least upper bound, if it exists, is unique up to equivalence. We show existence. Let $Z = \prod_i Y_i$. Then Z is compact by Tikhonov's Theorem (the Hausdorff version XIII.11.6.6. suffices). Define a map $\phi : X \to Z$ by $(\phi(x))_i = \phi_i(x)$ for $x \in X$. Then ϕ is continuous by the universal property of the product topology (). Let Y be the closure of $\phi(X)$ in Z. Then (Y, ϕ) obviously dominates (Y_i, ϕ_i) via coordinate projections. If (Y', ψ) dominates each (Y_i, ϕ_i) via a map $\omega_i : Y' \to Y_i$, define $\omega : Y' \to Y$ by $(\omega(t))_i = \omega_i(t)$ for $t \in Y'$. Then (Y', ψ) dominates (Y, ϕ) via ω. 👍

XIII.11.9.11. One then gets an apparent proof of the existence of a maximal compactification by simply taking the least upper bound of the "set" of all compact representations of X. However, there is a set-theoretic problem: this "set" is not a set. The problem can be circumvented by noting that if X has cardinality κ, then the cardinality of any compact representation is at most 2^{2^κ} (), so we could restrict to only compact representations whose underlying set is a subset of a fixed set of cardinality 2^{2^κ}, and these compact representations form a set. One can also restrict to one compact representation in each equivalence class in the construction. (This appears to require the AC (), but the AC can be avoided by an argument working through XIII.11.9.12.; anyway, use of the AC is not much of a problem here since some form of choice in the guise of Tikhonov's Theorem is necessary to do almost anything with compactifications.)

Call the maximal compact representation βX; it is a compactification since X has at least one compactification, which is dominated by βX. It is customary to identify X with its image in βX.

XIII.11.9.12. A more common and concrete construction of βX is to let $\mathcal{C}$ be the set of all continuous functions from X to $[0, 1]$, and embed X into $[0, 1]^{\mathcal{C}}$ by taking $\phi(x)_f = f(x)$ for $f \in \mathcal{C}$ and taking Y to be the closure of $\phi(X)$. Since each such f gives a compact representation of X, each of these is dominated by the βX constructed previously, hence (Y, ϕ) is dominated by βX. On the other hand, since βX is compact, it embeds into a cube (XIII.7.8.4., XIII.11.4.3.), and the coordinate functions of this embedding, restricted to X, are in $\mathcal{C}$; it follows that (Y, ϕ) dominates βX, and therefore (Y, ϕ) is equivalent to βX.

We summarize in a theorem:

XIII.11.9.13. Theorem. [Stone-Čech Compactification] Let X be a completely regular space. There is a compactum βX containing X as a dense subspace, unique up to homeomorphism which is the identity on X, with the property that if f is a continuous function from X to any compactum Y, then f extends uniquely to a continuous function $\tilde{f} : \beta X \to Y$.

XIII.11.9.14. We describe the ultrafilter version of the construction of βX only for X a set with the discrete topology. For the general version see [GJ76]. If X is a discrete space, βX can be identified with the set σX of all ultrafilters on X, with the hull-kernel topology (): the closure of a set $\mathcal{S}$ of ultrafilters is the set of all ultrafilters containing the filter which is the intersection of the ultrafilters in $\mathcal{S}$. It is easy to show

that this space is a compactum (), and that the set of fixed ultrafilters, which can be naturally identified with X, is dense, so σX is a compactification of X. To show that σX dominates βX,

XIII.11.9.15. If X is locally compact, then the Stone-Čech compactification dominates the one-point compactification $X^\dagger$. Since X is open in $X^\dagger$, its preimage in βX, the image of X in βX, is open. Conversely, if X is open in βX, then X is locally compact by XIII.11.8.2.(v).

The set $\beta X \setminus X$ is called the *corona* or *remainder* of X, especially if X is locally compact; in this case $\beta X \setminus X$ is compact.

XIII.11.9.16. The Stone-Čech compactification of a space is rather huge, and often has pathological properties. We list without proof some key properties of some standard spaces:

(i) If X is second countable and noncompact, then $card(\beta X) = 2^{2^{\aleph_0}}$. In particular, $card(\beta\mathbb{N}) = card(\beta\mathbb{Q}) = card(\beta\mathbb{R}) = 2^{2^{\aleph_0}}$. If X is any space which is locally compact and σ-compact, but not compact, then $card(\beta X \setminus X) \geq 2^{2^{\aleph_0}}$.

(ii) Every infinite closed subset of $\beta\mathbb{N} \setminus \mathbb{N}$ contains a copy of $\beta\mathbb{N}$, and in particular has cardinality $2^{2^{\aleph_0}}$.

(iii) $\beta\mathbb{N}$ is *extremally disconnected*: the closure of every open set is open.

(iv) $\beta\mathbb{R}$ is connected (as is βX for any connected X (XIII.13.1.10.)). $\beta\mathbb{R} \setminus \mathbb{R}$ has two connected components, one at $+\infty$ and the other at $-\infty$ (cf. XIII.13.3.6.). But $\beta\mathbb{R} \setminus \mathbb{R}$ is totally path disconnected, and not locally connected at any point. In fact, if X is locally compact and σ-compact, but not compact, then $\beta X \setminus X$ is an *F-space*: disjoint open F_σ's have disjoint closures. Any F-space is totally path disconnected. In $\beta\mathbb{N} \setminus \mathbb{N}$, $\beta\mathbb{R} \setminus \mathbb{R}$, and many other corona spaces, every nonempty G_δ has nonempty interior.

(v) Although any sequence in $\beta\mathbb{N}$ has a cluster point (in fact, $2^{2^{\aleph_0}}$ cluster points if it has infinite range), the only convergent sequences in $\beta\mathbb{N}$ are ones which are eventually constant. Thus any sequence of distinct points in $\beta\mathbb{N}$ has no convergent subsequence, and $\beta\mathbb{N}$ is compact but not sequentially compact. The same statements are true about $\beta\mathbb{R} \setminus \mathbb{R}$, so $\beta\mathbb{R} \setminus \mathbb{R}$ is also not sequentially compact.

For more detailed structure results about $\beta\mathbb{N}$ and related spaces, see [GJ76] and [vM84].

XIII.11.9.17. There are, however, locally compact noncompact spaces for which the one-point compactification and the Stone-Čech compactification coincide. The simplest example is the space ω_1 of ordinals less than the first uncountable ordinal, with the order topology, since every real-valued continuous function on ω_1 is eventually constant; cf. also XIII.20.5.6.. These spaces are *pseudocompact*: every real-valued continuous function is bounded.

XIII.11.9.18. Note that either Tikhonov's Theorem for Hausdorff spaces or the Ultrafilter Property is needed to prove Theorem XIII.11.9.13.. In fact, this theorem is equivalent to the BPI (II.6.3.11.).

XIII.11.9.19. THEOREM. The following theorems are logically equivalent:

(i) Tikhonov's Theorem for Hausdorff spaces.

(ii) Tikhonov's Theorem for copies of $[0,1]$.

(iii) Existence of Stone-Čech compactification for discrete spaces.

(iv) Existence of Stone-Čech compactification for arbitrary completely regular spaces.

PROOF: (i) ⇒ (ii) and (iv) ⇒ (iii) are trivial. (ii) ⇒ (iv) is the standard construction of the Stone-Čech compactification (XIII.11.9.12.). So we need only prove (iii) ⇒ (i).

Let $\{X_i : i \in I\}$ be an indexed collection of compacta, and let $X = \prod_i X_i$ with the product topology. Let Y be the set X with the discrete topology, and βY its Stone-Čech compactification. For each i, the coordinate map $\pi_i : X \to X_i$, regarded as a map from Y to X_i, extends uniquely to a continuous function f_i from βY to X_i, by the universal property of βY. The f_i define a continuous function from βY to X (), which is surjective since its restriction to Y is the identity map from Y to X. Thus X is a continuous image of the compact space βY, hence compact by ().

Note that (ii) ⇒ (i) is a simple consequence of the fact that every compact Hausdorff space can be embedded in a product of intervals, but we did not need to use this argument. 👍

XIII.11.9.20. There is a compactification of any T_1 space X called the *Wallman compactification*, which is similar to the Stone-Čech compactification but not Hausdorff if X is not normal (it coincides with the Stone-Čech compactification if X is normal). The Wallman compactification shows that every T_1 space can be embedded as a dense subset of a compact T_1 space (for a simpler proof of this, see XIII.11.14.14.).

XIII.11.9.21. According to [Dud02], TIKHONOV was the first to prove existence of the Stone-Čech compactification (in 1930), and ČECH elaborated on TIKHONOV's work; STONE's work appears to have been done independently, but later than TIKHONOV (as GELFAND once observed, TIKHONOV obtained his results "previously but independently.")

On the other hand, according to [Dud02], the general form (XIII.11.6.1.) of Tikhonov's Theorem was first proved by ČECH in 1936; TIKHONOV only proved that an arbitrary product of copies of $[0,1]$ is compact.

XIII.11.10. Paracompactness and Partitions of Unity

For some purposes, it is enough to be able to make an open cover of a space "locally finite" instead of actually (globally) finite, and it suffices to allow making a given open cover "finer" instead of just working with an actual subcover. It turns out that this is possible in the great majority of topological spaces encountered "naturally"; such spaces are called *paracompact*.

Paracompact spaces were only defined and named in 1944 by J. DIEUDONNÉ.

XIII.11.10.1. DEFINITION. Let $\mathcal{U}$ be an open cover of a topological space $(X, \mathcal{T})$. $\mathcal{U}$ is *point finite* if every $x \in X$ is contained in only finitely many $U \in \mathcal{U}$. $\mathcal{U}$ is *locally finite* if each $x \in X$ has an open neighborhood W such that $W \cap U \neq \emptyset$ for only finitely many $U \in \mathcal{U}$.

A locally finite cover is obviously point finite; the converse is false (Exercise ().) A finite cover is obviously locally finite.

XIII.11.10.2. DEFINITION. Let $\mathcal{U}$ and $\mathcal{V}$ be open covers of a topological space $(X, \mathcal{T})$. $\mathcal{V}$ is a *refinement* of $\mathcal{U}$ if every $V \in \mathcal{V}$ is contained in some $U \in \mathcal{U}$.

Note that $\mathcal{U}$ itself, or any subcover of $\mathcal{U}$, is a refinement of $\mathcal{U}$. If X is a metric space, $\epsilon, \delta > 0$, and $\mathcal{U}$ and $\mathcal{V}$ are the collections of all balls of radius ϵ and δ respectively, then $\mathcal{V}$ is a refinement of $\mathcal{U}$ if $\delta \leq \epsilon$. If $\mathcal{V}$ is a refinement of $\mathcal{U}$ and $\mathcal{W}$ is a refinement of $\mathcal{V}$, then $\mathcal{W}$ is a refinement of $\mathcal{U}$.

It can happen that two distinct open covers are each refinements of the other. For example, if $\mathcal{U}$ is the set of (bounded) open intervals in $\mathcal{R}$ of rational length, and $\mathcal{V}$ the set of open intervals of irrational length, then $\mathcal{U}$ and $\mathcal{V}$ are each refinements of the other.

XIII.11.10.3. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. Then $(X, \mathcal{T})$ is *paracompact* if it is Hausdorff and every open cover of $(X, \mathcal{T})$ has a locally finite refinement.

Note that a paracompact space is by definition Hausdorff. It is clear that any compactum is paracompact. But many noncompact spaces are also paracompact, as the next theorem shows (see also XIII.11.10.21.). This result was proved by A. H. STONE [Sto48] (cf. XIII.11.10.5.); the slick proof we give is due to M. E. RUDIN [Rud69].

XIII.11.10.4. THEOREM. Every metrizable space is paracompact.

PROOF: Let X be a metrizable space, and fix a metric ρ on X. Let $\mathcal{U} = \{U_i : i \in I\}$ be an open cover of X. We may well-order the index set I (using AC). Define, inductively on n, a collection of open sets V_{in} for $i \in I$ as follows. Suppose V_{im} has been defined for all i and for all $m < n$. Let V_{in} be the union of all open balls $B_{2^{-n}}(x)$ such that

(1) i is the smallest index such that $x \in U_i$.

(2) $x \notin V_{jm}$ for any j and any $m < n$.

(3) $B_{3 \cdot 2^{-n}}(x) \subseteq U_i$.

If $x \in X$, then there is a smallest i such that $x \in U_i$ and an n large enough that (3) holds. Then, by (2), $x \in V_{im}$ for some $m \leq n$. Thus $\mathcal{V} = \{V_{in} : i \in I, n \in \mathbb{N}\}$ is an open cover of X. Obviously $V_{in} \subseteq U_i$ for each i and n, so $\mathcal{V}$ refines $\mathcal{U}$.

To show that V is locally finite, fix $x \in X$, and let i be the smallest index for which $x \in V_{in}$ for some n. Fix the smallest n for which $x \in V_{in}$. Choose k so that $B_{2^{-k}}(x) \subseteq V_{in}$.

We first claim that if $m \geq n + k$, then $B_{2^{-n-k}}(x)$ intersects no V_{jm} for any $j \in I$. Fix j. Since $m > n$, every ball of radius 2^{-m} used to define V_{jm} has center y outside V_{in}, hence $\rho(x, y) \geq 2^{-k}$ by the choice of k. But $m \geq k + 1$ and $n + k \geq k + 1$, so $2^{-n-k} + 2^{-m} \leq 2^{-k}$, so $B_{2^{-n-k}}(x) \cap B_{2^{-m}}(y) = \emptyset$.

Finally, we claim that if $m < n + k$, then $B_{2^{-n-k}}(x)$ intersects V_{jm} for at most one j; this will complete the proof that $B_{2^{-n-k}}(x)$ intersects at most $n + k$ V_{jm}, and, since $x \in X$ is arbitrary, that $\mathcal{V}$ is locally finite. Suppose $p \in V_{j_1 m}$ and $q \in V_{j_2 m}$ for $j_1 < j_2$; we will show that $\rho(p, q) > 2^{-n-k+1}$, so both cannot be in $B_{2^{-n-k}}(x)$. There are points y and z such that $p \in B_{2^{-m}}(y) \subseteq V_{j_1 m}$ and $q \in B_{2^{-m}}(z) \subseteq V_{j_2 m}$. By (3), $B_{3 \cdot 2^{-m}}(y) \subseteq U_{j_1}$. But $z \notin U_{j_1}$ by (2), so $\rho(y, z) \geq 3 \cdot 2^{-m}$ and $\rho(p, q) > 2^{-m} \geq 2^{-n-k+1}$. 👍

The AC is used in the proof. It is known that this theorem cannot be proved in ZF+(DC).

XIII.11.10.5. Stone's main result from [Sto48] was the following theorem. Combining this with the result from [Tuk40] that metrizable spaces are fully normal, one obtains that all metrizable spaces are paracompact.

XIII.11.10.6. Theorem. A topological space is paracompact if and only if it is fully normal.

See XIII.7.7.8. for the essence of one direction of the proof.

Paracompact spaces share nice separation properties with compact spaces:

XIII.11.10.7. Proposition. Every paracompact space is normal.

Thus there are Hausdorff spaces, even completely regular spaces, which are not paracompact. There are even locally compact completely normal spaces which are not paracompact (Exercise XIII.11.14.12..)

XIII.11.10.8. A subset of a paracompact space is not paracompact in general (e.g. a subset of a compact Hausdorff space need not be normal). Even an open subset of a paracompact space need not be paracompact (Exercise XIII.11.14.12..) But:

XIII.11.10.9. Proposition. A closed subspace of a paracompact space is paracompact.

The proof is essentially identical to the proof of XIII.11.1.6. with "finite subcover" replaced by "locally finite refinement." In fact, an F_σ in a paracompact space is paracompact.

XIII.11.10.10. Proposition. Let X be a (Hausdorff) topological space, and $\{X_i : i \in I\}$ a partition of X into clopen subsets. Then X is paracompact if and only if each X_i is paracompact.

Proof: If X is paracompact, then each X_i is paracompact by XIII.11.10.9.. Conversely, suppose each X_i is paracompact. If $\mathcal{U}$ is an open cover of X, then, for each i, $\mathcal{U}_i = \{U \cap X_i\}$ is an open cover of X_i, which has a locally finite refinement $\mathcal{V}_i$. Then $\cup_{i\in I}\mathcal{V}_i$ is a locally finite open cover of X which refines $\mathcal{U}$. 👍

Paracompact spaces are the natural setting for some constructions of algebraic topology.

Partitions of Unity

For analysis purposes, the most important feature of paracompact spaces is the existence of "partitions of unity." These provide a systematic and controlled way to write a general real- or complex-valued function as a sum of functions of "small" support, and allow "local" phenomena to be stitched together globally.

XIII.11.10.11. DEFINITION. Let $(X, \mathcal{T})$ be a topological space. A *weak partition of unity* on X is a collection of continuous functions $\{f_i : i \in I\}$ from X to $[0, 1]$ such that

(i) For each $x \in X$, $\sum_{i \in I} f_i(x) = 1$.

The collection is a *partition of unity* if in addition

(ii) Each $x \in X$ has a neighborhood U such that $f_i \neq 0$ on U for only finitely many i.

The collection is a *strong partition of unity* if in addition

(iii) For each i, there is an $x \in X$ with $f_i(x) = 1$.

If $\mathcal{U}$ is an open cover of $(X, \mathcal{T})$, and $\{f_i : i \in I\}$ is a partition of unity on $(X, \mathcal{T})$, then $\{f_i\}$ is *subordinate to* $\mathcal{U}$ if for each $i \in I$ there is a $U \in \mathcal{U}$ such that $Z^c_{f_i} \subseteq U$, where for $f : X \to \mathbb{C}$ continuous Z^c_f is the *cozero set* of f, i.e.

$$Z^c_f = \{x : f(x) \neq 0\}$$

(note that Z^c_f is open for any continuous f). The closure of Z^c_f is called the *support* of f.

XIII.11.10.12. If $\{f_i : i \in I\}$ is a partition of unity, set $U_i = Z^c_{f_i} = \{x \in X : f_i(x) > 0\}$. Then $\{U_i : i \in I\}$ is a locally finite open cover of X. If $\mathcal{U}$ is an open cover of X, then the partition of unity is subordinate to $\mathcal{U}$ if and only if this locally finite open cover is a refinement of $\mathcal{U}$. If a partition of unity is subordinate to a refinement of $\mathcal{U}$, it is subordinate to $\mathcal{U}$.

XIII.11.10.13. There is some lack of uniformity in this terminology; what we call "weak partitions of unity" are often called "partitions of unity." Our term "strong partition of unity" is nonstandard, but will be useful; it is a desirable nondegeneracy requirement. It is equivalent to saying that $\{U_i\}$ is a minimal open cover of X, i.e. no proper subcollection covers X.

XIII.11.10.14. EXAMPLE. Let $X = \mathbb{R}$. *****

XIII.11.10.15. THEOREM. Let $(X, \mathcal{T})$ be a paracompact space. If $\mathcal{U}$ is any open cover of $(X, \mathcal{T})$, then there is a strong partition of unity on X subordinate to $\mathcal{U}$.

PROOF: Passing to a refinement if necessary, we may assume that $\mathcal{U}$ is locally finite. We first construct a smaller open cover by transfinite induction. If $\mathcal{U}$ is countable (or has a countable locally finite refinement, e.g. if X is second countable, or locally compact and σ-compact), the transfinite induction can be simplified to an ordinary induction.

Choose a well ordering $\{U_\alpha\}$ of $\mathcal{U}$. We will inductively define an open cover $\{V_\alpha\}$ such that $\overline{V_\alpha} \subseteq U_\alpha$ for each α. Suppose that V_β has been defined for all $\beta < \alpha$ so that $\{V_\beta : \beta < \alpha\} \cup \{U_\beta : \beta \geq \alpha\}$ is an open cover of X (this assumption is vacuous if $\alpha = 0$). Set

$$F_\alpha = X \setminus [(\cup_{\beta<\alpha} V_\beta) \cup (\cup_{\beta>\alpha} U_\beta)] \ .$$

Then F_α is a closed subset of U_α since $U_\alpha \cup F^c_\alpha = X$. Since X is normal (XIII.11.10.7.), there is an open set V_α with $F_\alpha \subseteq V_\alpha$ and $\overline{V_\alpha} \subseteq U_\alpha$. If $F_\alpha = \emptyset$, choose $V_\alpha = \emptyset$. We have $X = (\cup_{\beta<\alpha} V_\beta) \cup (\cup_{\beta>\alpha} U_\beta) \cup V_\alpha$, so the inductive step is finished.

We claim $\mathcal{V} = \{V_\alpha\}$ is an open cover of X. Suppose not; let $x \in X \setminus (\cup_\alpha V_\alpha)$. Then x is in only finitely many U_α, so there is an α such that $x \notin \cup_{\beta \geq \alpha} U_\beta$. For this α, $x \notin (\cup_{\beta<\alpha} V_\beta) \cup (\cup_{\beta \geq \alpha} U_\beta)$. But $\{V_\beta : \beta < \alpha\} \cup \{U_\beta : \beta \geq \alpha\}$ is an open cover of X, a contradiction. Note that since $V_\alpha \subseteq U_\alpha$ for each α, $\{U_\alpha\}$, $\mathcal{V}$ is locally finite. Also note that if $x \in F_\alpha$, then $x \notin V_\beta$ for $\beta < \alpha$ by construction, and $x \notin V_\beta$ for $\beta > \alpha$ since $V_\beta \subseteq U_\beta$ and $x \notin U_\beta$ by construction. Thus $x \notin V_\beta$ for any $\beta \neq \alpha$.

Now repeat the construction using the V_α in place of the U_α to inductively define open sets W_α and

$$E_\alpha = X \setminus [(\cup_{\beta<\alpha} W_\beta) \cup (\cup_{\beta>\alpha} V_\beta)]$$

such that $E_\alpha \subseteq W_\alpha$ and $\overline{W_\alpha} \subseteq V_\alpha$, and $\{W_\alpha\}$ is an open cover of X. Note that $F_\alpha \subseteq E_\alpha$ for each α.

Discard any α for which $V_\alpha = \emptyset$, i.e. $F_\alpha = \emptyset$ (since $F_\alpha \subseteq W_\alpha$, these are also the α for which $W_\alpha = \emptyset$). For each remaining α let g_α be a continuous function from X to $[0,1]$ such that $g_\alpha \equiv 1$ on $\overline{W_\alpha}$ and $g_\alpha \equiv 0$ on $X \setminus V_\alpha$ (XIII.7.6.13.). For $x \in X$, set $g(x) = \sum_\alpha g_\alpha(x)$. Since $\mathcal{V}$ is locally finite, only finitely many g_α are nonzero in a neighborhood of x, so $g(x)$ is finite, and g is continuous at x. Since $g_\alpha(x) = 1$ for at least one α, $g(x) \geq 1$. For each x set $f_\alpha(x) = g_\alpha(x)/g(x)$; then f_α is continuous, $0 \leq f_\alpha(x) \leq 1$ for all α and x, $f_\alpha \equiv 0$ on $X \setminus V_\alpha$, and for each x, only finitely many $f_\alpha(x)$ are nonzero, and $\sum_\alpha f_\alpha(x) = 1$. For any α, if $x \in F_\alpha$, then $x \in W_\alpha$ but $x \notin V_\beta$ for any $\beta \neq \alpha$, so $f_\alpha(x) = 1$, and thus the f_α satisfy (iii). Thus $\{f_\alpha\}$ is a partition of unity subordinate to $\mathcal{U}$. 👍

Note that this proof is a little more complex than the standard proof of existence of partitions of unity, because of condition (iii). What the proof really shows is that if $\mathcal{U}$ is a locally finite cover of a normal space X, then there is a strong partition of unity subordinate to $\mathcal{U}$. Note that this proof requires the Axiom of Choice.

XIII.11.10.16. Conversely, a partition of unity subordinate to an open cover gives a locally finite refinement of the cover, so if every open cover of X supports a partition of unity subordinate to it, X must be paracompact. There is a better result:

XIII.11.10.17. THEOREM. Let X be a T_1 space. If for every open cover $\mathcal{U}$ of X there is a weak partition of unity subordinate to $\mathcal{U}$, then X is paracompact.

PROOF: [Eng89] We first note that X must be completely regular. If E is a closed set in X and $p \notin E$, choose a partition of unity (f_i) subordinate to the open cover $\{X \setminus E, X \setminus \{p\}\}$. There is an i such that $f_i = 0$ on E and $f_i(p) > 0$, and it gives an Urysohn function for (E, p).

Now let $\mathcal{U}$ be an open cover of X, and $\{f_i : i \in I\}$ a partition of unity subordinate to $\mathcal{U}$. If ϕ is a strictly positive real-valued continuous function on X, then for each $x_0 \in X$ there is an open neighborhood U of x_0 and a finite set $I_0 \subseteq I$ such that $f_i(x) < \phi(x)$ for all $x \in U$ and $i \notin I_0$ [Choose I_0 so that $1 - \sum_{i \in I_0} f_i(x_0) < \phi(x_0)$, and $U = \{x \in X : 1 - \sum_{i \in I_0} f_i(x) < \phi(x)\}$.]

For $x \in X$, set $\phi(x) = \sup_{i \in I} f_i(x)$. Then ϕ is strictly positive, and since the supremum is a finite supremum in a neighborhood of each point, ϕ is continuous. For each i set

$$V_i = \left\{ x \in X : f_i(x) > \frac{1}{2}\phi(x) \right\} .$$

Then $\{V_i : i \in I\}$ is a locally finite open cover of X which refines $\mathcal{U}$. 👍

Here is a sample application of XIII.11.10.15.:

XIII.11.10.18. Corollary. Let (X, ρ) be a metric space. Then for any $\epsilon > 0$ there is a strong partition of unity $\{f_i : i \in I\}$ such that the diameter of $Z^c_{f_i}$ is less than ϵ for all i.

Proof: X is paracompact (XIII.11.10.4.), and the balls of radius $< \epsilon/2$ form an open cover of X. 👍

Paracompactness and Local Compactness

XIII.11.10.19. Proposition. Let X be a topological space, K a compact subset of X, and $\{U_i : i \in I\}$ a locally finite open cover of X. Then $K \cap U_i \neq \emptyset$ for only finitely many i.

Proof: For each $x \in K$, there is an open neighborhood V_x of x such that $V_x \cap U_i \neq \emptyset$ only for i in a finite subset F_x of I. Then $\{V_x : x \in K\}$ is an open cover of K, so there is a finite subcover $\{V_{x_1}, \ldots, V_{x_n}\}$. The set U_i can have nonempty intersection with K only if i is in the finite subset $F_{x_1} \cup \cdots \cup F_{x_n}$ of I. 👍

XIII.11.10.20. Proposition. Let X be a (necessarily Hausdorff) topological space which is locally compact, connected, and paracompact. Then X is σ-compact.

Proof: For each $x \in X$, let U_x be an open neighborhood of x with $\bar{U}_x$ compact. Let $\{V_i : i \in I\}$ be a locally finite refinement of $\{U_x : x \in X\}$. Then $\bar{V}_i$ is compact for each i.
Fix $i_1 \in I$ with V_{i_1} nonempty, and set $W_1 = V_{i_1}$, $K_1 = \bar{V}_{i_1}$. There are only finitely many i for which $V_i \cap K_1 \neq \emptyset$ (); let W_2 be their union, and $K_2 = \bar{W}_2$. Then K_2 is also compact, and $K_1 \subseteq W_2$. Repeat the process to get W_n and K_n for each n. Set $W = \cup_n W_n = \cup_n K_n$. Then W is clearly open in X and σ-compact. If $i \in I$ and $V_i \cap W \neq \emptyset$, then $V_i \cap K_n \neq \emptyset$ for some n, so $V_i \subseteq W_{n+1} \subseteq W$. Thus $X \setminus W$ is the union of the V_i which are disjoint from W, hence open; thus W is clopen. Since X is connected and $W \neq \emptyset$, $W = X$. 👍

Conversely:

XIII.11.10.21. Theorem. A locally compact, σ-compact Hausdorff space is paracompact. In fact, any open cover of a locally compact, σ-compact Hausdorff space has a countable locally finite refinement.

XIII.11.11. Compactly Generated Spaces

The great majority of topological spaces encountered naturally, if not compact, at least have their topology completely determined by their compact subsets.

XIII.11.11.1. Definition. Let X be a topological space. X is *compactly generated* if any subset U of X is open (in X) if (and only if) $U \cap K$ is relatively open in K for every compact subset K of X. Equivalently, X is compactly generated if any subset E of X is closed (in X) if (and only if) $E \cap K$ is relatively closed in K for every compact subset K of X.

In other words, X has the weak topology generated by the inclusion maps of its compact subsets. A set $U \subseteq X$ such that $U \cap K$ is relatively open in K for every compact $K \subseteq X$ is called a *k-open* set in X (cf. XIII.11.14.16.), and a compactly generated space is called a *k-space* in some references like [Dug78].

XIII.11.11.2. If X is a Hausdorff space, then X is compactly generated if and only if, for every subset E of X, E is closed if and only if $E \cap K$ is compact for every compact subset K of X.

Any compact space is obviously compactly generated. More generally:

XIII.11.11.3. PROPOSITION. Every locally compact (Hausdorff) space is compactly generated.

PROOF: Let X be locally compact, and $U \subseteq X$ k-open. Let $x \in U$. Then x has an open neighborhood V_x in X such that $\bar{V}_x$ (closure in X) is compact. $U \cap \bar{V}_x$ is relatively open in $\bar{V}_x$, i.e. there is an open set W_x in X such that $U \cap \bar{V}_x = W_x \cap \bar{V}_x$ (in particular, $x \in W_x$). Then $V_x \cap W_x$ is an open neighborhood of x in X which is contained in U. Since $x \in U$ is arbitrary, U is open in X. 👍

Perhaps more significantly, every sequential space (XIII.3.5.2.), and in particular every first countable space, is compactly generated. We need a result of independent interest:

XIII.11.11.4. PROPOSITION. Let X be a topological space and (x_n) a sequence in X. If $x_n \to x$, then $K = \{x\} \cup \{x_n : n \in \mathbb{N}\}$ is a compact subset of X.

PROOF: Let $\mathcal{U}$ be an open cover of K. Then $x \in U$ for some $U \in \mathcal{U}$. Since $x_n \to x$, $x_n \in U$ for all but finitely many n. Finitely many other sets in $\mathcal{U}$ cover the x_n which are not in U. 👍

XIII.11.11.5. **Caution:** This statement is false in general if "sequence" is replaced by "net." ✋

XIII.11.11.6. PROPOSITION. Every sequential space is compactly generated. In particular, every first countable space is compactly generated.

PROOF: Let X be sequential, and $U \subseteq X$ k-open. We show that U^c (complement in X) is closed. Since X is sequential, it suffices to show that if (x_n) is a sequence in U^c and $x_n \to x$, then $x \in U^c$. Let $K = \{x\} \cup \{x_n : n \in \mathbb{N}\}$; then K is compact, so $U \cap K$ is relatively open, i.e. $U^c \cap K$ is relatively closed. Since $x_n \in U^c \cap K$ for all n and $x_n \to x \in K$, $x \in U^c \cap K \subseteq U^c$. 👍

XIII.11.11.7. EXAMPLE. There are topological spaces which are not compactly generated. For example, if X is the space of XIII.11.14.10.(b), and $U = \{\alpha\}$, then $U \cap K$ is relatively open for every compact (i.e. finite) subset K of X. But U is not open in X.

This example is countable, hence σ-compact, and completely normal.

XIII.11.11.8. DEFINITION. Let X and Y be topological spaces, and $f : X \to Y$ a function. Then f is *continuous on compact sets* if, for every compact $K \subseteq X$, $f|_K$ is a continuous function from K to Y.

XIII.11.11.9. Any continuous function is continuous on compact sets. The converse is false in general: if X is the space of XIII.11.14.10.(b) and Y is any topological space, any function from X to Y is continuous on compact (i.e. finite) sets. However:

XIII.11.11.10. PROPOSITION. Let X and Y be topological spaces, with X compactly generated, and $f : X \to Y$ a function. If f is continuous on compact sets, it is continuous.

PROOF: Let V be an open set in Y, and $U = f^{-1}(V)$. If K is any compact subset of X, then $U \cap K = (f|_K)^{-1}(V)$ is relatively open in K since $f|_K$ is continuous. Thus U is open in X since X is compactly generated. 👍

XIII.11.12. Proper Maps

Proper maps can be defined between arbitrary topological spaces, but are particularly important for maps between locally compact spaces, especially manifolds.

XIII.11.12.1. DEFINITION. Let X and Y be topological spaces. A *proper map* from X to Y is a continuous function $f : X \to Y$ such that the preimage of every compact subset of Y is a compact subset of X.

There are variations in this definition in the literature, e.g. some authors require proper maps to be closed; but all the variations are equivalent for reasonable spaces (e.g. compactly generated Hausdorff spaces). ✋

XIII.11.12.2. Since singletons are compact, if $f : X \to Y$ is proper, then $f^{-1}(\{y\})$ is compact for every $y \in Y$.

If X is compact and Y is Hausdorff, then every continuous function from X to Y is proper. If Y is compact and $f : X \to Y$ is proper, then X is also compact.

XIII.11.12.3. EXAMPLE. The following example shows that the continuity condition is not automatic, even for the nicest spaces. Let $Y = \mathbb{N} \cup \{0\}$ with the discrete topology, and $X = \left\{\frac{1}{n} : n \in \mathbb{N}\right\} \cup \{0\} \subseteq \mathbb{R}$. Define $f : X \to Y$ by $f\left(\frac{1}{n}\right) = n$ for each n and $f(0) = 0$. Then f is a bijection, the preimage of every compact subset of Y is compact since every compact subset of Y is finite, but f is not continuous.

Proper maps need not be closed (XIII.11.14.29.), but they are under mild conditions:

XIII.11.12.4. PROPOSITION. Let X be a topological space and Y a compactly generated Hausdorff space, and $f : X \to Y$ a proper map. Then f is a closed mapping.

PROOF: Let E be a closed subset of X. If K is a compact subset of Y, then $f^{-1}(K)$ is compact, so $E \cap f^{-1}(K)$ is compact. Thus $f(E \cap f^{-1}(K))$ is compact since f is continuous. But $f(E \cap f^{-1}(K)) = f(E) \cap K$. Thus $f(E) \cap K$ is compact for every compact $K \subseteq Y$. By XIII.11.11.2., $f(E)$ is closed in Y. 👍

For locally compact spaces, proper maps have an alternate characterization:

XIII.11.12.5. PROPOSITION. Let X and Y be locally compact, and $f : X \to Y$ a continuous function. Let $X^\dagger$ and $Y^\dagger$ be the one-point compactifications, and define $f^\dagger : X^\dagger \to Y^\dagger$ by $f^\dagger(x) = f(x)$ if $x \in X$ and $f(\infty) = \infty$. Then f is proper if and only if $f^\dagger$ is continuous.

XIII.11.12.6. Informally, f is proper if and only if it sends points "close to infinity" to points in Y "close to infinity." If X and Y are locally compact and σ-compact, this has an alternate formulation. Say a sequence (x_n) in a topological space *escapes to infinity* if every compact subset of the space contains only finitely many of the x_n. Then $f : X \to Y$ (continuous) is proper if and only if, whenever (x_n) escapes to infinity in X, $(f(x_n))$ escapes to infinity in Y.

XIII.11.12.7. Proper maps have other equivalent characterizations (cf. XIII.11.14.25.), and closed surjective proper maps preserve much structure. See e.g. [Dug78, IX.5] (cf. XIII.11.14.26.) or [Bou98, I.10].

XIII.11.13. Realcompact Spaces

There is another generalization of compactness, realcompactness. As with paracompactness, the vast majority of reasonable topological spaces are realcompact. There are also a large number of equivalent characterizations, of a varied enough nature that it is a rather natural notion.

Realcompact spaces were first defined in 1948 by E. HEWITT [**?**]. They have been called by a number of names (HEWITT'S was *Q-space*); the term *realcompact* was introduced in [GJ76], which is a prime reference about these spaces. Other useful references are (), (), ()

XIII.11.13.1. Our beginning approach is an algebraic one. In this section, if X is a completely regular topological space (the theory works satisfactorily only in this case), we let $C(X)$ be the ring of real-valued continuous functions on X and $BC(X)$ the subring of bounded real-valued continuous functions on X. In some references $BC(X)$ is denoted $C^*(X)$, but as an operator algebraist I find this notation unsuitable for reasons discussed in (). We could equally well work with complex-valued continuous functions, but there are convenient technical simplifications in the real-valued case (e.g. unital ring-homomorphisms between such $C(X)$ are automatically $\mathbb{R}$-algebra homomorphisms).

XIII.11.13.2. If $p \in X$, then $M_p = \{f \in C(X) : f(p) = 0\}$ is a maximal ideal in $C(X)$, and $C(X)/M_p \cong \mathbb{R}$; the map $f \mapsto f(p)$ is a homomorphism from $C(X)$ to $\mathbb{R}$ with kernel M_p. This homomorphism is called *point evaluation at* p, and M_p is called a *fixed* maximal ideal. Not every maximal ideal in $C(X)$ is fixed in general; indeed, every maximal ideal is fixed if and only if X is compact ().

Similarly, $\tilde{M}_p = M_p \cap BC(X)$ is a maximal ideal in $BC(X)$, the kernel of the evaluation map at p. We again have that every maximal ideal of $BC(X)$ is fixed if and only if X is compact (in this case, $BC(X) = C(X)$).

XIII.11.13.3. If M is a maximal ideal of $C(X)$, then $C(X)/M$ is an ordered field. The image of the constant functions is a subfield isomorphic to $\mathbb{R}$. If this subfield is the whole quotient, the maximal ideal is *real*; otherwise it is *hyperreal* and the quotient is a nonarchimedean ordered field. $C(X)$ will have hyperreal maximal ideals unless $C(X) = BC(X)$ (X is *pseudocompact*).

XIII.11.13.4. DEFINITION. Let X be a completely regular space. Then X is *realcompact* if every real maximal ideal of $C(X)$ is fixed, i.e. every homomorphism from $C(X)$ to $\mathbb{R}$ is a point evaluation at a point of X.

XIII.11.13.5. EXAMPLES. $\mathbb{N}$, $\mathbb{R}$

One of the main characterizations of realcompactness is the following, which we will gradually prove (the proof is ()) obtaining an interesting construction along the way. It is an analog of () for compact spaces.

XIII.11.13.6. THEOREM. Let X be a completely regular space. Then X is realcompact if and only if it is homeomorphic to a closed subspace of a product of copies of $\mathbb{R}$.

XIII.11.14. Exercises

XIII.11.14.1. (a) Show that any set with the indiscrete topology () is compact.

(b) Show that any set with the finite complement topology () is compact and T_1.

(c) Show that every subset of the examples in (a) and (b) is compact. Thus a compact subset of a compact T_1-space need not be closed.

XIII.11.14.2. The aim of this exercise is to show without AC that a compact subset of a Hausdorff space is closed. Let X be a Hausdorff space, and $Y \subseteq X$. Suppose Y is not closed, and let $p \in \bar{Y} \setminus Y$. Fix a net (y_i) in Y converging to p (XIII.3.2.21.). Then the net (y_i) cannot have a cluster point in Y (XIII.7.4.1.), so Y is not compact (XIII.11.2.1.).

XIII.11.14.3. Here is an alternate proof of XIII.11.3.10. which gives an explicit Lebesgue number. Let (X, ρ) be a compact metric space, and $\mathcal{U}$ an open cover of X. Let $\{U_1, \ldots, U_n\}$ be a finite subcover of $\mathcal{U}$. For $x \in X$, define

$$f(x) = \frac{1}{n} \sum_{k=1}^{n} \rho(x, U_k^c) \ .$$

Then f is a continuous function from X to $(0, \infty)$ ().

(a) Let δ be the minimum of f on X. Show that $\delta > 0$.

(b) Show that δ is a Lebesgue number for $\mathcal{U}$.

XIII.11.14.4. Let $X = F([0,1],[0,1]) \cong [0,1]^{\mathfrak{c}}$ be the space of all functions from $[0,1]$ to itself, with the product topology (topology of pointwise convergence) (). Recall (XIII.6.3.2.–XIII.6.3.3.) that X is separable but not first countable. X is compact by Tikhonov's Theorem (XIII.11.6.6. suffices), and Hausdorff.

Let Y be the subset of X consisting of all nondecreasing functions from $[0,1]$ to $[0,1]$, with the relative topology (also the topology of pointwise convergence). Y is called the *Helly space*.

(a) Show that Y is a closed subspace of X, hence compact and Hausdorff.

(b) Show that the set of nondecreasing step functions () with only rational discontinuities, taking only rational values, is dense in Y. Conclude that Y is separable.

(c) Show that Y is first countable. [Use the fact () that a nondecreasing function has only jump discontinuities, and only finitely many jumps of more than ϵ for any fixed $\epsilon > 0$.]

(d) For each $t \in [0,1]$, let f_t be the function with $f_t(x) = 0$ for $x < t$, $f_t(x) = 1$ for $x > t$, and $f_t(t) = 1/2$. Show that $\{f_t : t \in [0,1]\}$ is an uncountable subset of Y which is discrete in the relative topology. Conclude that Y is not second countable, and hence not metrizable.

(e) Although $F([0,1],[0,1])$ has cardinality $2^{2^{\aleph_0}}$, show that Y has cardinality $2^{\aleph_0}$ (cf. XIII.7.10.6.).

XIII.11.14.5. (a) Let X be the unit square $[0,1] \times [0,1]$. Give X the order topology () from the lexicographic ordering (). Show that X is compact, Hausdorff, first countable, but not separable, hence not metrizable.

(b) Do the same as in (a) with the second copy of $[0,1]$ replaced by a three-point set with a total ordering. What if a two-point set is used instead?

XIII.11.14.6. The proof of Tikhonov's theorem (XIII.11.6.1.) uses the Axiom of Choice. Conversely, the Axiom of Choice can be proved from Tikhonov's Theorem for T_1 spaces, showing that the two are equivalent.

(a) Let X be a set. Define the *T_1-compactification of $\mathbb{N}$ by X* to be the disjoint union $\mathbb{N} \cup X$, topologized by taking as the open sets any subset with finite complement, and any subset of $\mathbb{N}$. Show that $\mathbb{N} \cup X$ is compact if X is nonempty, always T_1, but not Hausdorff if X has more than one element.

(b) Let $\{X_i : i \in I\}$ be an indexed collection of nonempty sets, and let $X = \prod_{i \in I} X_i$. Show using Tikhonov's Theorem that X is nonempty, as follows. For each i, let Y_i be the T_1 compactification of $\mathbb{N}$ by X_i, and let $Y = \prod_{i \in I} Y_i$. Then $X \subseteq Y$, and Y is compact by Tikhonov's Theorem.

(c) $Y \neq \emptyset$: for each $n \in \mathbb{N}$ let $y_n \in Y$ be the point with i'th coordinate n for all i. By compactness, the sequence (y_n) has a cluster point p in Y. Show that for each i, the i'th coordinate p_i of p is in X_i, so $p \in X$. (Actually every point of X is a cluster point of this sequence.)

Note that Tikhonov's Theorem for *Hausdorff* spaces is strictly weaker than the Axiom of Choice (it is equivalent to the BPI). See II.6.3.11. for a discussion of this and related matters.

XIII.11.14.7. **The Tikhonov Plank.** Let X be the set of ordinals less than or equal to the first uncountable ordinal ω_1, and Y the set of ordinals less than or equal to ω. Then X and Y are compact Hausdorff spaces in the order topology (). So $X \times Y$ is a compactum, hence normal. Let $Z = (X \times Y) \setminus (\omega_1, \omega)$ (Z is the "plank" $X \times Y$ with the "upper right-hand corner" removed).

(a) Z is an open set in $X \times Y$, hence locally compact; $Z^\dagger \cong X \times Y$. Show that $X \times Y$ is also βZ (cf. XIII.11.9.17.). Z is pseudocompact (XIII.11.9.17.).

(b) Let $T = \{(\alpha, \omega) : \alpha < \omega_1\}$ and $R = \{(\omega_1, \alpha) : \alpha < \omega\}$ be the "top and right edges" of Z. Show that T and R are disjoint closed subsets of Z which do not have disjoint neighborhoods in Z. Thus Z is not normal.

XIII.11.14.8. **Rational Sequence Topology.** [SS95, #65] Put a topology on $\mathbb{R}$ as follows. For each irrational $x \in \mathbb{R}$, fix a sequence $(q_{x,k})$ of rational numbers converging to x. Let $\mathcal{B}$ be the following collection of subsets of $\mathbb{R}$:

(i) $\{q\}$ for each $q \in \mathbb{Q}$.

(ii) $\{x\} \cup \{q_{x,k} : k \geq n\}$ for each $x \in \mathbb{J} = \mathbb{R} \setminus \mathbb{Q}$, $n \in \mathbb{N}$.

(a) Show that if $x, y \in \mathbb{J}$, $x \neq y$, then $\{q_{x,k} : k \in \mathbb{N}\} \cap \{q_{y,k} : k \in \mathbb{N}\}$ is finite.

(b) Show that $\mathcal{B}$ is a base for a topology $\mathcal{Q}$ on $\mathbb{R}$.

(c) Show that $\mathcal{Q}$ is stronger than the ordinary topology on $\mathbb{R}$. Thus $(\mathbb{R}, \mathcal{Q})$ is Hausdorff. Each rational number is an isolated point of $(\mathbb{R}, \mathcal{Q})$.

(d) $(\mathbb{R}, \mathcal{Q})$ is first countable.

(e) $\mathbb{Q}$ is a dense open subset of $(\mathbb{R}, \mathcal{Q})$; thus $(\mathbb{R}, \mathcal{Q})$ is separable.

(f) Each set in $\mathcal{B}$ is compact in $(\mathbb{R}, \mathcal{Q})$ (XIII.11.11.4.). Thus there is a base for $\mathcal{Q}$ consisting of compact open sets, so $(\mathbb{R}, \mathcal{Q})$ is locally compact and $ind(\mathbb{R}, \mathcal{Q}) = 0$ (XIII.22.3.2.). In particular, $(\mathbb{R}, \mathcal{Q})$ is completely regular (XIII.11.8.9.) and totally disconnected (XIII.13.6.1.).

(g) $\mathbb{J}$ is a discrete closed subset of $(\mathbb{R}, \mathcal{Q})$ of cardinality $2^{\aleph_0}$. Thus by XIII.7.10.13., $(\mathbb{R}, \mathcal{Q})$ is not normal.

(h) By XIII.13.6.5., $dim(\mathbb{R}, \mathcal{Q}) > 0$, i.e. $(\mathbb{R}, \mathcal{Q})$ is not zero-dimensional (XIII.13.6.3.). Also, $Ind(\mathbb{R}, \mathcal{Q}) > 0$.

XIII.11.14.9. Compactifications of $\mathbb{N}$. (a) Let X be a separable compactum. Fix a countable dense set (x_k) in X, and let (p_n) be a sequence of points of X with each p_n some x_k, and for each k $p_n = x_k$ for infinitely many n. Set

$$N = \left\{ \left(\frac{1}{n}, p_n \right) : n \in \mathbb{N} \right\} \subseteq [0,1] \times X$$

and $\tilde{X} = N \cup [\{0\} \times X]$. Show that $\tilde{X}$ is compact and Hausdorff, N is a dense open subset of $\tilde{X}$ homeomorphic to $\mathbb{N}$, and $\tilde{X} \setminus N$ is homeomorphic to X. Thus $\tilde{X}$ can be regarded as a compactification of $\mathbb{N}$.

(b) If X and Y are nonhomeomorphic separable compacta, say that the compactifications $\tilde{X}$ and $\tilde{Y}$ are *topologically distinct*. Topologically distinct compactifications of $\mathbb{N}$ are distinct compactifications.

(c) There are at least $2^{2^{\aleph_0}}$ pairwise nonhomeomorphic separable compact Hausdorff spaces (XIII.20.5.11.). Thus there are at least $2^{2^{\aleph_0}}$ topologically distinct compactifications of $\mathbb{N}$.

(d) Not every compactification of $\mathbb{N}$ is of this form: if Z is a compactification of $\mathbb{N}$, then $Z \setminus \mathbb{N}$ is not necessarily separable. [$\beta\mathbb{N} \setminus \mathbb{N}$ is not separable.]

(e) Distinct compactifications of $\mathbb{N}$ need not be topologically distinct. [There are $2^{\aleph_0}$ distinct two-point compactifications of $\mathbb{N}$ since there are $2^{\aleph_0}$ essentially different ways of splitting $\mathbb{N}$ into two infinite sets.]

(f) There are at most $2^{2^{\aleph_0}}$ distinct compactifications of $\mathbb{N}$. [Distinct compactifications define distinct uniformities on $\mathbb{N}$ (), and there are at most $2^{2^{\aleph_0}}$ uniformities on $\mathbb{N}$ (XIII.15.12.3.(a)).]

(g) Conclude that there are exactly $2^{2^{\aleph_0}}$ distinct compactifications of $\mathbb{N}$, and exactly $2^{2^{\aleph_0}}$ distinct separable compacta up to homeomorphism.

XIII.11.14.10. Let α be a point of $\beta\mathbb{N} \setminus \mathbb{N}$, and let $\mathcal{F}$ be the corresponding free ultrafilter on $\mathbb{N}$. Set $X = \mathbb{N} \cup \{\alpha\}$, with the relative topology. Then X is a countable completely regular (hence normal) space.

(a) Show that the open sets in X consist precisely of all subsets of $\mathbb{N}$ and all sets of the form $F \cup \{\alpha\}$, where $F \in \mathcal{F}$. (The space X could have been defined without reference to $\beta\mathbb{N}$ by just taking α to be a point not in $\mathbb{N}$ and defining the open sets this way for the free ultrafilter $\mathcal{F}$.)

(b) Show that α does not have a countable neighborhood base in X. If $x_n = n$, then α is a cluster point of the sequence (x_n), but no subsequence of (x_n) converges to α. [Every infinite subset of $\mathbb{N}$ contains an infinite subset which is not in $\mathcal{F}$.]

(c) Although every singleton subset of X is a G_δ, X is not first countable.

(d) X is not compact; in fact, the only compact subsets of X are the finite subsets. The identity map from $\mathbb{N}$ to X extends to a homeomorphism from $\beta\mathbb{N}$ to βX [GJ76,].

If γ is another point of $\beta\mathbb{N} \setminus \mathbb{N}$, corresponding to another free ultrafilter $\mathcal{G}$ on $\mathbb{N}$, and $Y = \mathbb{N} \cup \{\gamma\}$, then Y is not necessarily homeomorphic to X; in fact, there are $2^{2^{\aleph_0}}$ homeomorphism classes of such spaces since there are $2^{2^{\aleph_0}}$ points in $\beta\mathbb{N} \setminus \mathbb{N}$ and only $2^{\aleph_0}$ permutations of $\mathbb{N}$ [GJ76,].

(e) More generally, consider a topological space X which is countably infinite and in which every point but one is an isolated point. Identify the isolated points with $\mathbb{N}$ and call the nonisolated point α. Show that X is necessarily perfectly normal, and that there is a free filter $\mathcal{F}$ on $\mathbb{N}$ such that the open sets in X consist of all subsets of $\mathbb{N}$ and all sets of the form $F \cup \{\alpha\}$ for $F \in \mathcal{F}$. There is a weakest such topology on $\mathbb{N} \cup \{\alpha\}$, which is the one-point compactification of $\mathbb{N}$; the maximal such topologies are the ones described in (i) (i.e. where $\mathcal{F}$ is an ultrafilter). Every such topology is obtained by taking a compactification Y of $\mathbb{N}$ and letting α be a point of $Y \setminus \mathbb{N}$. [Consider βX.]

See XIII.6.3.4. for a more dramatic example of a countable normal space which is not first countable.

XIII.11.14.11. Let X be a Hausdorff space. Show that X is countably compact if and only if X does not contain a closed subspace homeomorphic to $\mathbb{N}$ (i.e. a countably infinite closed subset whose relative topology is the discrete topology).

XIII.11.14.12. Let X be the space ω_1 of ordinals less than the first uncountable ordinal (), with the order topology (). Let $\mathcal{U}$ be the open cover of X consisting of all sets of the form $[0, \alpha)$ for $\alpha \in X$. Show that $\mathcal{U}$ has no locally finite refinement. [If $\mathcal{V}$ is a refinement of $\mathcal{U}$, show that there is a sequence (V_n) of distinct sets in $\mathcal{V}$ and a strictly increasing sequence (α_n) in X with $\alpha_n \in V_n$, and thus that $\mathcal{V}$ cannot be locally finite at $\alpha = \sup \alpha_n$.] Thus X is locally compact and normal, even completely normal (XIII.16.1.1.), but not paracompact.

XIII.11.14.13. Show that a locally finite open cover of a separable topological space is countable. Conclude that every separable paracompact space is Lindelöf.

XIII.11.14.14. Let X be a topological space. Define a one-point compactification of X as follows. Let $X^\dagger$ be $X \cup \{\infty\}$, where ∞ is something not in X. A subset U of $X^\dagger$ is open if $U \subseteq X$ and U is open in X, or if $\infty \in U$ and $X \setminus U$ is closed and compact in X (regarding $\emptyset$ as a closed compact subset of X).

(a) Show that such open sets form a topology on $X^\dagger$.

(b) Show that the relative topology on X from $X^\dagger$ is the original topology on X.

(c) Show that $X^\dagger$ is compact.

(d) Show that X is dense in $X^\dagger$ if and only if X is not compact.

(e) Show that $\{\infty\}$ is closed in $X^\dagger$; hence $X^\dagger$ is T_1 if and only if X is T_1.

(f) Show that $X^\dagger$ is Hausdorff if and only if X is Hausdorff and locally compact.

XIII.11.14.15. An open cover $\mathcal{U}$ of a topological space X is *minimal* if no proper subset of $\mathcal{U}$ is a cover, i.e. for each $U \in \mathcal{U}$ there is at least one point of U not contained in any other set in $\mathcal{U}$. Note that if $\{f_i : i \in I\}$ is a strong partition of unity for X, then $\mathcal{U} = \{Z_{f_i}^c : i \in I\}$ is a minimal locally finite cover of X, so XIII.11.10.15. shows (under AC) that any open cover of a paracompact space has a minimal locally finite refinement.

Prove the following theorem:

THEOREM. Let $\mathcal{U}$ be a point-finite open cover of a topological space X. Suppose the index set of $\mathcal{U}$ can be well ordered (e.g. is countable). Then $\mathcal{U}$ has a minimal subcover.

(a) Write $\mathcal{U} = \{U_\alpha : \alpha < \sigma\}$ where σ is an ordinal. Transfinitely define subsets $\mathcal{U}_\alpha$ and $\mathcal{V}_\alpha$ of $\mathcal{U}$ as follows. Set $\mathcal{V}_\alpha = \cap_{\beta<\alpha}\mathcal{U}_\beta$ ($\mathcal{V}_0 = \mathcal{U}$) and let $\mathcal{U}_\alpha$ be $\mathcal{V}_\alpha \setminus \{U_\alpha\}$ if this is a cover of X and $\mathcal{U}_\alpha = \mathcal{V}_\alpha$ otherwise. (In other words, remove U_α from the cover if possible, still retaining a cover.)

(b) Using point-finiteness of $\mathcal{U}$, show that $\mathcal{V}_\alpha$ is a cover of X for each α (this is only an issue if α is a limit ordinal).

(c) Show that $\mathcal{V} = \cap_{\alpha<\sigma}\mathcal{V}_\alpha$ is a minimal subcover of $\mathcal{U}$. (Minimality is obvious, and the argument of (b) shows that $\mathcal{V}$ is a cover.)

(d) Show that the conclusion of the theorem is not necessarily true if $\mathcal{U}$ is not point-finite, even if it is countable. [Consider the open cover of $\mathbb{R}$ by open intervals of the form $(-\infty, n)$ for $n \in \mathbb{N}$.]

XIII.11.14.16. Let $(X, \mathcal{T})$ be a topological space, and let $\mathcal{T}^k$ be the set of k-open sets in X. Show that $\mathcal{T}^k$ is a topology on X which is stronger than $\mathcal{T}$; $(X, \mathcal{T})$ is compactly generated if and only if $\mathcal{T}^k = \mathcal{T}$. Is $(X, \mathcal{T}^k)$ always compactly generated? [Show that $\mathcal{T}$ and $\mathcal{T}^k$ have the same compact sets.]

XIII.11.14.17. (D. E. COHEN; cf. [Dug78, XI.9.4]) (a) Show that every compactly generated Hausdorff space is a quotient of a locally compact space. [Consider the separated union of the compact subsets of the space.]

(b) Show that any quotient space of a locally compact space is compactly generated. [If $f : X \to Y$ is a quotient map with X locally compact and U is a k-open set in Y, show that $f^{-1}(U) \cap V$ is open in X for any open set V in X with compact closure.]

(c) ([Dug78, XI.9.5], but note that a Hausdorff hypothesis is omitted there) Conclude that any quotient space of a compactly generated Hausdorff space is compactly generated.

XIII.11.14.18. Define a continuous function $f : \mathbb{R} \to \mathbb{R}$ as follows. Let $\{r_n : n \in \mathbb{Z}\}$ be an enumeration of $\mathbb{Q}$. For each n, set $f(x) = r_n$ if $x \in \left[n - \frac{1}{3}, n + \frac{1}{3}\right]$, and extend f to a continuous function on $\mathbb{R}$ by making it linear on $\left[n + \frac{1}{3}, n + \frac{2}{3}\right]$. If

$$U = \bigcup_{n\in\mathbb{Z}} \left(n - \frac{1}{3}, n + \frac{1}{3}\right)$$

then U is open, but $f(U) = \mathbb{Q}$ is not a G_δ (XIII.12.1.5.). Thus, even if X and Y are locally compact and σ-compact, the image of an open set in X under a continuous function $f : X \to Y$ is not necessarily a G_δ, so the image of a G_δ is *a fortiori* not always a G_δ. Contrast with XIII.11.7.11..

XIII.11.14.19. Let X and Y be topological spaces, A and B compact subsets of X and Y respectively, and W an open set in $X \times Y$ containing $A \times B$. Show that there are open sets U and V in X and Y containing A and B respectively, such that $U \times V \subseteq W$, as follows.

(a) For each $x \in A$ and $y \in B$ there are open neighborhoods $U_{x,y}$ of x and $V_{x,y}$ of y such that $U_{x,y} \times V_{x,y} \subseteq W$. [Open rectangles in $X \times Y$ form a base for the product topology.]

(b) Fix $y \in B$. There are open neighborhoods U_y of A and V_y of y such that $U_y \times V_y \subseteq W$. [The sets $\{U_{x,y} : x \in A\}$ form an open cover of A; let $\{U_{x_1,y}, \ldots, U_{x_n,y}\}$ be a finite subcover, and set $U_y = \cup_{j=1}^n U_{x_j,y}$ and $V_y = \cap_{j=1}^n V_{x_j,y}$.]

(c) The sets $\{V_y : y \in B\}$ form an open cover of B. Let $\{V_{y_1}, \ldots, V_{y_m}\}$ be a finite subcover, and set $U = \cap_{j=1}^m U_{y_j}$ and $V = \cup_{j=1}^m V_{y_j}$.

(d) Modify the argument as in () to avoid any use of the Axiom of Choice.

XIII.11.14.20. **Compacta without AC** ([L63], [Bru83]). (a) If no form of Choice is assumed, there exist infinite compacta on which every real-valued continuous function is constant ([L63]; the space constructed there is only locally compact, but its one-point compactification can be constructed without any Choice and has the same property, or, since the space is orderable, a compactification which is also orderable). Such a space is connected () and is called a *Läuchli continuum.* A Läuchli continuum is normal (XIII.11.4.3.) but not completely regular.

(b) Conclude that Urysohn's Lemma and XIII.7.8.3. require some form of Choice.

XIII.11.14.21. **Minimal Compactifications.** Let X be a completely regular space, and (Y, ϕ) a compactification of X.

(a) If $Y \setminus \phi(X)$ has two distinct points p and q, let Z be the identification space $Y/\{p, q\}$. Show that Z is a compactum [cf. XIII.8.3.6. or XIII.8.3.9..]

(b) If $\omega : Y \to Z$ is the quotient map and $\psi = \omega \circ \phi$, show that (Z, ψ) is a compactification of X which is strictly dominated by (Y, ϕ).

(c) Conclude that X has a minimal compactification if and only if it is locally compact, in which case $X^\dagger$ is the unique minimal compactification (actually the smallest compactification).

XIII.11.14.22. Show without any form of Choice that if X is a compact first countable topological space, then X is sequentially compact. [Use XIII.3.1.9. and XIII.11.2.1..] In particular, since it can be proved without any Choice that $[0, 1]^{\mathbb{N}}$ is compact (), and it is metrizable (), it is sequentially compact.

XIII.11.14.23. Show (in ZF) that the special case of Tikhonov's Theorem where all X_i are the same space is equivalent to the full Tikhonov Theorem, i.e. to prove the Tikhonov Theorem it suffices to show that if X is compact and I is any (nonempty) set, then X^I is compact (and similarly for the Hausdorff version), as follows.

(a) Let $\{X_i : i \in I\}$ be an indexed family of (nonempty) compact spaces. Let X be the one-point compactification () of the separated union of the X_i. Then $\prod_i X_i$ is a closed subset of X^I. Denote by ∞ the point at infinity of X.

(b) Let J be the set of all pairs (F, z), where F is a finite subset of I and $z \in \prod_{i \in F} X_i$. Then J is nonempty, and directed by setting $(F, z) \leq (E, w)$ if $F \subseteq E$. If $z \in \prod_{i \in F} X_i$, denote by $\tilde{z}$ the point of X^I whose

coordinates in F are the same as z and all other coordinates are ∞. Define a net in X^I by $x_{(F,z)} = \tilde{z}$, indexed by J.

(c) Suppose that X^I is compact. Then this net has a cluster point. Show that any cluster point is in $\prod_i X_i$, so $\prod_i X_i$ is nonempty, and compact since it is closed in X^I.

XIII.11.14.24. (a) [GJ76, 9.5] Show that if X is a noncompact completely regular topological space, then no point of $\beta X \setminus X$ is a G_δ. In particular, no point of $\beta X \setminus X$ has a countable local base.

(b) If $p \in X$ has a countable local base in X, then it has a countable local base in βX (XIII.7.10.11.).

(c) If X and Y are completely regular first countable topological spaces (this condition can be relaxed somewhat, cf. [GJ76, 9N]), then any homeomorphism from βX to βY sends X onto Y. In particular, if βX and βY are homeomorphic, then X and Y are homeomorphic.

Result (c) fails in general if X or Y is not first countable (XIII.11.14.10.(d)).

XIII.11.14.25. [Bou98, I.10] A continuous function $f : X \to Y$ is *universally closed* if $f \times id_Z : X \times Z \to Y \times Z$ is closed for every topological space Z. (Such a map is called "proper" in [Bou98].)

(a) Show that if $f : X \to Y$ is universally closed, and $A \subseteq Y$, then $g : f^{-1}(A) \to A$ is universally closed.

(b) Show that if the (unique) function from X to a one-point space is universally closed, then X is compact.

(c) Show that $f : X \to Y$ is universally closed if and only if it is closed and proper. (Thus, by XIII.11.12.4., if Y is compactly generated and Hausdorff, f is universally closed if and only if it is proper.)

XIII.11.14.26. [Dug78, IX.5.3] Let X and Y be topological spaces, and $f : X \to Y$ a continuous function. If f is closed and $f^{-1}(\{y\})$ is compact for every $y \in Y$, show that f is proper.

XIII.11.14.27. Show that if X is Hausdorff and Y is locally compact, and $f : X \to Y$ is proper, then X must also be locally compact. [Cover Y by open sets with compact closure and consider preimages.]

XIII.11.14.28. (AC) Let $\{X_i : i \in I\}$ and $\{Y_i : i \in I\}$ be indexed families of (nonempty) topological spaces, and $f_i : X_i \to Y_i$ a continuous function for each i. Show that $\times_i f_i : \prod_i X_i \to \prod_i Y_i$ is closed and proper if and only if each f_i is closed and proper. Can "closed" be deleted?

XIII.11.14.29. Let $X = \mathbb{N} \cup \{0\}$ with the discrete topology, and $Y = \mathbb{N} \cup \{\alpha\}$ as in XIII.11.14.10.. Let $f : X \to Y$ be the identity on $\mathbb{N}$ and $f(0) = \alpha$. Show that f is proper but not closed. [Every compact subset of Y is finite.]

XIII.11.14.30. A surjective continuous map f from a compact X to a (necessarily compact) Hausdorff Y is a closed mapping (XIII.11.5.1.). Show that f need not be open, even if X is metrizable. [Consider $f : [0,1] \to [0,1]$ defined by $f(t) = \max(0, 2t - 1)$.]

XIII.11.14.31. Prove (using Countable AC) the following slight generalization of a result of ALEKSANDROV:

THEOREM. Let (X, ρ) be a locally compact metric space. If any two points of X are contained in a compact connected subset of X, then X is σ-compact, hence separable and second countable.

Note that the hypotheses are satisfied if X is path-connected (the case considered by ALEKSANDROV). Some such hypothesis is needed, e.g. to rule out an uncountable discrete space.

(a) For $x \in X$, let $r(x)$ be the supremum of the r for which the closed ball of radius r centered at x is compact ($r(x) > 0$ since X is locally compact). Show that either $r(x) = \infty$ for all $x \in X$ or $r(x) < \infty$ for all $x \in X$..

(b) If $r(x) = \infty$ for some (hence every) $x \in X$, then X is the union of the closed balls of radius n around x for all n and is thus σ-compact.

(c) Suppose $r(x) < \infty$ for all $x \in X$. For each x let B_x be the closed ball of radius $r(x)/2$ centered at x. Then B_x is compact. Fix $x_0 \in X$. Set $A_1 = B_{x_0}$ and inductively set

$$A_{n+1} = \bigcup \{B_x : x \in A_n\} .$$

(d) Show inductively that each A_n is compact. [If (x_k) is a sequence in A_{n+1}, for each k there is a $y_k \in A_n$ with $\rho(x_k, y_k) \leq r(y_k)/2$. There is a convergent subsequence of (y_k) converging to some $y \in A_n$. Show that

$$\rho(x_k, y) \leq \frac{3r(y)}{4}$$

for infinitely many k, so infinitely many x_k lie in the compact ball of radius $3r(y)/4$ around y. Hence (x_k) has a convergent subsequence.]

(e) Use the connectedness hypothesis to show that $X = \cup_n A_n$.

XIII.11.14.32. Let (X, ρ) be a metric space.

(a) If $A \subseteq X$ is totally bounded, so is $\bar{A}$. [If $\epsilon > 0$, cover A by finitely many balls of radius $\epsilon/2$, and consider the balls of radius ϵ with the same centers.]

(b) If (X, ρ) is totally bounded, then its completion $(\bar{X}, \bar{\rho})$ is compact.

XIII.11.15. Dini's Theorem

Dini's Theorem is an assertion that certain sequences of functions which converge pointwise actually converge uniformly. Since uniform convergence has all sorts of nice consequences not shared by pointwise convergence, such a result is very useful.

XIII.11.15.1. Definition. Let X be a set, and (f_n) a sequence of real-valued functions on X. The sequence (f_n) is *nondecreasing* if, for each $x \in X$, the sequence $(f_n(x))$ of real numbers is nondecreasing. Similarly, (f_n) is *nonincreasing* if $(f_n(x))$ is nonincreasing for all $x \in X$. The sequence (f_n) is *monotone* if it is either a nondecreasing sequence or a nonincreasing sequence.

Note that it is the sequence itself that is nondecreasing (etc.), not the individual functions f_n. (Indeed, if X is a general set, it does not make sense to say an individual function from X to $\mathbb{R}$ is nondecreasing.) This notion of nondecreasing (etc.) is sometimes called *pointwise nondecreasing* (etc.) One can similarly define strictly increasing, strictly decreasing, and strictly monotone sequences of functions.

Here is the general version of Dini's Theorem:

XIII.11.15.2. Theorem. [Dini's Theorem, General Version] Let X be a compact topological space, and (f_n) a nondecreasing sequence of real-valued continuous functions on X converging pointwise to a function f (not assumed to be continuous). If g is a real-valued continuous function on X satisfying $g(x) < f(x)$ for all $x \in X$, then there is an n such that $g(x) < f_n(x)$ for all $x \in X$. The same statement holds with inequalities reversed if (f_n) is nonincreasing.

Proof: For each n, set $K_n = \{x \in X : g(x) \geq f_n(x)\}$. Then K_n is closed since g and f_n are continuous; hence K_n is compact (). Since $f_n(x) \leq f_{n+1}(x)$ for all $x \in X$, we have $K_{n+1} \subseteq K_n$ for all n. Also, $\cap_n K_n = \emptyset$, since if $x \in \cap_n K_n$, we have $g(x) \geq f_n(x)$ for all n, hence $g(x) \geq f(x)$, contrary to hypothesis. Thus (K_n) is a decreasing sequence of compact subsets of X with empty intersection, so $K_n = \emptyset$ for some n by (). The proof of the last statement is analogous. 👍

As a corollary, we get the commonly stated version of Dini's Theorem:

XIII.11.15.3. Corollary. [Dini's Theorem, Common Version] Let X be a compact topological space, and (f_n) a monotone sequence of real-valued continuous functions on X converging pointwise to a continuous function f. Then $f_n \to f$ uniformly.

Proof: Let $\epsilon > 0$. If (f_n) is nondecreasing, set $g(x) = f(x) - \epsilon$ for all x; if (f_n) is nonincreasing, set $g(x) = f(x) + \epsilon$. 👍

Combining this result with (), we obtain:

XIII.11.15.4. Corollary. Let X be a compact topological space, and (f_n) a monotone sequence of real-valued continuous functions on X converging pointwise to a function f. Then $f_n \to f$ uniformly if and only if f is continuous.

XIII.12. Baire Spaces and the Baire Category Theorem

The Baire Category Theorem is one of the two theorems from topology (along with Tikhonov's Theorem) which are most heavily used in advanced analysis. Some of the conclusions are remarkable, and can seem like "magic" or "too good to be true."

It will be most convenient to phrase the theorem by defining the class of topological spaces, the Baire spaces, for which the conclusion holds and then showing that complete metric spaces and locally compact Hausdorff spaces are Baire spaces.

XIII.12.1. Meager Sets and Baire Spaces

XIII.12.1.1. Definition. Let X be a topological space. A subset A of X is *nowhere dense* in X if the complement of $\bar{A}$ is dense in X.

A subset of X is *meager* in X, or *first category* in X, if it is a countable union of nowhere dense subsets of X. A subset is *second category* in X if it is not meager (first category) in X. A subset A is *residual* in X, or *comeager* or *generic* in X, if A^c is meager in X.

The (nonempty) topological space X is a *Baire space* if no nonempty open subset of X is meager in X, i.e. if every residual subset of X is dense in X.

The terms "nowhere dense", "meager", "first category", "second category", "residual", "generic", and "comeager" are relative terms, and the ambient topological space X must be mentioned or at least be understood. The term "Baire space", however, is intrinsic.

The terminology "first category" and "second category" is Baire's, and is still frequently used, but we prefer the more descriptive terminology "meager" and "residual." Note that "second category" and "residual" are not the same; in a Baire space, every residual set is second category, but there are many subsets A for which both A and A^c are second category. (In a non-Baire space, even $\emptyset$ can be residual.)

XIII.12.1.2. Examples. (i) The complement of a nowhere dense set is dense, but a set with dense complement is not nowhere dense in general. For example, $\mathbb{Q}$ is not nowhere dense in $\mathbb{R}$ (in fact, it is dense). A set $A \subseteq X$ is nowhere dense in X if and only if $\bar{A}$ is nowhere dense in X.

(ii) The complement of a dense open set in a topological space X is nowhere dense in X. In particular, the Cantor set is nowhere dense in $\mathbb{R}$.

(iii) $\mathbb{R}$ is not nowhere dense in $\mathbb{R}$ (in fact, no nonempty topological space is nowhere dense in itself). But $\mathbb{R}$ is nowhere dense in $\mathbb{C}$. Thus it is important to specify the ambient space.

(iv) If X is T_1 and $x \in X$ is not an isolated point, then $\{x\}$ is nowhere dense in X. Thus a countable subset of X containing no isolated points is meager. In particular, every countable subset of $\mathbb{R}$ (e.g. $\mathbb{Q}$) is meager.

(v) Any subset of a nowhere dense set is nowhere dense. Any subset of a meager set is meager. So any subset containing a dense G_δ is residual. The converse is true in a Baire space (XIII.12.1.4.).

(vi) Any discrete space is a Baire space, since a nonempty subset of a discrete space cannot be nowhere dense. On the other hand, by (iv), no countable T_1 space without isolated points is a Baire space.

(vii) A countable union of meager sets in X is meager in X (cf. II.5.4.8.).

XIII.12.1.3. PROPOSITION. Let X be a topological space. The following are equivalent:

(i) X is a Baire space.

(ii) The intersection of any sequence of dense open subsets of X is dense.

(iii) The intersection of any finite or countable collection of dense G_δ's is a dense G_δ.

PROOF: (i) $\Rightarrow$ (ii): If X is a Baire space and (A_n) is a sequence of dense open subsets of X, then A_n^c is closed and nowhere dense in X, so $\cup_{n=1}^\infty A_n^c$ is meager and $\cap_{n=1}^\infty A_n$ is residual, hence dense.

(ii) $\Rightarrow$ (i): Suppose the intersection of any sequence of dense open subsets of X is dense, and let A be a residual subset of X. Then $A^c = \cup_{n=1}^\infty B_n$, where B_n is nowhere dense in X. If A_n is the complement of $\bar{B}_n$, then A_n is a dense open set in X, and $\cap_{n=1}^\infty A_n \subseteq A$. By assumption, $\cap_{n=1}^\infty A_n$ is dense in X, so A is also dense in X.

For (iii), the only issue is whether the intersection is dense since in any topological space an intersection of countably many G_δ's is a G_δ. (iii) $\Rightarrow$ (ii) is trivial since an open set is a G_δ, and (ii) $\Rightarrow$ (iii) is nearly trivial since a dense G_δ is an intersection of a sequence of dense open sets. 👍

XIII.12.1.4. COROLLARY. Let X be a Baire space. A subset A of X is residual in X if and only if A contains a dense G_δ.

PROOF: One direction was observed in XIII.12.1.2.(v). Conversely, let A be a residual subset of a Baire space X. Then $A^c = \cup_{n=1}^\infty B_n$, where B_n is nowhere dense in X. If A_n is the complement of $\bar{B}_n$, then A_n is a dense open set in X, and $\cap_{n=1}^\infty A_n$ is a dense G_δ in X which is contained in A. 👍

XIII.12.1.5. If X is a T_1 Baire space and A is a countable dense subset of X containing no isolated points, then A is meager (XIII.12.1.2.), so cannot be a G_δ since if it were A^c would be an F_σ, hence also meager (Exercise XIII.12.5.1.). In particular, $\mathbb{Q}$ is not a G_δ in $\mathbb{R}$.

XIII.12.2. The Baire Category Theorem

It turns out that many standard topological spaces are Baire spaces. The next theorem is the principal result of this section. It is named after R. BAIRE, who proved it in 1899 [] (BAIRE actually only stated it for $\mathbb{R}^n$, but essentially the same proof gives the general result), and is called the "Baire Category Theorem" because of his category terminology (note that it has nothing to do with category theory ()). The case of $\mathbb{R}$ was actually proved earlier (1897) by W. OSGOOD [Osg97], who used different terminology; OSGOOD also independently obtained the result for $\mathbb{R}^n$ [Osg00].

XIII.12.2.1. THEOREM. [BAIRE CATEGORY THEOREM] (i) Every completely metrizable space (topological space whose topology is given by a complete metric) is a Baire space.

(ii) Every locally compact Hausdorff space is a Baire space.

The first sentence is often stated, "Every complete metric space is a Baire space." But note that the property of being a Baire space is a purely topological property, not a metric property. Thus, for example, $(0,1)$ is a Baire space since it is homeomorphic to $\mathbb{R}$, even though the usual metric on $(0,1)$ is not complete. See XIII.10.5. for more detailed results about completely metrizable spaces.

PROOF: The proofs of both parts are nearly identical. In both cases, we must show that if (U_n) is a sequence of dense open subsets of X, then $\cap_{n=1}^{\infty} U_n$ is also dense in X, i.e. if V is a nonempty open set in X, then $V \cap [\cap_{n=1}^{\infty} U_n]$ is nonempty.

First suppose the topology on X is given by a complete metric ρ. Since U_1 is dense in X, $V \cap U_1$ is nonempty. Choose $x_1 \in V \cap U_1$, and let B_1 be an open ball of radius ≤ 1 centered at x_1 whose closure is contained in $V \cap U_1$ (there exists such a ball since $V \cap U_1$ is open in X). Since U_2 is dense in X, there is an $x_2 \in B_1 \cap U_2$. Let B_2 be an open ball centered at x_2 of radius $\leq 1/2$ whose closure is contained in $B_1 \cap U_2$. Inductively define points $x_n \in B_{n-1} \cap U_n$ and balls B_n centered at x_n of radius $\leq 1/n$ whose closure is contained in $B_{n-1} \cap U_n$. By construction, x_m is in the closure of B_n for all $m > n$, and thus $\rho(x_n, x_m) \leq 1/n$ for all $m > n$. Thus (x_n) is a Cauchy sequence in (X, ρ). Since ρ is complete, (x_n) converges to some $x \in X$. We have that x is in the closure of B_n for each n; thus $x \in \cap_{n=1}^{\infty} U_n$. And $x \in \bar{B}_1 \subseteq V$. Thus $V \cap [\cap_{n=1}^{\infty} U_n]$ is nonempty.

The proof of (ii) is similar, but easier. Choose $x_1 \in V \cap U_1$ as before. There is an open neighborhood B_1 of x_1 with compact closure contained in $V \cap U_1$ (). Inductively choose points x_n and open neighborhoods B_n such that the closure of B_n is compact and contained in $B_{n-1} \cap U_n$. Then $(\bar{B}_n)$ is a decreasing sequence of nonempty compact subsets of X and hence there is an $x \in \cap_{n=1}^{\infty} \bar{B}_n$. This x is in U_n for all n, and also in V. 👍

XIII.12.2.2. Note that the proof of this theorem uses the Axiom of Choice to choose the x_n. In fact, only a sequence of choices must be made, each dependent on the previous ones, so the Axiom of Dependent Choice (DC) suffices. Actually, the Baire Category Theorem is logically equivalent to DC (Exercise II.6.4.7.). But if X is separable, no form of choice is necessary to prove the Baire Category Theorem (Exercise ()). (It can be proved without any form of choice that a separable metrizable space is second countable (); however, the Countable Axiom of Choice is needed to prove that a second countable topological space, even a second countable metrizable space, is separable, so Countable AC is needed to prove the Baire Category Theorem for second countable completely metrizable spaces.)

XIII.12.2.3. There are many Baire spaces, even metrizable ones, which do not satisfy the hypotheses of XIII.12.2.1.. In particular, a metrizable Baire space need not be completely metrizable (XIII.12.5.4.).

XIII.12.2.4. One of the important ways the Baire Category Theorem is used in analysis is to prove the existence of objects with a certain (usually pathological) property by showing that the set of objects without the property is a meager subset of some complete metric space. Such proofs are often regarded as "nonconstructive" or "pure existence proofs" which prove existence without exhibiting a specific example. The claim is made in [Oxt71] that this point of view is incorrect (at least in the separable case), and that the proof of the Baire Category Theorem is an explicit construction which in applications gives a

concrete example. This claim is correct, at least in principle: the proof of the theorem does give a recipe for constructing an example. But the actual example is the limit of a Cauchy sequence which may be almost impossible to describe explicitly. Even the terms in the Cauchy sequence can be quite difficult to explicitly describe in many cases. (This point is acknowledged in [Oxt71].)

Here is an interesting sample application of Baire Category. An explicit example of a continuous function which is not differentiable anywhere is given in (). Here is a simpler argument which shows that such functions exist; in fact, in a category sense almost all continuous functions fail to be differentiable anywhere!

XIII.12.2.5. THEOREM. Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$, and $C([a, b])$ the real-valued continuous functions on $[a, b]$, with supremum norm. If D is the set of functions in $C([a, b])$ which have a (finite) derivative at at least one point of $[a, b]$, then D is a meager subset of $C([a, b])$. Hence there is a residual set of functions on $[a, b]$ which are continuous everywhere but differentiable nowhere.

This argument in a sense shows more than the one in (), and in a sense shows less: this argument shows not only that continuous nondifferentiable functions exist, but that such functions are "typical" (despite one's impression from Calculus) and form a set which is uniformly very dense. However, it is hard to identify a specific example of such a function from the Baire Category argument.

There are some rather astounding variations on this result, such as the following theorem of J. MARCINKIEWICZ (1935).

XIII.12.2.6. THEOREM. Let $[a, b]$ be a closed bounded interval, and (h_n) a sequence of nonzero real numbers converging to zero. Then there is a continuous function $f : [a, b] \to \mathbb{R}$ with the following property: if $\phi : [a, b] \to \mathbb{R}$ is any measurable function, then there is a subsequence (h_{k_n}) of (h_n) such that

$$\lim_{n\to\infty} \frac{f(x + h_{k_n}) - f(x)}{h_{k_n}} = \phi(x)$$

for almost all $x \in [a, b]$. (Thus f is differentiable at most on a set of measure zero.) In fact, there is a residual set of such f.

See [Str81, p. 316-317] for an outline of the proof.

XIII.12.3. Meager Sets vs. Null Sets

Meager sets have a close formal resemblance to sets of Lebesgue measure zero. Both notions are reasonable interpretations of what it should mean to be "negligibly small." The relationship is more than just an analogy; the theme of the book [Oxt71] is the close connections between the concepts.

The two notions of smallness are not the same, however, as the next rather dramatic example shows:

XIII.12.3.1. EXAMPLE. There is a dense G_δ in $\mathbb{R}$ of Lebesgue measure 0. Thus $\mathbb{R}$ can be partitioned into a disjoint union of a meager set and a set of measure 0. This follows immediately from () and the fact that $\mathbb{Q}$ is a dense subset of $\mathbb{R}$ of Lebesgue measure 0, but we give an explicit example.

Let (r_n) be an enumeration of $\mathbb{Q}$. Let $I_{n,k}$ be the open interval of length $\frac{1}{2^n k}$ centered at r_n, and for each k let $A_k = \cup_{n=1}^{\infty} I_{n,k}$. Then A_k is open, and dense since $\mathbb{Q} \subseteq A_k$; and

$$\lambda(A_k) \leq \sum_{n=1}^{\infty} \frac{1}{2^n k} = \frac{1}{k} .$$

Thus $A = \cap_{k=1}^{\infty} A_k$ is a dense G_δ in $\mathbb{R}$ of Lebesgue measure 0.

An even more dramatic example is possible: there is a dense G_δ in $\mathbb{R}$ which not only has Lebesgue measure 0, but Hausdorff dimension 0, i.e. s-dimensional Hausdorff measure 0 for all $s > 0$. Such an example can be readily constructed by a refinement of the argument in XIII.12.3.1., but there is a natural and significant example: the Liouville numbers. See ().

XIII.12.4. The Baire Property

XIII.12.4.1. DEFINITION. Let X be a topological space. If $A, B \subseteq X$, write $A =^* B$ if $A \triangle B$ is a meager subset of X (i.e. "A and B differ by a meager set.")

A subset A of X has the *Baire property* if $A =^* U$ for some open set U in X, i.e. "A differs from an open set by a meager set."

XIII.12.4.2. It is easy to check that $=^*$ is an equivalence relation on subsets of X. Note that the Baire property is a relative property of A in X, not an intrinsic property of A. For example, every open set and every meager set in X has the Baire property. In particular, there is no direct connection between having the Baire property and being a Baire space.

XIII.12.4.3. PROPOSITION. Let X be a topological space. Then the collection of all subsets of X with the Baire property is a σ-algebra, and is the σ-algebra generated by the open sets and the meager sets. In particular, every Borel set in X has the Baire property.

PROOF: If U is open, then $U =^* \bar{U}$ since $\bar{U} \setminus U$ is nowhere dense, hence meager. Similarly, if E is closed, $E =^* E^\circ$.

If $A =^* U$ with U open, then $A^c \triangle U^c = A \triangle U$ is meager, so $A^c =^* U^c =^* (U^c)^\circ$, and A^c has the Baire property. And if $A_n =^* U_n$ for all n with each U_n open, then

$$\left(\bigcup_n A_n\right) \triangle \left(\bigcup_n U_n\right) \subseteq \bigcup_n (A_n \triangle U_n)$$

and a countable union of meager sets is meager, hence $\cup_n A_n =^* \cup_n U_n$, so $\cup_n A_n$ has the Baire property. Thus the collection $\mathcal{D}$ of subsets with the Baire property is a σ-algebra.

If $\mathcal{A}$ is the σ-algebra generated by the open sets and the meager sets, then $\mathcal{A} \subseteq \mathcal{D}$. Conversely, if $A \in \mathcal{D}$, then $A \triangle U = M$ is meager for some open set U, and $A = M \triangle U$, so $A \in \mathcal{A}$. Thus $\mathcal{A} = \mathcal{D}$.

XIII.12.4.4. Corollary. Let X be a topological space, and $A \subseteq X$. The following are equivalent:

(i) A has the Baire property.

(ii) $A = E \cup M$ with E a G_δ and M meager.

(iii) $A = F \setminus M$ with F an F_σ and M meager.

Proof: (ii) $\Rightarrow$ (i) and (iii) $\Rightarrow$ (i) by XIII.12.4.3.. For (i) $\Rightarrow$ (ii), suppose $A \triangle U = N$ for a meager set N. Then there is a meager F_σ F with $N \subseteq F$. Then $E = U \setminus F$ is a G_δ, $E \subseteq A$, and $A \setminus E \subseteq N$ is meager. For (i) $\Rightarrow$ (iii), apply (i) $\Rightarrow$ (ii) for A^c. 👍

Baire Measurable Functions

XIII.12.4.5. Definition. Let X and Y be topological spaces, and $f : X \to Y$ a function. Then f is *Baire measurable* if $f^{-1}(U)$ has the Baire property for every open set U in Y.

XIII.12.4.6. It follows from () that if f is Baire measurable, then $f^{-1}(B)$ has the Baire property for any Borel set B in Y. In particular, if f is Borel measurable, it is Baire measurable. (Of course, if f is continuous it is Borel measurable, hence Baire measurable.)

The next result is related to Luzin's Theorem (), at least in spirit. It holds for any topological spaces, but is most interesting if X is a Baire space:

XIII.12.4.7. Theorem. Let X and Y be topological spaces, and $f : X \to Y$ a Baire measurable function. Then there is a residual G_δ E in X such that $f|_E$ is continuous.

XIII.12.4.8. The set E is necessarily dense in X if X is a Baire space. Note that the conclusion is that the *restriction* of f to E is continuous, not that f is continuous at the points of E. In fact, f need not be continuous anywhere, even if X is a Baire space (e.g. $\chi_{\mathbb{Q}} : \mathbb{R} \to \mathbb{R}$; in this case we may take $E = \mathbb{J}$). ✋

In contrast, here is an important result in which actual continuity is obtained on a residual set.

XIII.12.4.9. Theorem. Let X, Y, and Z be metrizable topological spaces, and $f : X \times Y \to Z$ a function. Suppose f is separately continuous, i.e. for each $x \in X$ the function $f_x : Y \to Z$ given by $f_x(y) = f(x, y)$ is continuous, and for each $y \in Y$ the function $f^y : X \to Z$ given by $f^y(x) = f(x, y)$ is continuous. Then there is a residual G_δ E in $X \times Y$, with $E^y = \{x \in X : (x, y) \in E\}$ residual in X for each $y \in Y$, such that f is continuous (i.e. jointly continuous) at all points of E.

XIII.12.4.10. In particular, if $X \times Y$ is a metrizable Baire space, e.g. if X and Y are completely metrizable, then a separately continuous function from $X \times Y$ to any metric space Z is jointly continuous on a dense G_δ.

XIII.12.5. Exercises

XIII.12.5.1. Let X be a complete metric space, or a locally compact Hausdorff space, without isolated points. Show that no countable dense subset of X is a G_δ. (Hence $\mathbb{Q}$ is not a G_δ in $\mathbb{R}$.) In particular, X is uncountable. [The complement of a dense G_δ is meager.]

XIII.12.5.2. (a) Show that a dense G_δ in a Baire space is a Baire space in the relative topology.

(b) Prove that a compact Hausdorff space is a Baire space. [The proof is similar to and not essentially easier than the proof of the second half of XIII.12.2.1..] Conclude from (a) that a G_δ in a compact Hausdorff space is a Baire space.

(c) Deduce both statements in XIII.12.2.1. as corollaries of (b). [Use XIII.7.8.4. and **??** for completely metrizable spaces.]

The word "dense" cannot be eliminated in (a) (XIII.12.5.4.).

XIII.12.5.3. Prove the following strong converse to XIII.12.5.2.(a): If a topological space X contains a dense subspace which is a Baire space in the relative topology, then X is a Baire space. [Use XIII.2.3.9..]

XIII.12.5.4. Let $X = (\mathbb{R} \times [\mathbb{R} \setminus \{0\}]) \cup (\mathbb{Q} \times \{0\})$, with the relative topology from $\mathbb{R}^2$.

(a) Apply XIII.12.5.3. to show that X is a Baire space.

(b) X has a closed subspace, a G_δ, homeomorphic to $\mathbb{Q}$, which is not a Baire space (XIII.12.1.2.(vi)). Thus a closed subspace of a Baire space, even a closed G_δ, is not necessarily a Baire space.

(c) X is metrizable. But a closed subspace of a completely metrizable space is completely metrizable and hence a Baire space, so X is not completely metrizable.

Simpler non-Hausdorff examples of non-Baire closed subspaces of Baire spaces can be given.

XIII.12.5.5. [Boa96, p. 59–61] Use the Baire Category Theorem to show that on any closed bounded interval $[a, b]$, there is a continuous function which is not monotone on any subinterval, and in fact a residual set of such functions in $C([a, b])$. [Use subintervals with rational endpoints.] See XIII.12.5.8. for a closely related result.

XIII.12.5.6. [CSB54] Use the Baire Category Theorem to show that if f is a C^∞-function on an interval $[a, b]$, and for every $x \in [a, b]$ there is an n such that $f^{(n)}(x) = 0$, then f is a polynomial, i.e. there is an n such that $f^{(n)}(x) = 0$ for all $x \in [a, b]$.

XIII.12.5.7. [Wei76] Let $[a, b]$ be a closed bounded interval, and $\mathcal{D}'_b([a, b])$ the set of bounded functions on $[a, b]$ which are derivatives of differentiable functions (i.e. bounded functions on $[a, b]$ which have antiderivatives).

(a) Show that $\mathcal{D}'_b([a, b])$ is a closed subspace of the Banach space of bounded real-valued functions on $[a, b]$ with the supremum norm, hence also a Banach space (cf. V.8.5.7.).

(b) Show that if $f \in \mathcal{D}'_b([a, b])$, then $Z_f = \{x \in [a, b] : f(x) = 0\}$ is a G_δ (cf. V.12.1.8., ()).

(c) Let $\mathcal{D}'_0([a, b])$ be the set of all $f \in \mathcal{D}'_b([a, b])$ for which Z_f is dense in $[a, b]$. Use (b) and the Baire Category Theorem to show that $\mathcal{D}'_0([a, b])$ is a closed subspace of $\mathcal{D}'_b([a, b])$, and thus also a Banach space.

(d) There is a $g \in \mathcal{D}_0'([a,b])$ which is not identically zero (XIII.22.10.3.). By translating and scaling g, show that for any $x \in [a,b]$ and $\epsilon > 0$ there is an $h \in \mathcal{D}_0'([a,b])$ with $\|h\| < \epsilon$ and $h(x) \neq 0$.

(e) If I is a closed (nondegenerate) subinterval of $[a,b]$, let

$$E_I^+ = \{f \in \mathcal{D}_0'([a,b]) : f(x) \geq 0 \text{ for all } x \in I\} .$$

Show that E_I^+ is a closed subset of $\mathcal{D}_0'([a,b])$. Use (d) to show that E_I^+ is nowhere dense in $\mathcal{D}_0'([a,b])$. Do the same for $E_I^- = \{f \in \mathcal{D}_0'([a,b]) : f(x) \leq 0 \text{ for all } x \in I\}$.

(f) Let E be the set of all bounded derivatives of differentiable functions on $[a,b]$ which are monotone on some (nondegenerate) subinterval. Show that E is a meager subset of $\mathcal{D}_0'([a,b])$. [Let (I_n) be an enumeration of the closed subintervals of $[a,b]$ with rational endpoints. Show that $E = \cup_n (E_{I_n}^+ \cup E_{I_n}^-)$.

XIII.12.5.8. This problem is a continuation of XIII.12.5.7., using the same notation.

(a) Let $\mathcal{D}_0([a,b])$ be the set of differentiable functions on $[a,b]$ whose derivative is in $\mathcal{D}_0'([a,b])$. Show that $\mathcal{D}_0([a,b])$ is a subalgebra of $C_{\mathbb{R}}([a,b])$. (The Baire Category Theorem is needed to prove it is closed under multiplication.)

(b) Show that $\mathcal{D}_0([a,b])$ separates points of $[a,b]$. Thus by the Stone-Weierstrass Theorem (XX.10.2.2.) $\mathcal{D}_0([a,b])$ is uniformly dense in $C_{\mathbb{R}}([a,b])$.

(c) Give a direct proof of the result of (b) without using the Stone-Weierstrass Theorem, using:

Lemma. For every $\epsilon > 0$ there is an $f \in \mathcal{D}_0([0,1])$ with the following properties:

(i) $f(0) = 0$, $f(1) = 1$.

(ii) $f'(0) = f'(1) = 0$.

(iii) For every $x \in [0,1]$, $f'(x) < 1 + \epsilon$.

(iv) For every $x \in [0,1]$, $x - \epsilon < f(x) < x + \epsilon$.

(d) To prove the Lemma, let g be an element of $\mathcal{D}_0([a,b])$ for some $[a,b]$ with $1 < \sup g' < 1 + \epsilon$, and let $c, d \in [a,b]$, $c < d$, with $\frac{g(d)-g(c)}{d-c} > 1$. Replace c and d with nearby places where $g' = 0$. Rescale $[c,d]$ to $[0,1]$. Show that (iv) follows from (i), (iii), and the MVT.

(e) By translating and scaling the f of the lemma, show that any piecewise-linear continuous function on any closed bounded interval can be uniformly approximated by functions in $\mathcal{D}_0([a,b])$.

XIII.12.5.9. [HY88] A space X, even $[0,1]$ or $\mathbb{R}$, can have two subsets Y and Z which are homeomorphic in the relative topologies, such that Y is first category in X and Z is second category in X. The Y can even be nowhere dense. In this sense category is "nontopological."

(a) Let Y be the set of points in $[0,1]$ whose ternary expansion contains no 1's and does not end in a string of 0's or a string of 2's. Then Y is the Cantor set K with the endpoints of the complementary intervals removed. Since K is nowhere dense in $[0,1]$ (and in $\mathbb{R}$), Y is also nowhere dense in $[0,1]$ (and in $\mathbb{R}$), and in particular first category.

(b) Let Z be the set of points in $[0,1]$ whose binary expansion does not end in a string of 0's or a string of 1's. The complement of Z in $[0,1]$ is the set of dyadic rationals in $[0,1]$. Since $[0,1] \setminus Z$ is countable and thus first category in $[0,1]$, Z cannot be first category in $[0,1]$ (or in $\mathbb{R}$).

(c) Show that the map that converts 2's to 1's in the expansions is a homeomorphism from Y to Z with their relative topologies. (**Caution:** it is not one-to-one on K!)

(d) Show that if X is a topological space and ϕ is a homeomorphism from X to X, $Y \subseteq X$, and $Z = \phi(Y)$, then Z is first category [resp. nowhere dense, second category, residual] in X if and only if Y is. Thus in this sense all these notions are "topological."

XIII.12.5.10. (a) Show that an open set in a Baire space is a Baire space. [If W is open in X, and (U_n) is a sequence of dense open sets in W, set $V_n = U_n \cup \bar{W}^c$.]

(b) Conversely, if X is a topological space and every point of X has an open neighborhood which is a Baire space (it is enough to check that a dense set of points have this property), then X is a Baire space. ("The property of being a Baire space is a local property.")

XIII.12.5.11. (a) ([Oxt61], [FK78]) A product of two Baire spaces is not necessarily Baire, even if they are metrizable.

(b) [Oxt61] An arbitrary product of Baire spaces with countable pseudo-base (XIII.1.6.7.) is a Baire space.

XIII.12.5.12. Let X be a second countable topological space.

(a) Show that X is a disjoint union of a meager F_σ and a G_δ which is zero-dimensional in the relative topology. [If $\{U_n : n \in \mathbb{N}\}$ is a countable base for the topology, consider $\cup_n \partial U_n$.]

(b) If X is a Baire space, then X has a dense G_δ which is zero-dimensional in the relative topology.

XIII.13. Connectedness

The notion of a connected topological space, or of a connected subset of a topological space, is fundamental in topology and especially in applications in analysis.

The idea is simple. If a topological space can be divided into two disconnected pieces, each piece must be a closed set since no point of one of the sets can be a limit point of the other. Since each set is the complement of the other, the sets must be open as well as closed. Conversely, if a space can be partitioned into two subsets which are each both open and closed, these sets disconnect the space.

A second, more restrictive, notion of connectedness which is perhaps more intuitively appealing is that of *path-connectedness*, that any two points of the space can be connected by a path. Path-connectedness implies connectedness (once we know that intervals are connected), but not conversely. Both types of connectedness are useful in topology and analysis.

XIII.13.1. Connected Spaces

XIII.13.1.1. A subset U of a topological space X is *clopen* if U is both open and closed. ($X \setminus U$ is then also clopen.) In any topological space X, $\emptyset$ and X are clopen; in many interesting topological spaces, these are the only clopen subsets.

Finite unions, finite intersections, and set-theoretic differences of clopen sets are clopen.

The term "clopen" is regarded as a language abomination in some circles and is somewhat unfashionable; but the term is useful and there is no good alternative, so we will use it.

XIII.13.1.2. Definition. Let X be a topological space. A *separation* of X is a decomposition $X = U \cup V$ of X into two nonempty disjoint clopen subsets. X is *connected* if it has no separation, i.e. if the only clopen subsets of X are $\emptyset$ and X.

Connectedness is obviously a topological property of a space, preserved under homeomorphisms.

XIII.13.1.3. Examples. (i) Any space with the indiscrete topology is connected.

(ii) A space with the discrete topology and more than one point is not connected.

(iii) The Sierpiński space () is connected.

(iv) An infinite set with the finite complement topology () is connected.

(v) $\mathbb{Q}$ with the relative topology from $\mathbb{R}$ is not connected: for example, $U = \{x \in \mathbb{Q} : x < \sqrt{2}\}$ and $V = \{x \in \mathbb{Q} : x > \sqrt{2}\}$ form a separation of $\mathbb{Q}$.

Examples (i), (iii), and (iv) show that connectedness can behave strangely in non-Hausdorff spaces. The notion is most interesting for spaces with nice separation properties, although even there some of the behavior can be surprising.

By far the most important example of a connected topological space is:

XIII.13.1.4. THEOREM. $\mathbb{R}$ (with its usual topology) is connected.

PROOF: Let $\mathbb{R} = U \cup V$ be a separation, and let $a \in U$, $b \in V$. We may assume $a < b$. Set

$$Y = \{x \in [a, b] : x \in U\}$$

and set $c = sup(Y)$. We have $c < b$ since $(b - \epsilon, b] \subseteq V$ for some $\epsilon > 0$. Either $c \in U$ or $c \in V$. If $c \in U$, then $[c, c + \epsilon) \subseteq U$ for some $\epsilon > 0$, contradicting that $c = \sup(Y)$. If $c \in V$, then $(c - \epsilon, c] \subset V$ for some $\epsilon > 0$, so $c - \epsilon$ would be an upper bound for Y, another contradiction. So no separation exists. 👍

Connected Subsets of a Space

XIII.13.1.5. If X is a topological space and $E \subset X$, we say E is a connected subset of X if E is a connected space in the relative topology. This can be phrased in several equivalent ways:

XIII.13.1.6. PROPOSITION. Let X be a topological space and $E \subseteq X$. The following are equivalent:

(i) E is not connected.

(ii) There are open sets U and V in X with $U \cap E$ and $V \cap E$ nonempty, such that $E \subseteq U \cup V$ and $U \cap V \cap E = \emptyset$.

(iii) There are nonempty subsets A and B of E such that $E = A \cup B$ and $\bar{A} \cap B = A \cap \bar{B} = \emptyset$.

The proof is an easy exercise, and is left to the reader. A pair $\{U, V\}$ of open sets as in (ii) is called a *separation* of E in X (cf. XIII.10.1.1.).

XIII.13.1.7. COROLLARY. A subset of $\mathbb{R}$ is connected if and only if it is an interval or a singleton.

An interval is connected by (ii) and essentially the same argument as in the proof of XIII.13.1.4.. A subset of $\mathbb{R}$ which is not an interval or singleton can be separated as in XIII.13.1.3.(v).

XIII.13.1.8. An interesting question is whether, if A and B are subsets of a topological space X which are separated as in (iii), then they have disjoint open neighborhoods (essentially, whether the U and V in (ii) can be chosen disjoint). The answer is: not always (Exercise XIII.13.9.4.). It is not even always true if X is a compact Hausdorff space [SS95, Example 86]. But it is true if X is metrizable (XIII.10.1.1.).

Closure of a Connected Set

XIII.13.1.9. PROPOSITION. Let X be a topological space, and E a connected subset of X. Then the closure $\bar{E}$ is also connected.

PROOF: Suppose $\{U, V\}$ is a separation of $\bar{E}$ as in XIII.13.1.6.(ii). Then $U \cap E$ or $V \cap E$ is empty, for otherwise $\{U, V\}$ would be a separation of E. If $V \cap E$ is empty, then $E \subseteq U$. Thus $\bar{E} \subseteq U$, since $\bar{E} \cap U = \bar{E} \cap V^c$ is closed. Thus $\bar{E} \cap V = \emptyset$, a contradiction. 👍

XIII.13.1.10. Corollary. Let X be a topological space. If X has a dense connected subset, then X is connected. In particular, if X is a topological space, E a connected subset of X, and Y a subset of X with $E \subseteq Y \subseteq \bar{E}$, then Y is connected.

Connected Boundary

The next result will be useful in XII.8.3.2..

XIII.13.1.11. Proposition. Let X be a connected topological space and U an open set in X. If ∂U is connected, then $U^c = X \setminus U$ is connected.

Proof: Set $V = (U^c)^\circ = X \setminus \bar{U}$. Then V is open, and X is the disjoint union of U, ∂U, and V. Suppose $U^c = \partial U \cup V$ is not connected, and let E and F be disjoint nonempty relatively clopen subsets of U^c with union U^c. Since U^c is closed, E and F are closed in X. Since ∂U is connected, either $\partial U \subseteq E$ or $\partial U \subseteq F$; suppose without loss of generality $\partial U \subseteq E$. Then $F \subseteq V$, and since F is relatively open in V and V is open in X, F is open in X. Hence F is a nontrivial clopen set in X, contradicting that X is connected. 👍

XIII.13.2. Continuous Images of Connected Sets

One of the most important properties of connected sets is:

XIII.13.2.1. Proposition. Let X and Y be topological spaces and $f : X \to Y$ a continuous function. If E is a connected subset of X, then $f(E)$ is a connected subset of Y.

Informally, a continuous image of a connected space is connected.

Proof: Suppose $\{U, V\}$ is a separation of $f(E)$ as in XIII.13.1.6.(ii). Then $\{f^{-1}(U), f^{-1}(V)\}$ is a separation of E. 👍

XIII.13.2.2. Corollary. Let X be a topological space, and $f : X \to \mathbb{R}$ a continuous function. If E is a connected subset of X, then $f(E)$ is an interval or singleton. In particular, if $f : \mathbb{R} \to \mathbb{R}$ is continuous and I is an interval in $\mathbb{R}$, then $f(I)$ is an interval or singleton.

XIII.13.2.3. Corollary. A connected completely regular space with more than one point has cardinality at least $2^{\aleph_0}$.

Proof: Such a space has a nonconstant real-valued continuous function, whose range has cardinality $2^{\aleph_0}$. 👍

XIII.13.2.4. More generally, a connected Urysohn space (XIII.7.10.21.) has cardinality at least $2^{\aleph_0}$ (same proof). This corollary can fail for spaces which are not completely regular. There are countably infinite connected Hausdorff spaces [SS95, Examples 60, 75], even countable connected completely Hausdorff spaces [SS95, Example 127]. A connected regular space must be uncountable, since a countable topological space is Lindelöf, and a regular Lindelöf space is normal, hence completely regular. There is a connected regular space of cardinality $\aleph_1$ [SS95, Example 92].

The Intermediate Value Theorem

As a corollary, we get one of the most important properties of continuous functions on the real line:

XIII.13.2.5. Corollary. [Intermediate Value Theorem] Let I be an interval in $\mathbb{R}$, and $f : I \to \mathbb{R}$ a continuous function. If $a, b \in I$ and d is any number between $f(a)$ and $f(b)$, then there is at least one $c \in I$ with $f(c) = d$. (See Figure ().)

XIII.13.3. Unions and Intersections

XIII.13.3.1. A union of two or more connected subsets of a topological space X frequently fails to be connected, e.g. the union of two disjoint open sets or of two sets with disjoint closures (e.g. two singletons in a T_1 space). However:

XIII.13.3.2. Proposition. Let X be a topological space, and $\{E_i : i \in I\}$ a collection of connected subsets of X. If $\cap_{i\in I} E_i \neq \emptyset$, then $\cup_{i\in I} E_i$ is connected.

Proof: Let $E = \cup_{i\in I} E_i$, and let $p \in \cap_{i\in I} E_i$. Suppose A and B are subsets of E with $E = A \cup B$ and $\bar{A} \cap B = A \cap \bar{B} = \emptyset$, with $p \in A$. For each i set $A_i = A \cap E_i$ and $B_i = B \cap E_i$. We have $E_i = A_i \cup B_i$ and $\bar{A}_i \cap B_i = A_i \cap \bar{B}_i = \emptyset$, so by XIII.13.1.6.(iii) A_i and B_i cannot both be nonempty. But $p \in A_i$ for all i, so $B_i = \emptyset$ for all i, and hence $B = \cup_i B_i = \emptyset$. Thus E is connected by XIII.13.1.6.(iii). 👍

XIII.13.3.3. Proposition XIII.13.3.2. can be used to "string together" connected subsets: the union of a sequence of connected subsets of a space is connected provided each successive pair has nonempty intersection.

XIII.13.3.4. Corollary. Every topological vector space is connected. In particular, $\mathbb{R}^n$ is connected for any n.

Proof: Any topological vector space is the union of all lines through the origin, each of which is homeomorphic to $\mathbb{R}$ and hence connected. 👍

See XIII.13.8.3.(ii) for another proof.

Intersections

XIII.13.3.5. The intersection of two connected subsets of a topological space is not connected in general. For example, if X is a circle, the closed upper and lower semicircles are connected since they are homeomorphic to intervals (the whole circle is also connected by XIII.13.3.2.), but the intersection consists of two points with the discrete topology.

Decreasing intersections are trickier. It turns out that a decreasing sequence of connected sets, even in $\mathbb{R}^2$, does not have connected intersection in general (Exercise XIII.13.9.5.). However, we have:

XIII.13.3.6. THEOREM. Let X be a Hausdorff space, and $\{E_i : i \in I\}$ be a collection of compact subsets of X. If $\cap_{i\in F} E_i$ is connected for each finite subset F of I, then $\cap_{i\in I} E_i$ is connected. In particular, the intersection of a decreasing sequence of connected compact subsets of X is connected.

"Hausdorff" cannot be weakened to T_1 (Exercise XIII.13.9.6.).

PROOF: Set $E = \cap_{i\in I} E_i$ and $E_F = \cap_{i\in F} E_i$ for $F \subseteq I$ finite. Then $E = \cap_F E_F$ is compact. Suppose E is not connected. Then there are disjoint nonempty relatively clopen subsets A, B of E with $E = A \cup B$. Since A and B are relatively closed in E, they are compact. There are disjoint open neighborhoods U and V of A and B in X (XIII.11.4.2.). By XIII.11.1.11. there is a finite subset F of I such that $E_F \subseteq U \cup V$. Since E_F is connected, either $E_F \cap U$ or $E_F \cap V$ is empty. But $A \subseteq E_F \cap U$ and $B \subseteq E_F \cap V$, a contradiction. Thus E is connected. 👍

XIII.13.4. Components of a Space

XIII.13.4.1. DEFINITION. Let X be a topological space, and $p \in X$. The *connected component* (or just *component*) of p in X is the union of all connected subsets of X containing p. (There is always at least one such set, $\{p\}$.)

XIII.13.4.2. PROPOSITION. Let X be a topological space.

(i) The connected component of any point p of X is closed and connected, and is the largest connected subset of X containing p.

(ii) If p and q are points of X, the connected components of p and q are either identical or disjoint.

PROOF: (i): The union P of all connected subsets of X containing p is nonempty and connected by XIII.13.3.2., and is obviously the largest connected subset of X containing p. By XIII.13.1.9., $\bar{P}$ is a connected set containing p, hence $\bar{P} \subseteq P$, i.e. P is closed.

(ii): If P and Q are the components of p and q in X respectively, and $P \cap Q \neq \emptyset$, then $P \cup Q$ is connected by XIII.13.3.2. and contains p and q, hence $P \cup Q = P = Q$. 👍

XIII.13.4.3. So the connected components of a space X partition X into disjoint closed subsets. Connected components can easily contain more than one point: if X is connected, the connected component of any point of X is all of X. At the other extreme, connected components can be singletons. This obviously happens for isolated points, but can happen for nonisolated points too: the components of the Cantor set K are connected subsets of $\mathbb{R}$, hence points or intervals; since K contains no intervals, the connected components of K are single points. Intermediate possibilities also occur in many spaces.

The connected component of a point p in a space X depends on X; if Y is a subspace of X containing p, the connected component of p in Y can be smaller than the connected component in X (it can even be smaller than the intersection of Y with the connected component in X, as easy examples show).

Quasi-Components

XIII.13.4.4. If X is a topological space, $p \in X$, and U is a clopen subset of X containing p, then any connected subset of X containing p must be contained in U (otherwise $\{U, U^c\}$ would give a separation), and in particular the component P of p in X is contained in U. Thus

$$P \subseteq Q = \bigcap \{U \subseteq X \text{ clopen } : p \in U\} .$$

The set Q is sometimes called the *quasi-component* of p in X. It can be strictly larger than the component of p in X (Exercises XIII.13.9.7. and XIII.13.9.8.). However:

XIII.13.4.5. Proposition. Let X be a compact Hausdorff space. Then the quasi-component of any point is just its component.

"Hausdorff" cannot be weakened to T_1 in this result (Exercise XIII.13.9.7.).

Proof: The argument is similar to the proof of XIII.13.3.6.. Let $p \in X$, and let P and Q be the component and quasi-component of p in X. Since $P \subseteq Q$, it suffices to show that Q is connected. Suppose Q is not connected, and write $Q = A \cup B$ for disjoint nonempty relatively clopen subsets with $p \in A$. Then A and B are compact, so they have disjoint open neighborhoods U and V in X. Since Q is the intersection of all clopen subsets of X containing p and each clopen set in X is compact, by XIII.11.1.11. there are finitely many clopen subsets $W_1, \ldots, W_n$ of X containing p such that $W = W_1 \cap \cdots \cap W_n$ is contained in $U \cup V$. W is clopen, and W is the disjoint union of $W \cap U$ and $W \cap V$. $W \cap U$ is open, and $W \cap U = W \cap V^c$ is also closed, i.e. $W \cap U$ is clopen and contains p, and thus $Q \subseteq W \cap U$, contradicting that B is nonempty. Thus Q is connected. 👍

XIII.13.4.6. Quasi-components, like components, are closed, and two quasi-components are either identical or disjoint. Thus the whole space partitions into quasi-components, each of which partitions into one or more components.

XIII.13.5. Products of Connected Spaces

XIII.13.5.1. Theorem. Let $\{X_i : i \in I\}$ be a collection of topological spaces. Then $\prod_i X_i$ is connected if and only if each X_i is connected.

PROOF: Let $X = \prod_i X_i$. If X is connected, since each coordinate projection is continuous it follows from XIII.13.2.1. that each X_i is connected. Conversely, suppose each X_i is connected. Let $p = (p_i)$ and $q = (q_i)$ be points of X. Let $F = \{i_1, \ldots, i_n\}$ be a finite subset of I. The subset $E_1 = \{(x_i) : x_i = p_i \text{ for } i \neq i_1\}$ is homeomorphic to X_{i_1}, hence connected, and contains the point a_1 with $x_{i_1} = q_{i_1}$ and $x_i = p_i$ for all $i \neq i_1$. Thus a_1 is in the connected component of p. Similarly, if a_2 is the point (x_i) with $x_{i_1} = q_{i_1}$, $x_{i_2} = q_{i_2}$, and $x_i = p_i$ for $i \neq i_1, i_2$, then a_2 is in the connected component of a_1 and hence in the connected component of p. Repeating the argument finitely many times, if q_F is the point (x_i) with $x_i = q_i$ for $i \in F$ and $x_i = p_i$ for $i \notin F$, then q_F is in the connected component of p. This is true for any finite F. Any open neighborhood of q contains q_F for some finite set F, so q is in the closure of the component of p. But components are closed, so q is in the component of p in X. This is true for any p and q, so X consists of only one connected component, i.e. is connected. 👍

XIII.13.6. Totally Disconnected and Zero-Dimensional Spaces

XIII.13.6.1. DEFINITION. Let X be a topological space. X is *totally disconnected* if for every pair $\{p, q\}$ of distinct points of X, there is a clopen set $U \subseteq X$ containing p but not q. (In other words, the quasi-components of X are single points.)

A totally disconnected topological space is Hausdorff: if U is a clopen set containing p but not q, then U and $X \setminus U$ are disjoint open neighborhoods of p and q.

XIII.13.6.2. Some references define a space to be totally disconnected if its connected components are single points (what we have called "totally disconnected" is often then called "totally separated"). This definition is equivalent to the one we have given for compact Hausdorff spaces, but not in general; see Theorem XIII.13.6.7.. Definition XIII.13.6.1. appears to be the usual definition used by topologists.

XIII.13.6.3. DEFINITION. Let X be a topological space. X is *zero-dimensional* if X is T_1 and every finite open cover of X has a refinement consisting of a partition of X into clopen subsets.

More precisely, this is the specification that the covering dimension (XIII.22.3.24.) of X is zero. Some references give a different definition (that there is a base for the topology of X consisting of clopen sets, i.e. that the *small inductive dimension* of X is zero; cf. XIII.22.3.2.). The two definitions are equivalent for T_1 Lindelöf spaces (in particular, for compact Hausdorff spaces and second countable metrizable spaces), but not in general; see Theorem XIII.13.6.7.. The T_1 condition cannot be reasonably weakened to T_0 (Exercise XIII.13.9.1.).

XIII.13.6.4. Actually, if X is zero-dimensional, every finite open cover of X has a refinement consisting of a *finite* partition of X into clopen sets. Even more is true: a finite open cover $\{V_1, \ldots, V_n\}$ has a *shrinking* consisting of a partition, i.e. there is a partition $\{U_1, \ldots, U_n\}$ of X into clopen sets with $U_j \subseteq V_j$ for each j (XIII.22.3.32.).

XIII.13.6.5. PROPOSITION. Let E and F be disjoint closed subsets of a zero-dimensional space X. Then there is a clopen set $U \subseteq X$ with $E \subset U$ and $F \subseteq X \setminus U$. In particular, a zero-dimensional topological space is normal.

PROOF: Let E and F be disjoint closed subsets of a zero-dimensional space X. The open cover $\{X \setminus E, X \setminus F\}$ has a refinement consisting of a partition of the space into clopen subsets. Let U be the union of all the sets in the partition which contain points of E; any such set is disjoint from F since the partition refines $\{X \setminus E, X \setminus F\}$. Then U is an open set containing E, and $F \subseteq X \setminus U$. $X \setminus U$ is the union of the remaining sets in the partition, hence is also open, i.e. U is clopen. 👍

There is a converse: if X is T_1 and any two disjoint closed sets can be separated by a partition, i.e. there is a clopen set containing one but not the other ($Ind(X) = 0$ (XIII.22.3.9.)), then X is zero-dimensional (XIII.22.3.33.).

XIII.13.6.6. PROPOSITION. Let X be a T_0 space for which there is a base for the topology consisting of clopen sets. Then X is completely regular.

PROOF: Let F be a closed set in X, and $p \in V = X \setminus F$. Then V is an open neighborhood of p, so there is a clopen neighborhood U of p contained in V. Define $f : X \to [0, 1]$ by setting $f(x) = 1$ if $x \in U$ and $f(x) = 0$ if $x \in X \setminus U$. Then f is continuous, $f(p) = 1$, and f is identically 0 on F. 👍

XIII.13.6.7. THEOREM. Let X be a compact Hausdorff space. The following are equivalent:

(i) Each component of X is a single point.

(ii) X is totally disconnected.

(iii) There is a base for the topology of X consisting of clopen sets ($ind(X) = 0$).

(iv) X is zero-dimensional ($dim(X) = 0$).

(v) Any two disjoint closed sets in X can be separated by a partition ($Ind(X) = 0$).

If X is metrizable, these are also equivalent to

(vi) X is homeomorphic to a subset of the Cantor set.

XIII.13.6.8. For general topological spaces, we have (vi) $\Rightarrow$ (iv) $\Rightarrow$ (ii) $\Rightarrow$ (i) and (iv) $\Longleftrightarrow$ (v) (XIII.22.3.34.); (iv) $\Rightarrow$ (iii) for T_1 spaces (XIII.22.8.1.), and (iii) $\Rightarrow$ (ii) for T_0 spaces (a space with the indiscrete topology and more than one point satisfies (iii) but not (ii)). (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) for compact Hausdorff spaces, but not in general: there are G_δ subsets of $\mathbb{R}^2$ satisfying (i) but not (ii), and ones satisfying (ii) but not (iii) ([**?**, §46], [**?**], Exercise XIII.13.9.3.); thus (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (iii) do not hold even in Polish spaces. See also Exercise XIII.13.9.7.. (iii) $\Rightarrow$ (iv) for T_0 Lindelöf spaces, but not for general normal spaces [GJ76, 16M], so (iii), (iv), and (v) are equivalent for T_1 Lindelöf spaces (which are then automatically normal). (iv) does not imply (vi) for compact Hausdorff spaces: an uncountable product of two-point spaces is a counterexample (the space from Exercise () is a more interesting counterexample). (iii) $\Rightarrow$ (vi) for arbitrary second countable T_0 spaces; thus (iii), (iv), (v), and (vi) are equivalent for second countable T_1 spaces (which are then automatically metrizable).

PROOF: (i) ⇒ (ii) is XIII.13.4.5., and (ii) ⇒ (i) is XIII.13.4.4..

(iii) ⇒ (ii): We use only that X is T_0. Let p and q be points of X, and let V be an open set containing p but not q. Since the clopen sets form a base for the topology, there is a clopen set U with $p \in U \subseteq V$. Then U separates p and q.

(ii) ⇒ (iii): Let $p \in X$ and V an open set containing p. For each $q \notin V$, let U_q be a clopen set containing p but not q. Then $\{U_q^c : q \in V^c\}$ is an open cover for V^c. Since V^c is closed and hence compact, there is a finite subcover $\{U_{q_1}^c, \dots, U_{q_n}^c\}$. Then $U = U_{q_1} \cap \cdots \cap U_{q_n}$ is a clopen set containing p and contained in V. Thus the clopen sets form a base for the topology.

(iv) ⇒ (iii): We use only that X is T_1. Let $p \in X$ and V an open set containing p. Since $\{p\}$ is closed, $\{V, X \setminus p\}$ is an open cover of X, so it has a refinement consisting of a partition into clopen sets. The clopen set containing p must be contained in V. Thus the clopen sets form a base for the topology.

(iii) ⇒ (iv): We will use only that X is T_0 and Lindelöf, and satisfies (iii). Then X is completely regular by XIII.13.6.6., and in particular T_1 (actually X is automatically normal by XIII.7.6.12., but we will not need this; after the fact X will be normal by XIII.13.6.5.). Let $\mathcal{V}$ be an open cover of X (not necessarily finite). Since the clopen sets form a base for the topology, there is a refinement $\mathcal{W}$ of $\mathcal{V}$ consisting of clopen sets. $\mathcal{W}$ has a countable subcover $\{W_1, W_2, \dots\}$. Set $U_1 = W_1$, $U_2 = W_2 \setminus W_1$, $U_3 = W_3 \setminus (W_1 \cup W_2)$, etc. (Any U_n which are empty can be discarded.) Then $\{U_1, U_2, \dots\}$ is a partition of X into clopen sets, refining $\mathcal{W}$ and hence refining $\mathcal{V}$.

Actually, any partition of a compact X into clopen sets is necessarily finite since it forms an open cover of X with no proper subcover.

(iii) ⇒ (vi): We will use only that X is T_0, second countable, and satisfies (iii).
We first show that there is a countable base for the topology of X consisting of clopen sets. Let $\mathcal{V}$ be a countable base for the topology of X. For every pair (V, W) in $\mathcal{V}$ with $\bar{V} \subseteq W$, let $\mathcal{C}(V, W)$ be the set of all clopen sets U with $V \subseteq U \subseteq W$. $\mathcal{C}(V, W)$ might potentially be empty for some (V, W) (after the fact it turns out that it is always nonempty); but if it is nonempty, choose a clopen set $U_{V,W} \in \mathcal{C}(V, W)$. The collection $\mathcal{U}$ of all such $U_{V,W}$ is countable. $\mathcal{U}$ is a base for the topology of X since if p is any point of X and N any open neighborhood of p, there is a $W \in \mathcal{V}$ with $p \in W$ and $W \subseteq N$; since the clopen sets form a base, there is a clopen set U with $p \in U$ and $U \subseteq W$; there is then a $V \in \mathcal{V}$ with $p \in V$ and $V \subseteq U$. Thus $\mathcal{C}(V, W) \neq \emptyset$ and $U_{V,W}$ is defined; then $p \in V \subseteq U_{V,W} \subseteq W \subseteq N$.
Now let $\{U_1, U_2, \dots\}$ be a countable base for the topology of X consisting of clopen sets. Define a function $f : X \to \mathbb{R}$ by letting $f(x)$ be the number with base 3 expansion

$$.d_1 d_2 d_3 \cdots$$

where $d_n = 2$ if $x \in U_n$ and $d_n = 0$ if $x \in U_n^c$. f is one-to-one since X is T_0. The range of f is contained in the Cantor set K. The function f is continuous since if $x_k \to x$, then for each n, if $x \in U_n$ we have $x_k \in U_n$ for all sufficiently large k, and similarly if $x \in U_n^c$, so the n'th decimal term of x_k agrees with the n'th decimal term of x for all sufficiently large k, and hence $f(x_k) \to f(x)$. Conversely, if $(x_k), x \in X$ and $f(x_k) \to f(x)$, then for each n we have that the n'th decimal term of x_k agrees with the n'th term of x for all sufficiently large k, so if $x \in U_n$ we have $x_k \in U_n$ for sufficiently large k. Since the U_n form a base for the topology of X, we have $x_k \to x$. Thus $f^{-1} : f(X) \to X$ is continuous and f is a homeomorphism onto its image.

(vi) ⇒ (iii) since the Cantor set satisfies (iii) and (iii) passes to subspaces. 👍

The Countable AC is needed in several places in this proof.

XIII.13.6.9. There is an extensive theory of dimension for topological spaces, discussed in (), for which the above theory is the dimension zero part; the theory is based on "higher-order connectedness."

XIII.13.7. Locally Connected Spaces

XIII.13.7.1. DEFINITION. Let X be a topological space, $p \in X$. Then X is *locally connected at* p if every open neighborhood of p contains a connected open neighborhood of p, i.e. if there is a local base for the topology at p consisting of connected sets. X is *locally connected* if it is locally connected at every point, i.e. if there is a base for the topology consisting of connected sets.

XIII.13.7.2. Simple examples of locally connected spaces are $\mathbb{R}$ and any discrete space. The latter example shows that a locally connected space need not be connected. Any open set in a locally connected space is locally connected.

The Cantor set is not locally connected at any point. In fact, a nondiscrete totally disconnected space cannot be locally connected.

The topologist's sine curve () is connected, but not locally connected at points on the y-axis. There are more extreme examples: there are compact subsets of Euclidean space, e.g. the pseudo-arc () or a solenoid (), which are connected but not locally connected at any point.

XIII.13.7.3. If X is locally connected, then the components of X are open, hence clopen. The converse is not true since there are connected spaces which are not locally connected. Spaces whose components are open are precisely ones in which each point has at least one connected open neighborhood. If X is locally connected, the connected components of any open subset U of X are also open, since U is also locally connected. Conversely, if every component of every open subset of X is open, then X is locally connected.

XIII.13.8. Path-Connected Spaces

An intuitively appealing variation of connectedness is path-connectedness. A space is path-connected if any two points can be connected by a continuous path; such a space is necessarily connected. Most familiar topological spaces such as (connected) manifolds and polyhedra are path-connected.

XIII.13.8.1. DEFINITION. Let X be a topological space, $p, q \in X$. A *path* in X from p to q is a continuous function $\gamma : [0,1] \to X$ with $\gamma(0) = p$, $\gamma(1) = q$.

The space X is *path-connected* if for any $p, q \in X$, there is a path from p to q.

A path-connected space is also called *arcwise connected*. (Some references make a technical distinction between paths and arcs; cf. XIII.13.8.11..)

Since an interval is connected (XIII.13.1.7.), we obtain:

XIII.13.8.2. PROPOSITION. Let X be a topological space, $p, q \in X$. If there is a path from p to q, then p and q are in the same connected component of X. In particular, a path-connected space is connected.

PROOF: If γ is a path from p to q, then by XIII.13.2.1. the set $\gamma([0,1])$ is a connected subset of X containing both p and q. 👍

XIII.13.8.3. EXAMPLES. (i) Any connected subset of $\mathbb{R}$ (interval or point) is path-connected.

(ii) Any convex set in $\mathbb{R}^n$ (or, more generally, in any topological vector space) is path-connected: the straight line segment between any two points is a path. In particular, any topological vector space, including $\mathbb{R}^n$, is path-connected.

(iii) The topologist's sine curve () is connected but not path-connected. There are more dramatic examples of such spaces (XIII.13.8.6.).

Define a relation $\sim$ on X by $p \sim q$ if there is a path from p to q.

XIII.13.8.4. PROPOSITION. The relation $\sim$ is an equivalence relation.

PROOF: If γ is a path from p to q, set $\bar{\gamma}(t) = \gamma(1-t)$ for $0 \leq t \leq 1$; then $\bar{\gamma}$ is a path from q to p, so $\sim$ is symmetric. The constant function from $[0,1]$ to X with value p is a path from p to p, so $\sim$ is reflexive. Now suppose γ_1 is a path from p to q and γ_2 a path from q to r. Define $\gamma : [0,1] \to X$ by

$$\gamma(t) = \begin{cases} \gamma_1(2t) & \text{if} \quad 0 \leq t \leq \frac{1}{2} \\ \gamma_2(2t-1) & \text{if} \quad \frac{1}{2} \leq t \leq 1 \end{cases} .$$

Then γ is a path from p to r, so $\sim$ is transitive. 👍

XIII.13.8.5. DEFINITION. The equivalence classes of $\sim$ are the *path components* of X.

As with components, the space X partitions into a disjoint union of path components. By XIII.13.8.2., each path component is contained in a connected component of X, i.e. each connected component partitions into one or more path components.

Unlike components, path components need not be closed. For example, the topologist's sine curve () consists of two path components, one closed and the other dense and open. A solenoid () has uncountably many path components, each dense (and not open).

XIII.13.8.6. A space X is *totally path disconnected* if its path components are single points, i.e. if every continuous function from $[0,1]$ into X is constant. There are connected compact metrizable spaces which are totally path disconnected, e.g. the pseudo-arc ().

XIII.13.8.7. DEFINITION. A topological space X is *locally path-connected* at p if there is a local base for the topology at p consisting of path-connected sets, i.e. any open neighborhood of p contains a path-connected open neighborhood of p. X is *locally path-connected* if it is locally path-connected at each point, i.e. if there is a base for the topology consisting of path-connected sets.

A locally path-connected space is locally connected. There are connected and locally connected compact Hausdorff spaces which are not locally path-connected [**?**, Examples 46, 48]; however, a connected compact metrizable space which is locally connected is locally path-connected (Exercise XIII.13.9.11.; in fact, the whole space is a continuous image of $[0,1]$ by the Hahn-Mazurkiewicz Theorem; cf. [HY88, 3.29–3.30], Exercise XIII.13.9.12.). This result is false for noncompact sets (XIII.13.9.14.).

XIII.13.8.8. There are spaces which are path-connected but not locally path-connected: for example, the extended topologist's sine curve () is not locally path-connected (not even locally connected) at points on the y-axis.

XIII.13.8.9. If X is locally path-connected at p, then the path component of p in X is open. If X is locally path-connected, then all its path components are open, hence also closed since the complement of a path component is the union of the other path components. Thus in a locally path-connected space, the path components are the same as the components (and the quasi-components). In particular, if X is connected and locally path-connected, it is path-connected.

XIII.13.8.10. Since $\mathbb{R}^n$ is locally path-connected (open balls form a base for the topology of path-connected sets), any open set in $\mathbb{R}^n$ is locally path-connected. Hence the connected components of any open set in $\mathbb{R}^n$ are the same as the path components, and are open.

The same is true in any locally convex topological vector space (in particular, any normed vector space). In fact, any topological vector space, locally convex or not, is locally path-connected (there is a base for the topology consisting of balanced open sets, which are path-connected).

Path-Connectedness vs. Arcwise Connectedness

XIII.13.8.11. DEFINITION. A topological space X is *arcwise connected* if, for every $p, q \in X$, $p \neq q$, there is a homeomorphism $\gamma : [0,1] \to X$ with $\gamma(0) = p$ and $\gamma(1) = q$.

Arcwise connectedness is more restrictive than path-connectedness, which only requires that there be a continuous γ. A homeomorphic image of $[0,1]$ in X is called an *arc* in X. The Sierpiński space is path-connected but not arcwise connected, as is a set of cardinality $\geq 2^{\aleph_0}$ with the finite complement topology.

The main technical result is that path-connectedness and arcwise connectedness are the same thing for Hausdorff spaces. We give a proof since this result is not discussed in most topology texts, and where it is the proofs given tend to use unnecessarily heavy machinery. Our proof, although somewhat complicated, is quite elementary.

XIII.13.8.12. THEOREM. Let X be a Hausdorff space, $p, q \in X$, $p \neq q$. If there is a path in X from p to q, then there is an arc in X from p to q.

PROOF: Set $p_0 = p$, $p_1 = q$. Let $\gamma : [0,1] \to X$ be a path from p_0 to p_1 (i.e. $\gamma(0) = p_0$, $\gamma(1) = p_1$). We want to find an arc $\tilde{\gamma} : [0,1] \to X$ from p_0 to p_1. In fact, we will find an arc with $\tilde{\gamma}([0,1]) \subseteq \gamma([0,1])$.

(i) Replacing X by $\gamma([0,1])$, we may assume X is compact and metrizable. [The fact that $\gamma([0,1])$ is metrizable is not essential to the rest of the proof, but makes part (vi) of the argument a little cleaner. The slickest proof I know of this fact is to apply XIII.10.1.11..] Fix a metric d on X. Then γ is uniformly continuous, i.e. for every $\epsilon > 0$ there is a $\delta(\epsilon) > 0$ such that $d(\gamma(t), \gamma(s)) < \epsilon$ whenever $t, s \in [0,1]$ and $|t - s| < \delta(\epsilon)$.

(ii) We may assume $\gamma(t) \neq p_0$ and $\gamma(t) \neq p_1$ for all $t \in (0,1)$: since X is Hausdorff, there is a largest $t \in [0,1]$ for which $\gamma(t) = p_0$; call it a_0. We have $0 \leq a_0 < 1$. Then there is a smallest $t \in [a_0, 1]$ with $\gamma(t) = p_1$; call it b_0. We have $a_0 < b_0 \leq 1$. Replace γ by $\gamma|_{[a_0,b_0]}$ (rescaled) if necessary.

(iii) Set $\gamma_1 = \gamma$. Let a_1 be the smallest t such that $\gamma_1(t) = \gamma_1(s)$ for some $s \in \left[\frac{1}{2}, 1\right]$ [there is such a t since $\gamma_1\left(\left[\frac{1}{2}, 1\right]\right)$ is compact and hence closed, so $\gamma_1^{-1}\left(\gamma_1\left(\left[\frac{1}{2}, 1\right]\right)\right)$ is closed]; then $0 < a_1 \leq \frac{1}{2}$. Let b_1 be

the largest t such that $\gamma_1(b_1) = \gamma_1(a_1)$; then $\frac{1}{2} \le b_1 < 1$. Set $p_{1/2} = \gamma_1(a_1) = \gamma_1(b_1)$, and let γ_2 be $\gamma_1|_{[0,a_1]}$ rescaled to $\left[0, \frac{1}{2}\right]$ and $\gamma_1|_{[b_1,1]}$ rescaled to $\left[\frac{1}{2}, 1\right]$. We have $\gamma_2([0,1]) \subseteq \gamma_1([0,1])$. The sets $A_{11} = \gamma_2\left(\left[0, \frac{1}{2}\right]\right)$ and $A_{12} = \gamma_2\left(\left[\frac{1}{2}, 1\right]\right)$ satisfy $A_{11} \cap A_{12} = \{p_{1/2}\}$.

(iv) Repeat the construction of (iii) for the restrictions of γ_2 to $\left[0, \frac{1}{2}\right]$ and $\left[\frac{1}{2}, 1\right]$. We obtain numbers $a_{21} \in \left(0, \frac{1}{4}\right]$, $b_{21} \in \left[\frac{1}{4}, \frac{1}{2}\right)$, $p_{1/4} = \gamma_2(a_{21}) = \gamma_2(b_{21})$ and $\gamma_{31} : \left[0, \frac{1}{2}\right] \to X$ which is rescaled γ_2 on $[0, a_{21}]$ and $[b_{21}, 1/2]$, and similarly $a_{22} \in \left(\frac{1}{2}, \frac{3}{4}\right]$, $b_{22} \in \left[\frac{3}{4}, 1\right)$, $p_{3/4} = \gamma_2(a_{22}) = \gamma_2(b_{22})$ and $\gamma_{32} : \left[\frac{1}{2}, 1\right] \to X$ which is rescaled γ_2 on $[1/2, a_{21}]$ and $[b_{21}, 1]$. Put γ_{31} and γ_{32} together to form $\gamma_3 : [0,1] \to X$, a path from p_0 to p_1 with $\gamma_3(1/2) = \gamma_2(1/2) = p_{1/2}$. We have $\gamma_3\left(\left[0, \frac{1}{2}\right]\right) \subseteq \gamma_2\left(\left[0, \frac{1}{2}\right]\right)$ and $\gamma_3\left(\left[\frac{1}{2}, 1\right]\right) \subseteq \gamma_2\left(\left[\frac{1}{2}, 1\right]\right)$, so $\gamma_3([0,1]) \subseteq \gamma_2([0,1])$.

(v) Continuing inductively in this way, a sequence (γ_n) of paths from p_0 to p_1 is constructed, with $\gamma_n(k/2^m) = p_{k/2^m}$ for $m \le n-1$. We have $\gamma_{n+1}([0,1]) \subseteq \gamma_n([0,1])$, and more generally

$$\gamma_{n+1}\left(\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]\right) \subseteq \gamma_n\left(\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]\right)$$

for any $m \le n-1$ and any k. If $A_{n+1,k} = \gamma_{n+1}\left(\left[\frac{k-1}{2^{-n}}, \frac{k}{2^{-n}}\right]\right)$, $A_{n+1,k} \cap A_{n+1,k+1} = \{p_{k/2^n}\}$ and $A_{n+1,k} \cap A_{n+1,k+j} = \emptyset$ for $j > 1$.

(vi) We claim that (γ_n) converges uniformly on $[0,1]$ to a path $\tilde{\gamma}$, i.e. that (γ_n) is a uniform Cauchy sequence. Let $\epsilon > 0$, and fix N with $2^{-N+1} < \delta(\epsilon)$. For any $c \in [0,1]$, there is a k such that $\frac{k}{2^{N-1}} \le c \le \frac{k+1}{2^{N-1}}$. Then $Z = \gamma_N\left(\left[\frac{k}{2^{N-1}}, \frac{k+1}{2^{N-1}}\right]\right)$ is $\gamma([a,b])$ for some interval $[a,b]$ of length $\le 2^{-N+1} < \delta(\epsilon)$, so Z has diameter less than ϵ. If $n, m \ge N$, we have $\gamma_n(c), \gamma_m(c) \in Z$, so $d(\gamma_n(c), \gamma_m(c)) < \epsilon$. Note that N depends only on ϵ and not on c.

We have

$$\tilde{\gamma}\left(\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]\right) \subseteq \gamma_n\left(\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]\right)$$

for any n, $m \le n-1$, and k.

(vii) The path $\tilde{\gamma}$ is an arc. To show this, we need only show that $\tilde{\gamma}$ is one-to-one (since X is Hausdorff). So suppose $\tilde{\gamma}(c) = \tilde{\gamma}(d)$ for some c, d, $c < d$. Choose n, k, j with $j \ge 2$ and

$$\frac{k}{2^n} \le c < \frac{k+1}{2^n}, \quad \frac{k+j}{2^n} < d \le \frac{k+j+1}{2^n} .$$

But then

$$\tilde{\gamma}(c) \in \gamma_{n+1}\left(\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right]\right)$$

$$\tilde{\gamma}(d) \in \gamma_{n+1}\left(\left[\frac{k+j}{2^n}, \frac{k+j+1}{2^n}\right]\right)$$

and these sets are disjoint, a contradiction. 👍

XIII.13.8.13. COROLLARY. Let X be a Hausdorff space. If X is path-connected, then X is arcwise connected.

XIII.13.9. Exercises

XIII.13.9.1. Let X be the Sierpiński two-point space (). Every open cover of X has a refinement consisting of clopen sets, namely $\{X\}$, but there is no base for the topology consisting of clopen sets. X is T_0, but not T_1 and hence not zero-dimensional by Definition XIII.13.6.3.; in fact, it is connected, and even path-connected: the function from $[0,1]$ to X sending $[0,1)$ to a and 1 to b is continuous.

XIII.13.9.2. (Erdős) Let X be the set of points in ℓ^2 () with rational coordinates. Show that X is totally disconnected, but there is no base for the topology consisting of clopen sets (in fact, no clopen subset is bounded.)

XIII.13.9.3. [?, §27,VI] Let K be the Cantor set (). Define a function $f : K \to \mathbb{R}$ as follows. If $t \in K$, then t can be uniquely written as a finite or infinite sum

$$t = \sum_k \frac{2}{3^{n_k}}$$

(where $t = 0$ is the "empty sum"), i.e. n_k is the location of the k'th 2 in the base 3 decimal expansion of t. Then set

$$f(t) = \sum_k \frac{(-1)^{n_k}}{2^k}$$

(with $f(0) = 0$). Let X be the graph of f, i.e. $X = \{(t, f(t)) : t \in K\} \subseteq \mathbb{R}^2$.

(a) Show that f is not continuous. Find the points of continuity of f.

(b) Show that X is totally disconnected.

(c) Show that there is no base for the topology of X consisting of clopen sets.

(d) Show that there is a point $p \in \mathbb{R}^2$ such that $X \cup \{p\}$ is not totally disconnected, although its connected components are single points.

(e) Show that X and $X \cup \{p\}$ are G_δ subsets of $\mathbb{R}^2$. [f is a pointwise limit of a sequence of continuous functions.]

XIII.13.9.4. Topologize $X = \{a, b, c\}$ by taking the open sets to be $\{\emptyset, \{a\}, \{a, b\}, \{a, c\}, X\}$. Then X is connected. The relative topology on $E = \{b, c\}$ is the discrete topology and E is not connected. However, there is no separation $\{U, V\}$ of E with U and V disjoint.

XIII.13.9.5. Let X be the subset $G \cup E$ of the plane, where $G = \{(x, y) : 0 < x \leq \frac{1}{\pi}, y = \sin\left(\frac{1}{x}\right)\}$ and $E = \{(0, 1), (0, -1)\}$. Since G is homeomorphic to $(0, 1]$, it is connected. G is dense in X, so X is connected. Let $E_n = E \cup \{(x, y) : 0 < x \leq \frac{1}{n\pi}, y = \sin\left(\frac{1}{x}\right)\}$. Then E_n is connected for all n, and $E = \cap_n E_n$, but E is not connected. Note that E is compact, although the E_n are not compact.

XIII.13.9.6. (a) Give $\mathbb{N}$ the finite complement topology. Then $\mathbb{N}$ is compact and T_1, and every infinite subset of $\mathbb{N}$ is connected. Set $E = \{1, 2\}$, and $E_n = \{1, 2, n + 1, n + 2, \dots\}$. Then E_n is compact and connected for all n, but $E = \cap_n E_n$ is not connected. (E_n is not, however, closed in $\mathbb{N}$.)

(b) Does the conclusion of XIII.13.3.6. hold for arbitrary X if the E_i are assumed to be *closed* compact subsets?

XIII.13.9.7. Let X be $\mathbb{N}$ with two points a, b added. Topologize X by taking the open subsets to be all subsets with finite complement and all subsets of $\mathbb{N}$.

(a) Show that X is a compact T_1 space.

(b) Show that the components of X are single points.

(c) Show that the quasi-component of a and b is $\{a, b\}$, i.e. there is no clopen set containing a but not b.

XIII.13.9.8. Let X be the subset of $\mathbb{R}^2$ consisting of $\left\{\left(\frac{1}{n}, t\right) : n \in \mathbb{N}, 0 \leq t \leq 1\right\}$ along with $(0,0)$ and $(0,1)$.

(a) Show that the components of X are $\{(0,0)\}$, $\{(0,1)\}$, and $\left\{\left(\frac{1}{n}, t\right) : 0 \leq t \leq 1\right\}$ for fixed n.

(b) Show that $(0,0)$ and $(0,1)$ are in the same quasi-component of X.

XIII.13.9.9. Let X be the subset of $\mathbb{R}^2$ obtained by beginning with the line segment between $(0,0)$ and $(1,0)$ and adding the closed line segments between $\left(\frac{1}{n}, 0\right)$ and $\left(\frac{1}{n+1}, \frac{1}{n(n+m)}\right)$ for all $m, n \in \mathbb{N}$. X is called the *infinite broom*. Is X locally connected at $(0,0)$? Note that if our usual convention that neighborhoods be open is relaxed, every neighborhood of $(0,0)$ in X contains a connected neighborhood. (This is false for points $(t,0)$ with $0 < t < 1$). See [HY88, p. 113]. (Note that this X is quite different from the ones in XIII.19.5.1..)

XIII.13.9.10. THEOREM. Let $Y \subseteq \mathbb{R}^n$ $(n > 1)$ be an absolute retract (e.g. Y is homeomorphic to a closed ball of some dimension; cf. ()). Then $\mathbb{R}^n \setminus Y$ is connected.

PROOF: Suppose $\mathbb{R}^n \setminus Y$ is not connected, i.e. it has a nonempty bounded component U. Let $p \in U$, and let $X = Y \cup U$. Then X is compact since $\partial U \subseteq Y$. The identity map on Y extends to a continuous function f from X onto Y (i.e. f is a retraction from X onto Y).

Let B be a closed ball centered at p large enough that X is contained in its interior. Let g be radial retraction from $B \setminus \{p\}$ to $S = \partial B$. Define $r : B \to S$ by

$$r(\boldsymbol{x}) = \begin{cases} g(f(\boldsymbol{x})) & \text{if} \quad \boldsymbol{x} \in X \\ g(\boldsymbol{x}) & \text{if} \quad \boldsymbol{x} \in B \setminus X \end{cases} .$$

Then r is a retraction from B onto S, contradicting the No-Retraction Theorem (). 👍

XIII.13.9.11. [HY88, Theorem 3.17] Show that every locally connected completely metrizable space is locally path-connected. Every open set in a locally connected Polish space is locally path-connected.

XIII.13.9.12. Let X be a compact space, Y a (necessarily compact) Hausdorff space, and f a continuous function from X onto Y. If X is locally connected, show that Y is also locally connected. [If U is an open set in Y, and C is a component of U, show that $f^{-1}(C)$ is a union of components of $f^{-1}(U)$ and hence is open in X (XIII.13.7.3.). Use that $X \setminus f^{-1}(C)$ is compact to conclude that C is open in Y.]

XIII.13.9.13. Let X be a connected locally compact Hausdorff space. Let p be a point of X and U an open neighborhood of p in X. Show that there is a nontrivial subcontinuum of X (i.e. a connected compact subset containing more than one point) containing p and contained in U. In fact, if V is an open neighborhood of p with compact closure, show that the connected component of p in $\bar{V}$ contains a point of ∂V. [Show that the quasi-component of p in $\bar{V}$ contains a point of ∂V, and apply XIII.13.4.5..] This is a weak form of local connectedness.

The statements can fail if X is not locally compact (XIII.13.9.5..)

A point in a connected locally compact metrizable space need not have a connected compact neighborhood: consider the union of the graph of $y = \frac{1}{x} \sin \frac{1}{x}$ $(x > 0)$ and the y-axis in $\mathbb{R}^2$.

XIII.13.9.14. (F. Bernstein [Ber07]; cf. [HY88, p. 110]) (a) Show that there are exactly $2^{\aleph_0}$ nontrivial closed connected subsets of $\mathbb{R}^2$. Note that each has cardinality $2^{\aleph_0}$.

(b) (AC) Well-order these nontrivial closed connected subsets $\{X_\alpha : \alpha < \sigma\}$ according to the first ordinal σ of cardinality $2^{\aleph_0}$. Inductively choose distinct points a_α, b_α in X_α, for each $\alpha < \sigma$ (i.e. a_α, b_α, a_β, b_β are all distinct for $\alpha \neq \beta$).

(c) Let $A = \{a_\alpha : \alpha < \sigma\}$ and $B = \mathbb{R}^2 \setminus A$. Then $b_\alpha \in B$ for all α.

(d) Show that A and B are connected and locally connected.

(e) Show that neither A nor B contains a subcontinuum of more than one point. In particular, they are totally path disconnected.

Thus $\mathbb{R}^2$ is a disjoint union of two connected, locally connected subsets. Compare with XVII.4.10.12..

XIII.13.9.15. Let X be a compact Hausdorff space. Define an equivalence relation $\sim$ on X by $x \sim y$ if x and y are in the same connected component of X. Let $X_c = X/\sim$ with the quotient topology; X_c is called the *component space* of X.

(a) Show that the relation $\sim$ is closed. [Use XIII.13.4.5..]

(b) Show that the decomposition is usc (XIII.8.2.1.). [If E is closed in X and x is not in the saturation of E, show there is a clopen set in X containing x disjoint from E.]

(c) Show that X_c is a compact Hausdorff space. [Use XIII.8.3.9..]

(d) Show that X_c is totally disconnected.

(e) Is the decomposition necessarily lsc?

(f) Extend the results to the case where X is locally compact and normal. [Show that X_c embeds in $(\beta X)_c$.]

(g) To what extent does this construction work if X is not locally compact?

(h) Does a similar construction work for path components?

XIII.13.9.16. **Topological Characterization of $\mathbb{Q}$.** (a) Show that every countable, first countable (hence second countable) regular space is zero-dimensional. [Such a space is normal. Consider continuous functions to $\mathbb{R}$ separating points.]

(b) Show that $\mathbb{Q} \cap [0,1]$ and $\mathbb{Q} \cap (0,1]$ are homeomorphic to $\mathbb{Q}$. [For $\mathbb{Q} \cap (0,1]$, let (c_n) be an increasing sequence of irrational numbers converging to 1, and "reverse" $\mathbb{Q} \cap (c_{2n}, c_{2n+1})$ to the right of 1. Consider the order topology, and use II.5.4.11..]

(c) Show that every countable, first countable regular space X without isolated points is homeomorphic to $\mathbb{Q}$. [Define a dense ordering on X by repeated bisection, using (a), and show the topology on X is homeomorphic to the order topology. Use II.5.4.11..] In particular, $\mathbb{Q}^n$ is homeomorphic to $\mathbb{Q}$ for any $n \in \mathbb{N}$.

(d) Show that every separable, first countable regular space without isolated points contains a dense subspace homeomorphic to $\mathbb{Q}$; in fact, every countable dense subspace with the relative topology is homeomorphic to $\mathbb{Q}$.

(e) A space, even a compact Hausdorff space, containing a dense subspace homeomorphic to $\mathbb{Q}$ is necessarily separable, but need not be first countable. [Consider $\beta\mathbb{Q}$ or $\beta\mathbb{R}$.]

XIII.13.9.17. A space X which is connected and locally path-connected is path-connected. [The path components are clopen sets.]

XIII.14. Direct and Inverse Limits

Direct and, especially, inverse limits are important topological constructions. Actually, they have little to do with topology specifically; they can be defined for systems of sets, or more generally within a category. But they are especially useful in topology.

Caution: The use of the term "limit" in connection with direct and inverse limits has almost nothing to do (at least on the surface) with the usual use of the term "limit" in topology, e.g. "limit of a sequence."

XIII.14.1. Direct Systems and Direct Limits of Sets

XIII.14.1.1. Definition. A *direct system of sets* over $\mathbb{N}$ is a sequence (X_n) of sets and, for each n, a function $\phi_{n,n+1} : X_n \to X_{n+1}$.

Pictorially, we have

$$X_1 \xrightarrow{\phi_{1,2}} X_2 \xrightarrow{\phi_{2,3}} X_3 \xrightarrow{\phi_{3,4}} \cdots$$

Both the sets X_n and the maps $\phi_{n,n+1}$ must be specified. Usually a direct system of sets over $\mathbb{N}$ is denoted $(X_n, \phi_{n,n+1})$.

If $n < m$, then there is an induced map $\phi_{n,m}$ from X_n to X_m defined by composition:

$$\phi_{n,m} = \phi_{m-1,m} \circ \phi_{m-2,m-1} \circ \cdots \circ \phi_{n+1,n+2} \circ \phi_{n,n+1} \ .$$

If $n < m < k$, then $\phi_{n,k} = \phi_{m,k} \circ \phi_{n,m}$.

We can more generally define a direct system over any directed set I ():

XIII.14.1.2. Definition. Let I be a (properly) directed set. A *direct system of sets* over I is an indexed collection $\{X_i : i \in I\}$ of sets and, for each i, j with $i < j$, a function $\phi_{i,j} : X_i \to X_j$ satisfying $\phi_{i,k} = \phi_{j,k} \circ \phi_{i,j}$ whenever $i < j < k$.

We usually write $(X_i, \phi_{i,j})$ or $(X_i, \phi_{i,j})_{i \in I}$ for a direct system of sets over I.

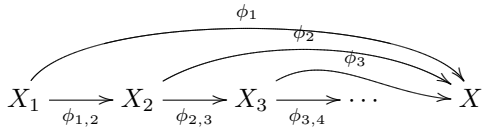
If working in a category such as the category of topological spaces, the sets are objects in the category, e.g. topological spaces in the category of topological spaces, and the functions are required to be morphisms in the category, e.g. continuous functions in the category of topological spaces.

XIII.14.1.3. We want to associate to a direct system a set X which "sits at $+\infty$". To see how this works, first consider the case where $I = \mathbb{N}$ and each $\phi_{n,n+1}$ is injective. Then X_n can be identified with the subset $\phi_{n,n+1}(X_n)$ of X_{n+1} for each n, and the direct limit X is then the "union" of the X_n. (It is not truly a union except in the rare case where each $\phi_{n,n+1}$ is an inclusion map of a subset.)

If the $\phi_{n,n+1}$ are not all injective, things are a little more complicated: X is not the "union" of the X_n since any elements which are identified (i.e. have the same image) in some X_m must be identified in X. Thus in general the direct limit X informally consists of all objects which appear at any stage, with objects identified if they are eventually identified at some stage.

Here is the formal property that the direct limit must have:

XIII.14.1.4. DEFINITION. Let $(X_i, \phi_{i,j})$ be a direct system over a directed set I. A *direct limit* for the direct system is a pair (X, ϕ_i), where X is a set and, for each i, $\phi_i : X_i \to X$ is a function such that $\phi_i = \phi_j \circ \phi_{i,j}$ for any i, j with $i < j$; pictorially, in the case $I = \mathbb{N}$, we have a commutative diagram:



with the following *universal property*:

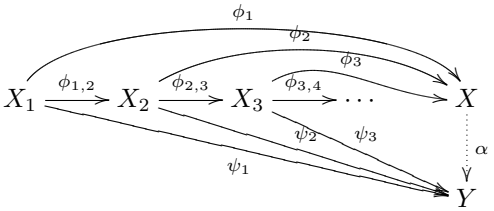
> If Y is a set and for each i there is a function $\psi_i : X_i \to Y$ such that $\psi_i = \psi_j \circ \phi_{i,j}$ for any i, j with $i < j$, then there is a unique function $\alpha : X \to Y$ such that $\psi_i = \alpha \circ \phi_i$ for all i.

We write

$$(X, \phi_i) = \varinjlim (X_i, \phi_{i,j})$$

for a direct limit. A direct limit is often called an *inductive limit*.

Pictorially, the universal property in the case $I = \mathbb{N}$ is the following commutative diagram:



XIII.14.1.5. THEOREM. Let $(X_i, \phi_{i,j})$ be a direct system of sets over a directed set I. Then

(i) There is a direct limit (X, ϕ_i).

(ii) If (X, ϕ_i) and (Y, ψ_i) are direct limits, there is a unique bijection $\alpha : X \to Y$ such that $\psi_i = \alpha \circ \phi_i$ for all i.

Thus up to natural identification there is a unique direct limit for any direct system.

PROOF: We begin with (ii). This is a simple standard argument that shows that any object with a universal property is unique up to isomorphism.

Since (X, ϕ_i) satisfies the universal property, there is a unique function $\alpha : X \to Y$ such that $\psi_i = \alpha \circ \phi_i$ for all i. Similarly, since (Y, ψ_i) satisfies the universal property, there is a unique function $\beta : Y \to X$ such that $\phi_i = \beta \circ \psi_i$. There is also a unique function $\gamma : X \to X$ such that $\phi_i = \gamma \circ \phi_i$ for all i. Since the identity map ι on X has this property, $\gamma = \iota$. But for each i we have

$$\phi_i = \beta \circ \psi_i = \beta \circ \alpha \circ \phi_i$$

so $\beta \circ \alpha$ has the property of γ and thus equals ι. By a similar argument we have that $\alpha \circ \beta$ is the identity on Y. Thus α and β are mutually inverse bijections.

We now show existence. Let Z be the subset of $[\cup_{i\in I} X_i] \times I$ defined by

$$Z = \{(x,i) : x \in X_i\} \subseteq [\cup_{i\in I} X_i] \times I$$

(Z is sometimes called the *disjoint union* of the X_i). Define a relation $\sim$ on Z by setting $(x,i) \sim (x',j)$ if and only if there is a k, $i < k$ and $j < k$, such that $\phi_{i,k}(x) = \phi_{j,k}(x')$. It is easily checked that $\sim$ is an equivalence relation. Let X be the set of equivalence classes. Write $[(x,i)]$ for the equivalence class of (x,i) in Z.

Fix i. Define a function $\phi_i : X_i \to X$ by $\phi_i(x) = [(x,i)]$ for $x \in X_i$. It is easily checked that $\phi_i = \phi_j \circ \phi_{i,j}$ for $i < j$, since $[(x,i)] = [(\phi_{i,j}(x), j)]$ for $i < j$ and $x \in X_i$.

Now suppose Y is a set and $\psi_i : X_i \to Y$ is a function for each i with $\psi_i = \psi_j \circ \phi_{i,j}$ for $i < j$. Define $\alpha : X \to Y$ by $\alpha([(x,i)]) = \psi_i(x)$ for $x \in X_i$. We need only show that this function is well defined, since then obviously $\psi_i(x) = \alpha([(x,i]]) = \alpha(\phi_i(x))$ for $x \in X_i$. If $[(x,i)] = [(x',j)]$, fix k such that $i < k$, $j < k$, and $\phi_{i,k}(x) = \phi_{j,k}(x')$. Then

$$\psi_i(x) = \psi_k(\phi_{i,k}(x)) = \psi_k(\phi_{j,k}(x')) = \psi_j(x')$$

so α is well defined.

For uniqueness of α, suppose β is another map from X to Y with $\psi_i = \beta \circ \phi_i$ for all i. Let $[(x,i)] \in X$. We have

$$\beta([(x,i)]) = \beta(\phi_i(x)) = \psi_i(x) = \alpha(\phi_i(x)) = \alpha([(x,i)])$$

and so $\beta = \alpha$. 👍

XIII.14.1.6. PROPOSITION. Let $(X, \phi_i) = \varinjlim(X_i, \phi_{i,j})$. Then

(i) For any $x \in X$, there is an i and $x_i \in X_i$ with $x = \phi_i(x_i)$.

(ii) If $x, x' \in X$, $x = \phi_i(x_i)$ and $x' = \phi_j(x'_j)$ for $x_i \in X_i$, $x'_j \in X_j$, then $x = x'$ if and only if there is a $k \in I$, $i < k$, $j < k$, such that $\phi_{i,k}(x_i) = \phi_{j,k}(x'_j)$.

The simple proof is left to the reader.

Passage to Subsystems

XIII.14.1.7. Suppose $(X_n, \phi_{n,n+1})$ is a direct system over $\mathbb{N}$. Let (k_n) be a strictly increasing sequence in $\mathbb{N}$, and let $Y_n = X_{k_n}$, $\psi_{n,n+1} = \phi_{k_n,k_{n+1}}$. If $(X, \phi_n) = \varinjlim(X_n, \phi_{n,n+1})$ and $(Y, \psi_n) = \varinjlim(Y_n, \psi_{n,n+1})$, we will show there is a natural bijection between X and Y. For each n, set $\alpha_n = \psi_n \circ \phi_{n,k_n} : X_n \to Y$; there is then a unique $\alpha : X \to Y$ such that $\alpha_n = \alpha \circ \phi_n$ for all n. If $\beta_n = \phi_{k_n} : Y_n = X_{k_n} \to X$ and $\beta : Y \to X$ is the unique function such that $\beta_n = \beta \circ \psi_n$, we have

$$\beta \circ \alpha \circ \phi_n = \beta \circ \alpha_n = \beta \circ \psi_n \circ \phi_{n,k_n} = \beta_n \circ \phi_{n,k_n} = \phi_{k_n} \circ \phi_{n,k_n} = \phi_n$$

$$\alpha \circ \beta \circ \psi_n = \alpha \circ \beta_n = \alpha \circ \phi_{k_n} = \alpha_{k_n} = \psi_{k_n} \circ \phi_{k_n,k_{k_n}} = \phi_{k_{k_n}} \circ \phi_{k_n,k_{k_n}} = \phi_{k_n} = \psi_n$$

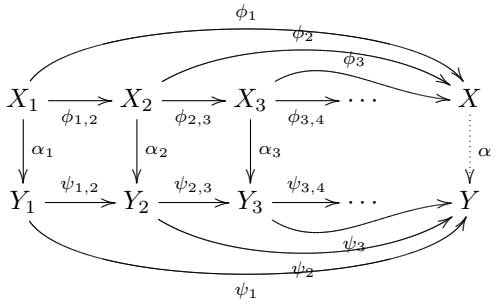
so we have $\beta \circ \alpha = \iota_X$ and $\alpha \circ \beta = \iota_Y$, i.e. α and β are mutually inverse bijections.

Thus the direct limit over a subsystem is the "same" as the direct limit over the whole system. This is an analog (in a quite different setting) of the fact that the limit of a subsequence is the same as the limit of the whole sequence (if it exists).

More generally, the same result holds for direct limits over arbitrary directed sets (XIII.14.1.10.).

Induced Functions

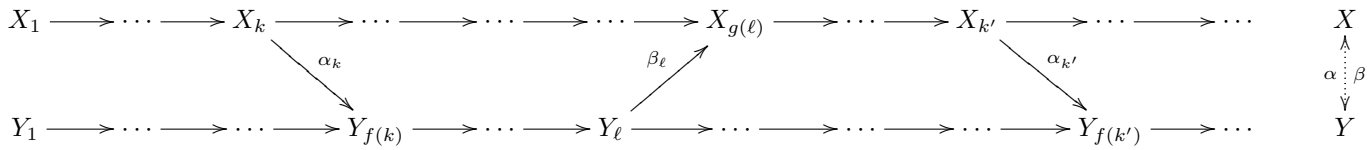
XIII.14.1.8. Let $(X_i, \phi_{i,j})$ and $(Y_i, \psi_{i,j})$ be direct systems over the same index set I. Suppose there is for each i a function $\alpha_i : X_i \to Y_i$ such that $\alpha_j \circ \phi_{i,j} = \psi_{i,j} \circ \alpha_i$ for all $i < j$. Let $(X, \phi_i) = \varinjlim(X_i, \phi_{i,j})$, $(Y, \psi_i) = \varinjlim(Y_i, \psi_{i,j})$. The maps $\psi_i \circ \alpha_i$ are a coherent family of maps from the X_i to Y, so by the universal property there is a unique function $\alpha : X \to Y$ such that $\alpha \circ \phi_i = \psi_i \circ \alpha_i$ for all i. The function α is called the *function induced by the* α_i. For $I = \mathbb{N}$ we have a commutative diagram:



XIII.14.1.9. Here is a general procedure including both XIII.14.1.7. and XIII.14.1.8.. Let I and J be directed sets, and $(X_i, \phi_{i,i'})$ and $(Y_j, \psi_{j,j'})$ be inductive systems over I and J respectively. (Note that we have slightly changed our standard notation.) Let K be a cofinal subset of I, and f a directed function () from K to J, which for simplicity we will take to be order-preserving (this assumption can be eliminated at some technical cost, in the manner of Exercise II.4.5.3.). Suppose we have a compatible set (α_k) of functions $\alpha_k : X_k \to Y_{f(k)}$ (i.e. if $k < k'$ we have $\alpha_{k'} \circ \phi_{k,k'} = \psi_{f(k),f(k')} \circ \alpha_k$). For $i \in I$, define $\theta_i : X_i \to Y$ by $\theta_i = \psi_{f(k)} \circ \alpha_k \circ \phi_{i,k}$ for any $k \in K$, $i < k$ (by compatibility this is well defined independent of the choice of k). The θ_i are a compatible family of functions from the X_i to Y, so there is a unique induced map $\alpha : X \to Y$ such that $\alpha \circ \phi_i = \theta_i$ for all i. This is the function induced by f and the α_k.

XIII.14.1.10. PROPOSITION. Let I and J be directed sets, and $(X_i, \phi_{i,i'})$ and $(Y_j, \psi_{j,j'})$ be direct systems over I and J respectively, with direct limits (X, ϕ_i) and (Y, ψ_j). Let K and L be cofinal subsets of I and J respectively, and $f : K \to J$ and $g : L \to I$ order-preserving directed maps. Let $\alpha_k : X_k \to Y_{f(k)}$ and $\beta_\ell : Y_\ell \to X_{g(\ell)}$ be compatible families of maps, with $\beta_\ell \circ \psi_{f(k),\ell} \circ \alpha_k = \phi_{k,g(\ell)}$ whenever $k \in K$, $\ell \in L$, $f(k) < \ell$ and $k < g(\ell)$, and $\alpha_k \circ \phi_{g(\ell),k} \circ \beta_\ell = \psi_{\ell,f(k)}$ whenever $g(\ell) < k$ and $\ell < f(k)$. Then the induced functions $\alpha : X \to Y$ and $\beta : Y \to X$ are mutually inverse bijections.

The proof is very similar to the proof of XIII.14.1.7., and is left to the reader. Pictorially, if $I = J = \mathbb{N}$, we have a commutative diagram:



XIII.14.1.11. Let $(X_i, \phi_{i,j})$ be a direct system of sets over I, and for each i suppose $Y_i \subseteq X_i$ and $\phi_{i,j}(Y_i) \subseteq Y_j$ whenever $i < j$. Set $\psi_{i,j} = \phi_{i,j}|_{Y_i} : Y_i \to Y_j$ for $i < j$. Then $(Y_i, \psi_{i,j})$ is a direct system of sets over I, and the inclusion maps $\beta_i : Y_i \to X_i$ induce a function $\beta : Y \to X$, where $(X, \phi_i) = \varinjlim(X_i, \phi_{i,j})$ and $(Y, \psi_i) = \varinjlim(Y_i, \psi_{i,j})$.

XIII.14.1.12. Proposition. The map $\beta : Y \to X$ is injective; thus we may write $\varinjlim(Y_i, \psi_{i,j}) \subseteq \varinjlim(X_i, \phi_{i,j})$. (Informally, direct limits of sets commute with taking subsets.)

Proof: Suppose $\beta(y) = \beta(y')$. Let $y = \psi_i(z)$, $y' = \psi_j(z')$ for some $z \in Y_i$, $z' \in Y_j$. Then there is a k with $i < k$, $j < k$, and $\phi_{i,k}(z) = \phi_{j,k}(z')$. But then $\psi_{i,k}(z) = \psi_{j,k}(z')$, so $y = \psi_i(z) = \psi_j(z') = y'$. 👍

XIII.14.2. Inverse Systems and Inverse Limits of Sets

Inverse systems and limits work similarly to direct ones, but the arrows are reversed. The details turn out to be subtly different.

XIII.14.2.1. Definition. An *inverse system of sets* over $\mathbb{N}$ is a sequence (X_n) of sets and, for each n, a function $\phi_{n+1,n} : X_{n+1} \to X_n$.

Pictorially, we have

$$\cdots \xrightarrow{\phi_{4,3}} X_3 \xrightarrow{\phi_{3,2}} X_2 \xrightarrow{\phi_{2,1}} X_1$$

If $n < m$, then there is an induced map $\phi_{m,n}$ from X_m to X_n defined by composition:

$$\phi_{m,n} = \phi_{n+1,n} \circ \phi_{n+2,n+1} \circ \cdots \circ \phi_{m-1,m-2} \circ \phi_{m,m-1} \ .$$

If $n < m < k$, then $\phi_{k,n} = \phi_{m,n} \circ \phi_{k,m}$.

We can again work over any directed set:

XIII.14.2.2. Definition. Let I be a (properly) directed set. An *inverse system of sets* over I is an indexed collection $\{X_i : i \in I\}$ of sets and, for each i, j with $i < j$, a function $\phi_{j,i} : X_j \to X_i$ satisfying $\phi_{k,i} = \phi_{j,i} \circ \phi_{k,j}$ whenever $i < j < k$.

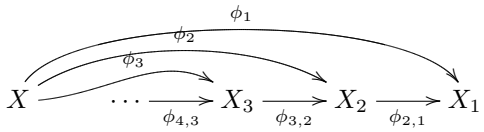
We usually write $(X_i, \phi_{j,i})$ or $(X_i, \phi_{j,i})_{i \in I}$ for an inverse system of sets over I.

XIII.14.2.3. We want to associate to an inverse system a set X which "sits at $-\infty$"; X will map to each X_i. To see how this works, first consider the case where $I = \mathbb{N}$ and each $\phi_{n+1,n}$ is injective (this is not actually a very interesting case for inverse limits). Then X_{n+1} can be identified with the subset $\phi_{n+1,n}(X_{n+1})$ of X_n for each n, and the inverse limit X is then the "intersection" of the X_n. (It is not truly an intersection except in the rare case where each $\phi_{n+1,n}$ is an inclusion map of a subset.)

If the connecting maps are surjective but not injective (the most interesting case), the inverse limit will be "larger" than the X_i, since the preimage of any element of X_i will be at least as large as the preimage under $\phi_{j,i}$, which increases in size as j increases.

Here is the universal property the inverse limit must have:

XIII.14.2.4. DEFINITION. Let $(X_i, \phi_{j,i})$ be an inverse system over a directed set I. An *inverse limit* for the inverse system is a pair (X, ϕ_i), where X is a set and, for each i, $\phi_i : X \to X_i$ is a function such that $\phi_i = \phi_{j,i} \circ \phi_j$ for any i, j with $i < j$; pictorially, in the case $I = \mathbb{N}$, we have a commutative diagram:



with the following *universal property*:

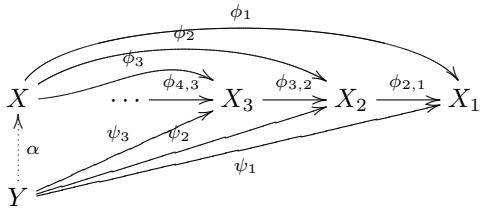
> If Y is a set and for each i there is a function $\psi_i : Y \to X_i$ such that $\psi_i = \phi_{j,i} \circ \psi_j$ for any i, j with $i < j$, then there is a unique function $\alpha : Y \to X$ such that $\psi_i = \phi_i \circ \alpha$ for all i.

We write

$$(X, \phi_i) = \varprojlim (X_i, \phi_{j,i})$$

for an inverse limit. An inverse limit is often called a *projective limit*.

Pictorially, the universal property in the case $I = \mathbb{N}$ is the following commutative diagram:



XIII.14.2.5. THEOREM. Let $(X_i, \phi_{j,i})$ be an inverse system of sets over a directed set I. Then

(i) There is an inverse limit (X, ϕ_i).

(ii) If (X, ϕ_i) and (Y, ψ_i) are inverse limits, there is a unique bijection $\alpha : Y \to X$ such that $\psi_i = \phi_i \circ \alpha$ for all i.

Thus up to natural identification there is a unique inverse limit for any inverse system.

PROOF: The proof of (ii) is essentially identical to the proof of XIII.14.1.5.(ii), and is left to the reader.

The argument for (i) is somewhat simpler than for direct limits once infinite Cartesian products have been defined (). Let Z be the Cartesian product of the X_i, and define

$$X = \{(\cdots x_i \cdots) : x_i = \phi_{j,i}(x_j) \text{ for all } i, j \in I, i < j\} \subseteq \prod_{i \in I} X_i = Z$$

(the elements of X are called *coherent threads* in Z). For each i, let ϕ_i be the restriction to X of the i'th coordinate map, i.e.

$$\phi_i((\cdots x_i \cdots)) = x_i \ .$$

It is obvious from the coherence that $\phi_i = \phi_{j,i} \circ \phi_j$ if $i < j$.

To show the universal property, let (Y, ψ_i) satisfy the coherence condition. Simply define $\alpha : Y \to X$ by

$$\alpha(y) = (\cdots \psi_i(y) \cdots) \ .$$

Then clearly $\psi_i = \phi_i \circ \alpha$ for all i, and α is the only possible function from Y to X with this property. 👍

XIII.14.2.6. PROPOSITION. Let $(X, \phi_i) = \varprojlim(X_i, \phi_{j,i})$. If $x, x' \in X$, then $x = x'$ if and only if $\phi_i(x) = \phi_i(x')$ for all i.

XIII.14.2.7. EXAMPLES. (i) Let X be a set and ι the identity map on X. Let I be a directed set. For $i \in I$ set $X_i = X$, and for $i < j$ set $\phi_{j,i} = \iota$. Then $(X_i, \phi_{j,i})$ is an inverse system over I, and (X, ι) is an inverse limit for $(X_i, \phi_{j,i})$.

(ii) Let Y be a set, and (X_n) a decreasing sequence of subsets of Y. Let $\phi_{n+1,n}$ be the inclusion map from X_{n+1} to X_n. Let $X = \cap_n X_n$ and ϕ_n the inclusion map from X to X_n. Then $(X_n, \phi_{n+1,n})$ is an inverse system over $\mathbb{N}$, and (X, ϕ_n) is an inverse limit for $(X_n, \phi_{n+1,n})$. Here $\mathbb{N}$ can be replaced by an arbitrary index set i, if $X_j \subseteq X_i$ whenever $i < j$.

(iii) As a special case of (ii), let $X_n = \{k \in \mathbb{N} : k \geq n\}$. Then $(X_n, \phi_{n+1,n})$ is an inverse system of nonempty sets with $\varprojlim(X_n, \phi_{n+1,n}) = \emptyset$.

It can even happen that the inverse limit of an inverse system of nonempty sets with surjective connecting maps is empty (Exercise XIII.14.5.5.(b)), but not if the index set is $\mathbb{N}$ (Exercise XIII.14.5.5.(a)).

We will see much more interesting examples later.

Induced Functions and Passage to Subsystems

Induced functions and passage to subsystems in inverse limits works almost identically to the direct limit case.

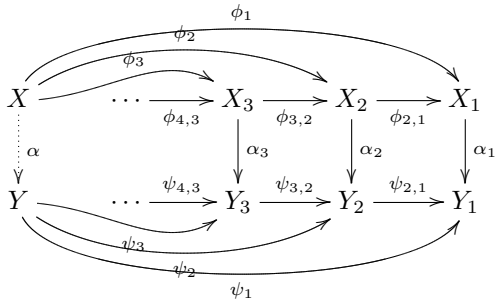
XIII.14.2.8. Suppose $(X_n, \phi_{n+1,n})$ is an inverse system over $\mathbb{N}$. Let (k_n) be a strictly increasing sequence in $\mathbb{N}$, and let $Y_n = X_{k_n}$, $\psi_{n+1,n} = \phi_{k_{n+1},k_n}$. If $(X, \phi_n) = \varprojlim(X_n, \phi_{n+,n})$ and $(Y, \psi_n) = \varprojlim(Y_n, \psi_{n,n+1})$, we will show there is a natural bijection between X and Y. For each n, set $\alpha_n = \phi_{k_n} : X \to Y_n = X_{k_n}$; there is then a unique $\alpha : X \to Y$ such that $\alpha_n = \psi_n \circ \alpha$ for all n. If $\beta_n = \phi_{k_n,n} \circ \psi_n : Y \to X_n$ and $\beta : Y \to X$ is the unique function such that $\beta_n = \phi_n \circ \beta$, we have

$$\phi_n \circ \beta \circ \alpha = \beta_n \circ \alpha = \phi_{k_n,n} \circ \psi_n \circ \alpha = \phi_{k_n,n} \circ \alpha_n = \phi_{k_n,n} \circ \phi_{k_n} = \phi_n$$

$$\psi_n \circ \alpha \circ \beta = \alpha_n \circ \beta = \phi_{k_n} \circ \beta = \beta_{k_n} = \phi_{k_{k_n},k_n} \circ \psi_{k_n} = \phi_{k_{k_n},k_n} \circ \phi_{k_{k_n}} = \phi_{k_n} = \psi_n$$

so we have $\beta \circ \alpha = \iota_X$ and $\alpha \circ \beta = \iota_Y$, i.e. α and β are mutually inverse bijections.

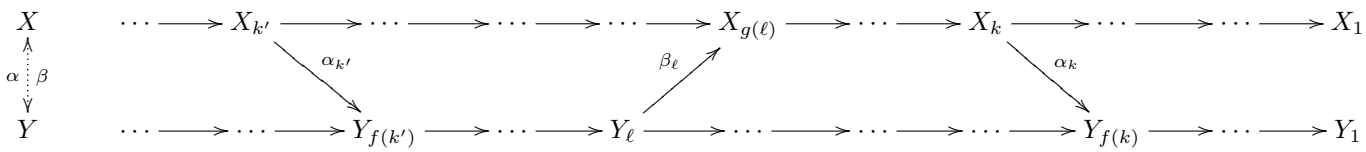
XIII.14.2.9. Let $(X_i, \phi_{j,i})$ and $(Y_i, \psi_{j,i})$ be inverse systems over the same index set I. Suppose there is for each i a function $\alpha_i : X_i \to Y_i$ such that $\psi_{j,i} \circ \alpha_j = \alpha_i \circ \phi_{j,i}$ for all $i < j$. Let $(X, \phi_i) = \varprojlim(X_i, \phi_{j,i})$, $(Y, \psi_i) = \varprojlim(Y_i, \psi_{j,i})$. The maps $\alpha_i \circ \psi_i$ are a coherent family of maps from X to the Y_i, so by the universal property there is a unique function $\alpha : X \to Y$ such that $\psi_i \circ \alpha = \alpha_i \circ \phi_i$ for all i. The function α is called the *function induced by the* α_i. For $I = \mathbb{N}$ we have a commutative diagram:



XIII.14.2.10. Here is the general procedure including both XIII.14.2.8. and XIII.14.2.9.. Let I and J be directed sets, and $(X_i, \phi_{i',i})$ and $(Y_j, \psi_{j',j})$ be inverse systems over I and J respectively. Let K be a cofinal subset of I, and f a directed function () from K to J, which for simplicity we will take to be order-preserving (this assumption can be eliminated at some technical cost, in the manner of Exercise II.4.5.3.). Suppose we have a compatible set (α_k) of functions $\alpha_k : X_k \to Y_{f(k)}$ (i.e. if $k < k'$ we have $\alpha_k \circ \phi_{k',k} = \psi_{f(k'),f(k)} \circ \alpha_{k'}$). For $j \in J$, define $\theta_j : X \to Y_j$ by $\theta_j = \psi_{f(k),j} \circ \alpha_k \circ \phi_k$ for any $k \in K$, $j < f(k)$ (by compatibility this is well defined independent of the choice of k). The θ_j are a compatible family of functions from X to the Y_j, so there is a unique induced map $\alpha : X \to Y$ such that $\psi_j \circ \alpha = \theta_j$ for all j. This is the function induced by f and the α_k.

XIII.14.2.11. Proposition. Let I and J be directed sets, and $(X_i, \phi_{i',i})$ and $(Y_j, \psi_{j',j})$ be inverse systems over I and J respectively, with inverse limits (X, ϕ_i) and (Y, ψ_j). Let K and L be cofinal subsets of I and J respectively, and $f : K \to J$ and $g : L \to I$ order-preserving directed maps. Let $\alpha_k : X_k \to Y_{f(k)}$ and $\beta_\ell : Y_\ell \to X_{g(\ell)}$ be compatible families of maps, with $\beta_\ell \circ \psi_{f(k),\ell} \circ \alpha_k = \phi_{k,g(\ell)}$ whenever $k \in K$, $\ell \in L$, $\ell < f(k)$ and $k < g(\ell)$, and $\alpha_k \circ \phi_{g(\ell),k} \circ \beta_\ell = \psi_{\ell,f(k)}$ whenever $k < g(\ell)$ and $\ell < f(k)$. Then the induced functions $\alpha : X \to Y$ and $\beta : Y \to X$ are mutually inverse bijections.

The proof is very similar to the proof of XIII.14.1.7., and is left to the reader. Pictorially, if $I = J = \mathbb{N}$, we have a commutative diagram:



XIII.14.2.12. Like direct limits, inverse limits commute with taking subspaces. Let $(X_i, \phi_{j,i})$ be an inverse system of sets over I, and for each i suppose $Y_i \subseteq X_i$ and $\phi_{j,i}(Y_j) \subseteq Y_i$ whenever $i < j$. Set

$\psi_{j,i} = \phi_{j,i}|_{Y_j} : Y_j \to Y_i$ for $i < j$. Then $(Y_i, \psi_{j,i})$ is an inverse system of sets over I, and the inclusion maps $\beta_i : Y_i \to X_i$ induce a function $\beta : Y \to X$, where $(X, \phi_i) = \varprojlim(X_i, \phi_{j,i})$ and $(Y, \psi_i) = \varprojlim(Y_i, \psi_{j,i})$. The next proposition is obvious:

XIII.14.2.13. PROPOSITION. The map $\beta : Y \to X$ is injective; thus we may write $\varprojlim(Y_i, \psi_{j,i}) \subseteq \varprojlim(X_i, \phi_{j,i})$. (Informally, inverse limits of sets commute with taking subsets.)

XIII.14.3. Direct Limits in Topology

If $(X_i, \phi_{i,j})$ is a direct system over a directed set I, and each X_i is a topological space and all $\phi_{i,j}$ are continuous, then we want to put a topology on the direct limit X so that all the maps in Definition XIII.14.1.4. are continuous. So we need each ϕ_i continuous, and for the universal property to hold the topology must be the strongest one with this property, i.e. the topology on X is induced by the functions ϕ_i in the sense of ().

XIII.14.3.1. DEFINITION. Let $(X_i, \phi_{i,j})$ be a direct system of topological spaces over a directed set I, i.e. each X_i is a topological space and each $\phi_{i,j} : X_i \to X_j$ is continuous for $i < j$. The *inductive limit topology* on $X = \varinjlim(X_i, \phi_{i,j})$ is the topology on X induced by the ϕ_i, i.e. $U \subseteq X$ is open if and only if $\phi_i^{-1}(U)$ is open in X_i for all i.

It is easily checked (cf. ()) that this is a topology on X and that each ϕ_i is continuous.

XIII.14.3.2. PROPOSITION. Let $(X_i, \phi_{i,j})$ be a direct system of topological spaces, with $(X, \phi_i) = \varinjlim(X_i, \phi_{i,j})$. Give X the inductive limit topology. If Y is a topological space and for each i there is a continuous function $\psi_i : X_i \to Y$ such that $\psi_i = \psi_j \circ \phi_{i,j}$ for any i, j with $i < j$, then the unique function $\alpha : X \to Y$ such that $\psi_i = \alpha \circ \phi_i$ for all i is continuous.

PROOF: Let V be an open set in Y, and $U = \alpha^{-1}(V)$. Then $\phi_i^{-1}(U) = \psi_i^{-1}(V)$ is open in X_i for each i. Thus U is open in X by definition of the inductive limit topology. 👍

XIII.14.3.3. There is almost nothing one can say about properties of the inductive limit topology in general; it can be quite pathological (cf. Example XIII.14.3.4.(i)). See [], [], ... for other examples. But the inductive limit topology tends to be fairly strong, as Example XIII.14.3.4.(ii) indicates.

XIII.14.3.4. EXAMPLES. (i) Let X be a set, and $\mathcal{T}_1, \mathcal{T}_2, \ldots$ a sequence of topologies on X with $\mathcal{T}_{n+1}$ weaker than $\mathcal{T}_n$ for all n. Let X_n be X with the topology $\mathcal{T}_n$ for each n, and let $\phi_{n,n+1}$ be the identity map on X. Then $\phi_{n,n+1}$ is continuous. The direct limit is again X, and the inductive limit topology is $\mathcal{T} = \cap_n \mathcal{T}_n$. If each $\mathcal{T}_n$ is T_1, then $\mathcal{T}$ is also T_1; but if each $\mathcal{T}_n$ is Hausdorff (even metrizable), then $\mathcal{T}$ is not necessarily Hausdorff ().

(ii) Here is a common situation which can be generalized in several ways. Let (X_n, ρ_n) be a metric space for each n, and $\phi_{n,n+1}$ an isometry of X_n onto a closed subset of X_{n+1}. There is an induced metric ρ on the direct limit X, since we can identify X_n with a closed subset of X, $X_n \subseteq X_{n+1}$, $X = \cup_n X_n$. It might appear that the inductive limit topology is the topology induced by ρ. This is sometimes true, e.g. if

$X_n = \left[\frac{1}{n}, 1\right]$ with the usual metric, then $X \cong (0,1]$ with the usual topology. But it is false in general. In fact, the inductive limit topology is strictly stronger then the ρ-topology in general, and is often not even first countable, so sequences are inadequate to describe the topology. An example is obtained by taking $X_n = \mathbb{R}^n$ with the usual metric (or the p-metric for $1 \leq p \leq \infty$), identifying $\mathbb{R}^n$ with $\mathbb{R}^n \times \{0\} \subseteq \mathbb{R}^{n+1}$. The direct limit can be described in this case as c_{oo}, the set of sequences of real numbers which are eventually 0. XIII.14.3.5. implies that a sequence in c_{oo} (i.e. a sequence of sequences) converges in the inductive limit topology if and only if there is a fixed n such that all the terms in each sequence vanish after the n'th place and the sequence of k'th coordinates converges for $1 \leq k \leq n$. The inductive limit topology on c_{oo} is not first countable: if $\sigma = (\epsilon_n)$ is any sequence of positive numbers, let U_σ be the set of all sequences $(x_1, x_2, \dots)$ in c_{oo} such that $|x_n| < \epsilon_n$ for all n. Then $U_\sigma \cap \mathbb{R}^n$ is open in $\mathbb{R}^n$ for all n, hence U_σ is open in the inductive limit topology. In fact, the U_σ form a local base at 0 for the inductive limit topology. But there is no countable local base at 0, since for any countable set $\{\sigma_k : k \in \mathbb{N}\}$ of such sequences, there is a σ whose terms go to 0 faster than those of any σ_k ().

In the situation of XIII.14.3.4., we have the following result for sequences (which does not generalize to nets):

XIII.14.3.5. Proposition. Let (X_n, ρ_n) be a metric space for each n, and $\phi_{n,n+1}$ an isometry of X_n onto a closed subset of X_{n+1}. Let $X = \varinjlim(X_n, \phi_{n,n+1})$. Give X the inductive limit topology $\mathcal{T}$, and identify X_n with a (closed) subset of X for each n. If (x_k) is a sequence in X, $x \in X$, then $x_k \to x$ in the $\mathcal{T}$-topology if and only if there is an n such that $x_k \in X_n$ for all k, $x \in X_n$, and $x_n \to x$ in X_n.

XIII.14.3.6. **Caution:** Suppose $(X_i, \phi_{i,j})$ is a direct system of topological spaces, and for each i we have $Y_i \subseteq X_i$, with $\phi_{i,j}(Y_i) \subseteq Y_j$ for $i < j$. Set $\psi_{i,j} = \phi_{i,j}|_{Y_i} : Y_i \to Y_j$ for $i < j$. Give Y_i the induced topology from X_i. Then $(Y_i, \psi_{i,j})$ is a direct system of topological spaces. If β_i is the inclusion map of Y_i into X_i for each i, there is an induced map $\beta : Y \to X$, where $(X, \phi_i) = \varinjlim(X_i, \phi_{i,j})$ and $(Y, \psi_i) = \varinjlim(Y_i, \psi_{i,j})$. The map β is injective by XIII.14.1.12. and continuous by XIII.14.3.2..

However, β is not a homeomorphism onto its image in general. For example (from [Dug78]), let each X_n be $[0,1]$ with its usual topology, and let $\phi_{n,n+1}$ be the identity map. Let $Y_n = \{0\} \cup \left(\frac{1}{n}, 1\right]$. Then X and Y can both be identified with $[0,1]$ with β the identity map. X has the usual topology on $[0,1]$, but Y has a strictly stronger topology, since $\{0\}$ is open in Y.

Thus direct limits of topological spaces do *not* commute with taking subspaces in general.

XIII.14.4. Inverse Limits in Topology

XIII.14.4.1. Inverse limits in topology are somewhat better behaved than direct limits. Suppose $(X_i, \phi_{j,i})$ is an inverse system over a directed set I, and each X_i is a topological space and all $\phi_{j,i}$ are continuous. Since the inverse limit (X, ϕ_i) is constructed as a subset of the Cartesian product, the obvious topology to use on X is the restriction of the product topology; this coincides with the topology defined by the maps ϕ_i (weakest topology for which all ϕ_i are continuous; cf. ()). This topology is called the *inverse limit topology*, but is rarely referred to by name since it is the only reasonable topology on the inverse limit. Another way to describe this topology is: if (x_σ) is a net in X and $x \in X$, then $x_\sigma \to x$ in X if and only if $\phi_i(x_\sigma) \to \phi_i(x)$ in X_i for all i ().

XIII.14.4.2. PROPOSITION. Let $(X_i, \phi_{j,i})$ be an inverse system of topological spaces, with $(X, \phi_i) = \varprojlim(X_i, \phi_{i,j})$. Give X the inverse limit topology. If Y is a topological space and for each i there is a continuous function $\psi_i : Y \to X_i$ such that $\psi_i = \phi_{j,i} \circ \psi_j$ for any i, j with $i < j$, then the unique function $\alpha : Y \to X$ such that $\psi_i = \phi_i \circ \alpha$ for all i is continuous.

PROOF: Since the inverse limit topology on X is generated by sets of the form $\phi_i^{-1}(U)$, where U is open, it suffices to show that $\alpha^{-1}(\phi_i^{-1}(U))$ is open for any open $U \subseteq X_i$. But $\alpha^{-1}(\phi_i^{-1}(U)) = \psi_i^{-1}(U)$, which is open since ψ_i is continuous. 👍

XIII.14.4.3. Inverse limits behave better than direct limits with respect to taking subspaces. Suppose $(X_i, \phi_{j,i})$ is an inverse system of topological spaces, and for each i we have $Y_i \subseteq X_i$, with $\phi_{j,i}(Y_j) \subseteq Y_i$ for $i < j$. Set $\psi_{j,i} = \phi_{j,i}|_{Y_j} : Y_j \to Y_i$ for $i < j$. Give Y_i the induced topology from X_i. Then $(Y_i, \psi_{j,i})$ is an inverse system of topological spaces. If β_i is the inclusion map of Y_i into X_i for each i, there is an induced map $\beta : Y \to X$, where $(X, \phi_i) = \varprojlim(X_i, \phi_{j,i})$ and $(Y, \psi_i) = \varprojlim(Y_i, \psi_{j,i})$. The map β is injective by XIII.14.2.13. and continuous by XIII.14.4.2.. In contrast to the direct limit case (XIII.14.3.6.), β is a homeomorphism onto its image in the inverse limit case, as follows immediately from the fact that if $\prod_i Y_i$ is regarded as a subset of $\prod_i X_i$, the relative topology on $\prod_i Y_i$ from the product topology on $\prod_i X_i$ is exactly the product topology on $\prod_i Y_i$ (). Thus inverse limits of topological spaces *do* commute with taking subspaces.

XIII.14.4.4. If each X_i is Hausdorff (resp. T_0, T_1, regular, completely regular), so is the inverse limit, since these separation properties are preserved under products and taking subspaces. Normality is not, however, preserved in general. An inverse limit of a sequence of metrizable spaces is metrizable.

XIII.14.4.5. PROPOSITION. Let $(X_i, \phi_{j,i})$ be an inverse system of Hausdorff spaces. Then the inverse limit is a closed subset of $\prod_i X_i$.

PROOF: Let (x_σ) be a net in $X = \varprojlim(X_i, \phi_{j,i})$ with $x_\sigma \to x \in \prod_i X_i$. Write $x_\sigma = (\cdots x_i^\sigma \cdots)$, $x = (\cdots x_i \cdots)$. Suppose $i, j \in I$, $i < j$. Then, since coordinate projections are continuous, we have $x_i^\sigma \to x_i$ and $x_j^\sigma \to x_j$. Since $\phi_{j,i}$ is continuous, $x_i^\sigma = \phi_{j,i}(x_j^\sigma) \to \phi_{j,i}(x_j)$. By uniqueness of limits in Hausdorff spaces, $x_i = \phi_{j,i}(x_j)$. Thus $x \in X$. 👍

Combining this result with Tikhonov's theorem (), we obtain:

XIII.14.4.6. COROLLARY. Let $(X_i, \phi_{j,i})$ be an inverse system of compact Hausdorff spaces. Then $\varprojlim(X_i, \phi_{j,i})$ is also a compact Hausdorff space.

XIII.14.4.7. The conclusion of XIII.14.4.5. fails in general if the X_i are not Hausdorff (e.g. if they have the indiscrete topology). But it is nonetheless true that an inverse limit of compact spaces is compact (XIII.14.5.6.).

Inverse limits of compact Hausdorff spaces behave somewhat better than general inverse limits. For example:

XIII.14.4.8. PROPOSITION. Let $(X_i, \phi_{j,i})$ be an inverse system of compact Hausdorff spaces over a directed set I, and $(X, \phi_i) = \varprojlim(X_i, \phi_{i,j})$. For each i set $Y_i = \cap_{j>i}\phi_{j,i}(X_j) \subseteq X_i$. Then for each i we have that $\phi_i(X) = Y_i$. In particular, if each X_i is nonempty, then X is nonempty.

PROOF: (AC) Note that Y_i is compact, and nonempty if all X_i are nonempty. Assume all X_i are nonempty. Fix $i_0 \in I$, and let $\mathcal{F}$ be the collection of all finite subsets of I containing i_0. Let $x_{i_0} \in Y_{i_0}$. Set $Z = \prod_i X_i$, and for each $F \in \mathcal{F}$ let Z_F be the set of all $(\cdots x_i \cdots) \in Z$ such that, whenever $i, j \in F$, $i < j$, we have $x_i = \phi_{j,i}(x_j)$, and x_{i_0} is the specified element. Then Z_F is a nonempty compact subset of Z [choose $j \in I$ such that $i < j$ for all $i \in F$, and let $x_j \in X_j$ with $\phi_{j,i_0}(x_j) = x_{i_0}$; set $x_i = \phi_{j,i}(x_j)$ for each $i \in F$]. The collection $\{Z_F : F \in \mathcal{F}\}$ has the finite intersection property, so the intersection is nonempty, and any point $x \in \cap_F Z_F$ is in the inverse limit X and satisfies $\phi_{i_0}(x) = x_{i_0}$. Thus $\phi_i(X)$ contains Y_i, and the reverse containment is trivial (this containment holds in general inverse limits). 👍

XIII.14.4.9. This result fails for general inverse limits (Exercise XIII.14.5.6.).

As a companion, it is easy to recognize inverse limits in the compact Hausdorff case:

XIII.14.4.10. PROPOSITION. Let $(X_i, \phi_{j,i})$ be an inverse system of compact Hausdorff spaces over a directed set I, and $(X, \phi_i) = \varprojlim(X_i, \phi_{i,j})$. For each i set $Y_i = \cap_{j>i}\phi_{j,i}(X_j) \subseteq X_i$. Let Y be a compact Hausdorff space, and ψ_i a compatible family of continuous functions from Y onto Y_i (i.e. $\psi_i = \phi_{j,i} \circ \psi_j$ for $i < j$ and $\psi_i(Y) = Y_i$ for all i), which separate points, i.e. if $y \neq y'$ there is an i with $\psi_i(y) \neq \psi_i(y')$. Then (Y, ψ_i) is an inverse limit for $(X_i, \phi_{j,i})$.

PROOF: Let $(X, \phi_i) = \varprojlim(X_i, \phi_{j,i})$, and $\alpha : Y \to X$ the induced map. It suffices to show that α is a bijection since it is continuous (). Since the ψ_i separate points, α is injective. To show surjectivity, let $x = (\cdots x_i \cdots) \in X$. Then $x_i \in Y_i$ for all i. For each i let $y_i \in Y$ with $\psi_i(y_i) = x_i$. Then (y_i) is a net in Y; let y be a cluster point. Since

$$\psi_i(y_j) = \phi_{j,i}(\psi_j(y_j)) = \phi_{j,i}(x_j) = x_i$$

for $i < j$, we have $\psi_i(y) = x_i$ for all i by continuity, so $\alpha(y) = x$. 👍

XIII.14.4.11. EXAMPLE. For each n let X_n be a finite set with 2^n points, with the discrete topology. We may conveniently take X_n to be the set of all n-tuples of 0's and 1's. Let $\phi_{n+1,n} : X_{n+1} \to X_n$ be a function for which the inverse image of each point of X_n consists of exactly two points. For example, we may take

$$\phi_{n+1,n}((d_1, \ldots, d_n, d_{n+1})) = (d_1, \ldots, d_n)$$

where $(d_1, \ldots, d_{n+1})$ is an $(n+1)$'tuple of 0's and 1's. The inverse image $(X, \phi_n) = \varprojlim(X_n, \phi_{n+1,n})$ is a compact Hausdorff space which is naturally homeomorphic to the Cantor set: it can be thought of as the set of all sequences of 0's and 1's, with the topology of coordinatewise convergence, with

$$\phi_n((d_1, d_2, \ldots)) = (d_1, \ldots, d_n) \ .$$

In fact it is easy to conclude from XIII.14.4.10. that the Cantor set with these maps is an inverse limit.

XIII.14.4.12. EXAMPLES. Some of the examples from () can be easily written as inverse limits of "nicer" spaces by projection.

(i) Let X be one of the spaces $X_{\boldsymbol{k}}$ of XIII.19.5.1.. Let X_n be the subset of X consisting of the horizontal line and the "whiskers" whose basepoint has x-coordinate $\geq \frac{1}{n}$. Define $\phi_{n+1,n} : X_{n+1} \to X_n$ by mapping all the points on the whiskers with basepoint $\left(\frac{1}{n+1}, 0\right)$ to the point $\left(\frac{1}{n+1}, 0\right)$. Then $\varprojlim(X_n, \phi_{n+1,n})$ is homeomorphic to X, and the map $\phi_n : X \to X_n$ just sends all the points of each whisker whose basepoint has x-coordinate $< \frac{1}{n}$ to the basepoint of the whisker.

(ii) Similarly, let X be the Hawaiian earring (XIII.19.5.8.). Let X_n be the subset of X consisting only of the first n circles (with radius $\geq \frac{1}{n}$). Define $\phi_{n+1,n} : X_{n+1} \to X_n$ by mapping all the points on the $(n+1)$'st circle to $(0,0)$. Then $\varprojlim(X_n, \phi_{n+1,n})$ is homeomorphic to X, and the map $\phi_n : X \to X_n$ just sends all the points of each circle of radius $< \frac{1}{n}$ to $(0,0)$.

(iii) Let X be the infinite broom (Exercise XIII.13.9.9.). For each n let X_n be the subset of X consisting only of the line segments from the origin to $\left(1, \frac{1}{k}\right)$ for $1 \leq k \leq n$. Define $\phi_{n+1,n} : X_{n+1} \to X_n$ by projecting the $(n+1)$'st line segment upwards onto the n'th and leaving the points of X_n fixed. Then X is naturally homeomorphic to $\varprojlim(X_n, \phi_{n+1,n})$, with ϕ_n the projection of all segments with right endpoint below $\left(1, \frac{1}{n}\right)$ upwards onto the n'th segment.

(iv) Let X be the topologist's sine curve (XIII.19.5.5.). Let X_n be the subset of $\mathbb{R}^2$ obtained by taking the part of X to the right of the vertical line $x = \frac{2}{4n+1}$ and the vertical line segment between $\left(\frac{2}{4n+1}, -1\right)$ and $\left(\frac{2}{4n+1}, 1\right)$. Then X_n is homeomorphic to $[0,1]$. Define $\phi_{n+1,n} : X_{n+1} \to X_n$ by projecting the points with x-coordinates $< \frac{2}{4n+1}$ horizontally to the right onto the line $x = \frac{2}{4n+1}$ and leaving the other points fixed. Then X is naturally homeomorphic to $\varprojlim(X_n, \phi_{n+1,n})$, with ϕ_n the projection of all points with x-coordinates $< \frac{2}{4n+1}$ horizontally to the right onto the line $x = \frac{2}{4n+1}$. Thus X is an inverse limit of copies of $[0,1]$. See XIII.14.4.18. for more examples of spaces which can be written as inverse limits of copies of $[0,1]$.

XIII.14.4.13. EXAMPLES. For $m \in \mathbb{N}$, write $\mathbb{Z}/m\mathbb{Z}$ for the cyclic group with m elements (we usually denote this group by $\mathbb{Z}_m$, but this notation will be reserved for something else here). Let $\boldsymbol{a} = (a_1, a_2, \dots)$ be a sequence in $\mathbb{N} \setminus \{1\}$ with $a_k | a_{k+1}$ for all k. Consider the inverse system

$$\cdots \xrightarrow{\pi_{4,3}} \mathbb{Z}/a_3\mathbb{Z} \xrightarrow{\pi_{3,2}} \mathbb{Z}/a_2\mathbb{Z} \xrightarrow{\pi_{2,1}} \mathbb{Z}/a_1\mathbb{Z}$$

where $\pi_{n+1,n}$ is the quotient map from $\mathbb{Z}/a_{n+1}\mathbb{Z}$ to $\mathbb{Z}/a_n\mathbb{Z}$. Let $(\mathbb{Z}_{\boldsymbol{a}}, \pi_n)$ be the inverse limit. Then $\mathbb{Z}_{\boldsymbol{a}}$ is a compact totally disconnected abelian topological group. If $a_n = p^n$ for a fixed prime p, we write $\mathbb{Z}_p$ for $\mathbb{Z}_{\boldsymbol{a}}$; $\mathbb{Z}_p$ is called the *p-adic integers* (cf. ()).

There is a natural quotient map ρ_n from $\mathbb{Z}$ to $\mathbb{Z}/a_n\mathbb{Z}$ for each n. The ρ_n define a homomorphism $\rho : \mathbb{Z} \to \mathbb{Z}_{\boldsymbol{a}}$. If m is in the kernel of ρ, then $a_n | m$ for all n, hence $m = 0$ and ρ is injective. But ρ is not surjective: in fact, $\mathbb{Z}_{\boldsymbol{a}}$ has cardinality $2^{\aleph_0}$ (it is homeomorphic to the Cantor set by ()). However, $\rho(\mathbb{Z})$ is dense in $\mathbb{Z}_{\boldsymbol{a}}$ (XIII.14.4.19.).

The group $\mathbb{Z}_{\boldsymbol{a}}$ is often, but not always, torsion-free (XIII.14.5.9.). In particular, $\mathbb{Z}_p$ is torsion-free. $\mathbb{Z}_p$ and more generally $\mathbb{Z}_{\boldsymbol{a}}$ has a natural structure as a topological ring, but this structure is not easily definable from the inverse limit construction.

There are several results about writing a general space as an inverse limit of "nice" spaces. For example:

XIII.14.4.14. THEOREM. Let X be a compact metrizable space. Then there is an inverse system $(X_n, \phi_{n+1,n})$ over $\mathbb{N}$, with each X_n a smooth compact manifold with boundary and $\phi_{n+1,n}$ a smooth map for each n, with X homeomorphic to $\varprojlim(X_n, \phi_{n+1,n})$.

PROOF: We may identify X with a (compact) subset of the Hilbert cube (). Fix a sequence (ϵ_n) with $0 < \epsilon_{n+1} < \frac{\epsilon_n}{4}$ for all n. Let Y_n be the projection of X onto the first n coordinates, and regard Y_n as a subset of $[0,1]^n \subseteq \mathbb{R}^n$. Let U_n be a neighborhood of Y_n in $\mathbb{R}^n$ as in XIII.23.10.1. with $\epsilon = \epsilon_n$, and set $X_n = \bar{U}_n$. Since the projection from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ is a contraction, it maps X_{n+1} into X_n. Let $\phi_{n+1,n}$ be the restriction of the projection map to X_{n+1}. Then $(X_n, \phi_{n+1,n})$ is an inverse system of smooth compact manifolds with boundary, with smooth connecting maps.

We show that X is homeomorphic to $\varprojlim(X_n, \phi_{n+1,n})$. For each n let ψ_n be the restriction to X of the projection from the Hilbert cube to $[0,1]^n$; then $\psi_n(X) = Y_n$ by definition of Y_n, and the conditions of XIII.14.4.10. are satisfied. 👍

If X is a compact subset of $\mathbb{R}^m$, a simplified version of the argument shows that the X_n can be taken to be m-dimensional.

XIII.14.4.15. There are two drawbacks to this result, however:

(i) The connecting maps are not surjective.

(ii) If X is finite-dimensional, the dimensions of the X_n may be much larger than $dim(X)$. (If $dim(X) = d$, X embeds in $\mathbb{R}^{2d+1}$ by the Whitney Embedding Theorem (), so the X_n can be chosen to have dimension $\leq 2d+1$.)

These difficulties cannot be avoided if one wants to write X as an inverse limit of smooth manifolds with boundary. But if we slightly more generally allow the X_n to be polyhedra (), we can do better:

XIII.14.4.16. THEOREM. Let X be a compact metrizable space. Then there is an inverse system $(X_n, \phi_{n+1,n})$ over $\mathbb{N}$, with each X_n a polyhedron and $\phi_{n+1,n}$ a surjective piecewise-linear map for each n, with X homeomorphic to $\varprojlim(X_n, \phi_{n+1,n})$. If $dim(X) = d < \infty$, then the X_n can be chosen so that $dim(X_n) = d$ for all n.

The proof is involved and uses some machinery, and will be given in (). This result was first proved by H. FREUDENTHAL in 1937 [Fre37]. The examples in XIII.14.4.11. and XIII.14.4.12. are examples of writing compact metrizable spaces as inverse limits of polyhedra.

XIII.14.4.17. EXAMPLES. **(Solenoids.)** Let $\mathbb{T}$ be the unit circle in the plane, regarded as the compact topological group of complex numbers of absolute value 1. For each n, let $X_n = \mathbb{T}$ and $\phi_{n+1,n} : X_{n+1} \to X_n$ defined by $\phi_{n+1,n}(z) = z^2$. Then $(X_n, \phi_{n+1,n})$ is an inverse system of compact Hausdorff spaces. Let $(X, \phi_n) = \varprojlim(X_n, \phi_{n,n+1})$. Then X has a natural structure as a compact Hausdorff space which is an abelian topological group (XIII.14.5.2.). In fact, it is easily seen using XIII.14.4.10. that X is homeomorphic to the solenoid (XIII.19.5.16.) of type 2^∞. Thus a solenoid is homeomorphic to an inverse limit of circles.

The circle $\mathbb{T}$ can be regarded as $\mathbb{R}/m\mathbb{Z}$ for any $m \in \mathbb{N}$. Thus we may regard the above X as

$$\varprojlim(\mathbb{R}/2^{n-1}\mathbb{Z}, \pi_{n+1,n})$$

where $\pi_{n+1,n} : \mathbb{R}/2^n\mathbb{Z} \to \mathbb{R}/2^{n-1}\mathbb{Z}$ is the natural quotient map. We have a commutative diagram

$$\begin{array}{ccccccccccc}
\cdots & \xrightarrow{\iota} & \mathbb{R} & \xrightarrow{\iota} & \cdots & \xrightarrow{\iota} & \mathbb{R} & \xrightarrow{\iota} & \mathbb{R} & \xrightarrow{\iota} & \mathbb{R} \\
 & & \downarrow{\scriptstyle \pi} & & & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} \\
\cdots & \xrightarrow{\pi_{n+1,n}} & \mathbb{R}/2^{n-1}\mathbb{Z} & \xrightarrow{\pi_{n,n-1}} & \cdots & \xrightarrow{\pi_{4,3}} & \mathbb{R}/4\mathbb{Z} & \xrightarrow{\pi_{3,2}} & \mathbb{R}/2\mathbb{Z} & \xrightarrow{\pi_{2,1}} & \mathbb{R}/\mathbb{Z}
\end{array}$$

which induces a continuous group homomorphism $\alpha : \mathbb{R} \to X$, which is injective. The range of α is dense in X by XIII.14.4.19., but is not all of X: the range consists of all threads (x_n), where $x_n \to 1$ in $\mathbb{T}$ when $\mathbb{R}/2^{n-1}\mathbb{Z}$ is identified with $\mathbb{T}$. Also, α is not a homeomorphism onto its image: for example, we have $\alpha(2^n) \to \alpha(0)$ in X. In fact, the range of α, which is a subgroup of X, is exactly the path component of the identity of X. The other path components of X are exactly the cosets of $\alpha(\mathbb{R})$ in X.

The other solenoids can be constructed by taking $\phi_{n+1,n}(z) = z^{k_n}$ for a sequence (k_n) in $\mathbb{N} \setminus \{1\}$.

XIII.14.4.18. Examples. (i) **Knaster bucket-handle.** (cf. XIII.19.5.31.) The Knaster bucket-handle can be written as an inverse limit of copies of $[0,1]$ by taking for each $\phi_{n+1,n}$ the piecewise-linear map from $[0,1]$ to $[0,1]$ with $\phi_{n+1,n}(0) = \phi_{n+1,n}(1) = 0$ and $\phi_{n+1,n}\left(\frac{1}{2}\right) = 1$:

Other bucket-handles can be constructed similarly as inverse limits by taking a sequence (k_n) of natural numbers ≥ 2 and letting $\phi_{n+1,n}$ be the sawtooth function taking values 0 and 1 alternately at the numbers $\frac{m}{k_n}$. The homeomorphism class of the bucket-handle is exactly determined by the equivalence class (XIII.19.5.24.) of the supernatural number $\prod_n k_n$ []. There is a qualitative difference between bucket-handles for which infinitely many k_n are even and ones with only finitely many even k_n (up to equivalence, ones with all k_n odd): the latter have two "endpoints" while the former have only one.

(ii) **Pseudo-Arc.** The pseudo-arc (XIII.19.5.32.) can also be written as an inverse limit of copies of $[0,1]$. The maps $\phi_{n+1,n}$ can also be taken to be piecewise-linear, but zigzag up and down similarly to the definition of a crooked embedding:

See [] for details.

XIII.14.4.19. Proposition. Let $(X_i, \phi_{j,i})$ be an inverse system of topological spaces, and $(X, \phi_i) = \varprojlim(X_i, \phi_{j,i})$. If $D \subseteq X$, then D is dense in X if and only if $\phi_i(D)$ is dense in $\phi_i(X) \subseteq X_i$ for all i.

Proof: If D is dense in X, then $\phi_i(D)$ is dense in $\phi_i(X)$ for all i by (). Conversely, suppose $\phi_i(D)$ is dense in $\phi_i(X)$ for each i. Fix $x \in X$. Let J be the set of all triples (i, U, y), where $i \in I$, U is an open neighborhood of $\phi_i(x)$ in X_i, and $y \in D$, $\phi_i(y) \in U$. Set $(i, U, y) < (i', V, y')$ if $i < i'$ and $\phi_{i',i}(V) \subseteq U$. Then J is a directed set. For $j = (i, U, y) \in J$ set $y_j = y$. Then (y_j) is a net in D, and $y_j \to x$ since $\phi_i(y_j) \to \phi_i(x)$ for all i. Thus D is dense in X. ☝

In particular, if $\phi_i(D) = X_i$ for all i, then D is dense in X (this has a simplified direct proof using AC); but D can be a proper subset of X (XIII.14.4.17.).

XIII.14.5. Exercises

XIII.14.5.1. Modify the definition of direct and inverse systems and limits by also defining $\phi_{i,i}$ to be the identity map on X_i for each i, and replacing $<$ by $\leq$ throughout. Show that this modified definition also works over directed sets which are not properly directed. Describe the limit if the index set is not properly directed.

XIII.14.5.2. (a) Let I be a directed set and G_i a group for each i, with $\phi_{i,j} : G_i \to G_j$ a homomorphism for $i < j$. Let $(G, \phi_i) = \varinjlim(G_i, \phi_{i,j})$. Show that G has a unique group structure making each ϕ_i a homomorphism.

(b) If each G_i is abelian, show that G is abelian. What about the converse?

(c) Do the same if each G_i is a ring, vector space, or topological group.

(d) Do the analog of (a)–(c) for inverse systems and limits.

XIII.14.5.3. [Dug78] Let $\mathbb{T}$ be the unit circle in the plane, regarded as the compact topological group of complex numbers of absolute value 1. For each n, let $X_n = \mathbb{T}$ and $\phi_{n,n+1} : X_n \to X_{n+1}$ defined by $\phi_{n,n+1}(z) = z^2$. Then $(X_n, \phi_{n,n+1})$ is a direct system of compact Hausdorff spaces. Let $(X, \phi_n) = \varinjlim(X_n, \phi_{n,n+1})$. Then X has a natural structure as a topological group (XIII.14.5.2.).

(a) Let H be the subgroup of $\mathbb{T}$ consisting of all roots of unity whose order is a power of 2. Show that X can be naturally identified with $\mathbb{T}/H$.

(b) Show that the inductive limit topology on X is the indiscrete topology.

(c) Repeat the exercise if $\phi_{n,n+1}(z) = z^{k_n}$ for an arbitrary sequence (k_n) in $\mathbb{N} \setminus \{1\}$.

XIII.14.5.4. Construct an explicit countable subset A of c_{oo} such that 0 is in the closure of A in the inductive limit topology, but there is no sequence in A converging to 0. [For $n, m \in \mathbb{N}$, consider the sequence with first term $\frac{1}{n}$, n'th term $\frac{1}{m}$, and all other terms 0.]

XIII.14.5.5. (a) Show (using DC) that if $(X_n, \phi_{n+1,n})$ is an inverse system of nonempty sets over $\mathbb{N}$, and each $\phi_{n+1,n}$ is surjective, and $(X, \phi_n) = \varprojlim(X_n, \phi_{n+1,n})$, then X is nonempty and each ϕ_n is surjective.

(b) (I. Farah) For each $\alpha < \omega_1$ let X_α be the set of all injective functions from α to $\mathbb{N}$ for which the complement of the range is infinite. If $\alpha < \beta$, let $\phi_{\beta,\alpha} : X_\beta \to X_\alpha$ be defined by $\phi_{\beta,\alpha}(f) = f|_\alpha$. Then $(X_\alpha, \phi_{\beta,\alpha})$ is an inverse system of nonempty sets over the directed (well-ordered) set ω_1. Show that each connecting map $\phi_{\beta,\alpha}$ is surjective, but $\varprojlim(X_\alpha, \phi_{\beta,\alpha}) = \emptyset$.

XIII.14.5.6. (a) If $(X_i, \phi_{j,i})$ is an inverse system of nonempty compact spaces, show that the inverse limit can be empty if the X_i are not Hausdorff. [Give the sets in XIII.14.2.7.(iii) or XIII.14.5.5.(b) the indiscrete topology or the finite complement topology.]

(b) Prove (using AC) that an inverse limit of compact spaces is compact (regarding $\emptyset$ as a compact space) by adapting the proof of Tikhonov's theorem.

XIII.14.5.7. Let $(X_n, \phi_{n+1,n})$ be an inverse system over $\mathbb{N}$ of nonempty finite sets with the discrete topology, and $(X, \phi_n) = \varprojlim(X_n, \phi_{n+1,n})$.

(a) If $\phi_{n+1,n}^{-1}(p)$ has at least two elements for each n and each $p \in X_n$, show that X is homeomorphic to the Cantor set [use ()].

(b) Let $X_n = \{k \in \mathbb{N} : k \leq n\}$, and define $\phi_{n+1,n}(k) = k$ for $k \leq n$, $\phi_{n+1,n}(n+1) = n$. Show that $\varprojlim(X_n, \phi_{n+1,n})$ is naturally homeomorphic to the one-point compactification of $\mathbb{N}$ (equivalently, to $\{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\} \subseteq \mathbb{R}$).

XIII.14.5.8. Let G be a group, I a directed set, and $\{H_i : i \in I\}$ a collection of normal subgroups of G with $H_j \subseteq H_i$ for $i < j$. Set $G_i = G/H_i$ for all i. Then there is a natural quotient map $\pi_j, i : G_j = G/H_j \to G/H_i = G_i$ for $i < j$, and $(G_i, \pi_{j,i})$ is an inverse system of groups. Set $(\bar{G}, \pi_i) = \varprojlim(G_i, \pi_{j,i})$. Then $\bar{G}$ is a group, and there is a homomorphism $\alpha : G \to \bar{G}$.

(a) Show that the kernel of α is exactly $\cap_i H_i$. In particular, α is injective if (and only if) $\cap_i H_i = \{e_G\}$.

(b) Show that α is not surjective in general (cf. XIII.14.4.13., XIII.14.4.17.).

(c) If each G_i has a topology making it into a compact (Hausdorff) topological group, with the $\pi_{j,i}$ continuous, then $\bar{G}$ is a compact (Hausdorff) topological group in the inverse limit topology.

(d) If each G_i is finite, give G_i the discrete topology. Then $\bar{G}$ is a compact totally disconnected (zero-dimensional) topological group. Every compact totally disconnected topological group arises in this manner as an inverse limit of finite groups ().

We may regard $\bar{G}$ as a "completion" of G with respect to the subgroups H_i. The examples of XIII.14.4.13., XIII.14.4.17., and **??** are special cases.

XIII.14.5.9. Let $\mathbb{Z}_{\boldsymbol{a}}$ be the $\boldsymbol{a}$-adic numbers as defined in XIII.14.4.13.. Show that $\mathbb{Z}_{\boldsymbol{a}}$ is torsion-free if and only if the supernatural number $a_1 \prod_n \frac{a_{n+1}}{a_n}$ (="$\lim_{n\to\infty} a_n$") has all exponents either 0 or ∞. [Use that $\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ if m and n are relatively prime () and XIII.14.5.11..]

XIII.14.5.10. Work out the details of writing the Cantor set, the Hawaiian earring, the infinite broom, the topologist's sine curve, solenoids, Knaster bucket-handles, and the pseudo-arc as inverse limits as described in XIII.14.4.11., XIII.14.4.12., XIII.14.4.17., and XIII.14.4.18..

XIII.14.5.11. Let I be an infinite index set (not ordered or directed). For each $i \in I$, let X_i be a set. Let $\mathcal{F}$ be the set of all finite subsets of I, directed by inclusion. For each $F \in \mathcal{F}$, define

$$Z_F = \prod_{i \in F} X_i$$

and if $F \subseteq G \in \mathcal{F}$, let $\pi_{G,F} : Z_G \to Z_F$ be the projection onto the coordinates in F. Show that

$$\left(\prod_{i \in I} X_i, \pi_F \right)$$

is an inverse limit for (Z_F, π_F), where π_F is the natural projection from $\prod_{i \in I} X_i$ onto $\prod_{i \in F} X_i$. If each X_i is a topological space, show that the inverse limit topology is the product topology.

Thus, once finite Cartesian products have been defined, infinite Cartesian products may be regarded as a special case of inverse limits. (However, this is not a logically sound way to *define* infinite Cartesian products, since infinite Cartesian products are used to construct inverse limits so there is a circularity problem.)

XIII.14.5.12. Restricted Direct Products. Let I be an infinite index set (not ordered or directed). For each $i \in I$, let X_i be a locally compact Hausdorff space and Y_i a compact open (hence clopen) subset of X_i. Let $\mathcal{F}$ be the set of all finite subsets of I, directed by inclusion. For each $F \in \mathcal{F}$, define

$$Z_F = \left\{ (\cdots x_i \cdots) \in \prod_i X_i : x_i \in Y_i \text{ for all } i \notin F \right\} \subseteq \prod_{i \in I} X_i$$

and give Z_F the restriction of the product topology. Then

$$Z_F \cong \prod_{i \in F} X_i \times \prod_{i \in I \setminus F} Y_i$$

and thus Z_F is locally compact by (). If $F \subseteq G \in \mathcal{F}$, the inclusion $\phi_{F,G}$ of Z_F into Z_G is a homeomorphism onto a clopen subset of Z_G. Let

$$(Z, \phi_F) = \varinjlim (Z_F, \phi_{F,G})$$

with the inductive limit topology. Z is called the *restricted direct product of the X_i with respect to the Y_i*, denoted

$$Z = \prod_{i \in I}{}' (X_i, Y_i) \ .$$

(a) As a set, Z can be identified with the subset $\cup_F Z_F$ of $\prod_i X_i$, the set of tuples $(\cdots x_i \cdots)$ with $x_i \in Y_i$ for all but finitely many i.

(b) Show that Z is locally compact and the image $\phi_F(Z_F)$ of Z_F is a clopen subset (homeomorphic to Z_F) for each $F \in \mathcal{F}$.

(c) Show that the inductive limit topology on Z is strictly stronger than the restriction to Z of the product topology if $Y_i \neq X_i$ for infinitely many i. [The image $\phi_F(Z_F)$ is not open in Z in the product topology.]

Restricted direct products are often used where the X_i have additional structure such as a topological group (). Here is one especially important application. The completions of the rational numbers $\mathbb{Q}$ to topological fields are precisely $\mathbb{R}$ and the p-adic numbers $\mathbb{Q}_p$ for primes p. Each of these is locally compact. The topological field $\mathbb{Q}_p$ contains the ring $\mathbb{Z}_p$ of p-adic integers as a compact open subring. Form

$$\mathbb{A}_{\mathbb{Q}} = \mathbb{R} \times \prod_p{}' (\mathbb{Q}_p, \mathbb{Z}_p)$$

which is naturally a locally compact topological ring, called the *ring of adeles of* $\mathbb{Q}$. There is a natural embedding of $\mathbb{Q}$ into $\mathbb{A}_{\mathbb{Q}}$, whose image turns out to be dense. A similar construction can be done for any algebraic number field (finite extension of $\mathbb{Q}$). Rings of adeles are important in number theory; cf. [].

XIII.14.5.13. (a) If X and Y are sets, use the notation Y^X for the set of all functions from X to Y. If $(X_i, \phi_{i,j})$ is a direct system of sets over I, and Y is a set, for $i < j$ define $\phi^*_{j,i} : Y^{X_j} \to Y^{X_i}$ by

$$\phi^*_{j,i}(f) = f \circ \phi_{i,j} \ .$$

Then $(Y^{X_i}, \phi^*_{j,i})$ is an inverse system of sets over I. If (X, ϕ_i) is a direct limit for $(X_i, \phi_{i,j})$, show that (Y^X, ϕ^*_i) is an inverse limit for $(Y^{X_i}, \phi^*_{j,i})$, where $\phi^*_i(f) = f \circ \phi_i$. Thus we may write

$$\varprojlim (Y^{X_i}, \phi^*_{j,i}) \cong Y^{\varinjlim (X_i, \phi_{i,j})} \ .$$

(b) If X and Y are topological spaces, use the notation Y^X for the set of all continuous functions from X to Y. If $(X_i, \phi_{i,j})$ is a direct system of topological spaces over I, and Y is a topological space, for $i < j$ define $\phi^*_{j,i} : Y^{X_j} \to Y^{X_i}$ by

$$\phi^*_{j,i}(f) = f \circ \phi_{i,j} \ .$$

Then $(Y^{X_i}, \phi^*_{j,i})$ is an inverse system of topological spaces over I. If (X, ϕ_i) is a set-theoretic direct limit for $(X_i, \phi_{i,j})$, show that (Y^X, ϕ^*_i) is a set-theoretic inverse limit for $(Y^{X_i}, \phi^*_{j,i})$, where $\phi^*_i(f) = f \circ \phi_i$. If all the spaces of functions are given the topology of pointwise convergence, show that the topology on Y^X is the inverse limit topology.

(c) If $(X_i, \phi_{j,i})$ is an inverse system of sets over I, and Y is a set, for $i < j$ define $\phi^*_{i,j} : Y^{X_i} \to Y^{X_j}$ by

$$\phi^*_{i,j}(f) = f \circ \phi_{j,i} \ .$$

Then $(Y^{X_i}, \phi^*_{i,j})$ is a direct system of sets over I. If (X, ϕ_i) is an inverse limit for $(X_i, \phi_{j,i})$, is it true that (Y^X, ϕ^*_i) is a direct limit for $(Y^{X_i}, \phi^*_{i,j})$, where $\phi^*_i(f) = f \circ \phi_i$? What is the relation between the two? What if the sets are topological spaces with modifications as in (b)?

XIII.14.5.14. Consider inverse limits $\varprojlim (X_n, \phi_{n+1,n})$ of Hausdorff spaces over $\mathbb{N}$. Let P be a property which is preserved under countable products and closed subspaces. Then if each X_n has P, so does the inverse limit. Examples of such properties are: first countability, second countability, metrizability, and compactness (cf. ()).

XIII.14.5.15. Uniqueness of the Cantor Set. A compactum is *perfect* if it has no isolated points. Prove the following theorem:

THEOREM. Every perfect zero-dimensional metrizable compactum is homeomorphic to the Cantor set.

(a) If X is a perfect zero-dimensional compactum, prove by induction that for each $k \in \mathbb{N}$, X can be partitioned into exactly k nonempty perfect clopen subsets.

(b) Let (X, ρ) be a perfect zero-dimensional compact metric space. For any $\epsilon > 0$, for all sufficiently large k X can be partitioned into exactly k nonempty perfect clopen subsets, each of diameter $< \epsilon$.

(c) Let (X, ρ) be a perfect zero-dimensional compact metric space. Let $\mathcal{P}_1$ be a partition of X into k_1 nonempty perfect clopen subsets $\{X_{1,1}, \ldots, X_{1,k_1}\}$, each of diameter < 1, where k_1 is a power of 2. Now divide each $X_{1,r}$ into k_2 nonempty perfect clopen sets, each of diameter $< \frac{1}{2}$, where k_2 is a power of 2. To avoid unwieldy notation, denote these sets by $\mathcal{P}_2 = \{X_{2,1}, \ldots, X_{2,m_2}\}$, where $m_2 = k_1 k_2$ is also a power of 2. Continue inductively to successively partition X into $\mathcal{P}_n = \{X_{n,1}, \ldots, X_{n,m_n}\}$ consisting of nonempty

perfect clopen sets of diameter $< \frac{1}{n}$, with m_n a power of 2, each partition refining the previous one and each set in the previous partition divided into the same number of subsets (a power of 2) in the new one.

(d) For each n let Y_n be a finite discrete space with m_n points, and let α_n be the map from X to Y_n sending the points of $X_{n,r}$ to the r'th point, i.e. identifying Y_n with the sets in $\mathcal{P}_n$. Let ϕ_{n+1} be the obvious map from Y_{n+1} to Y_n sending a point corresponding to a set in $\mathcal{P}_{n+1}$ to the point corresponding to the containing set in $\mathcal{P}_n$. Obtain a commutative diagram

$$\begin{array}{ccccccccc} X & & \cdots & \xrightarrow{id} & X & \xrightarrow{id} & X & \xrightarrow{id} & X \\ \vdots\,\alpha & & & & \downarrow{\scriptstyle \alpha_3} & & \downarrow{\scriptstyle \alpha_2} & & \downarrow{\scriptstyle \alpha_1} \\ Y & & \cdots & \xrightarrow{\phi_4} & Y_3 & \xrightarrow{\phi_3} & Y_2 & \xrightarrow{\phi_2} & Y_1 \end{array}$$

(e) Show that the induced map α from X to $Y = \varprojlim(Y_n, \phi_{n+1})$ is a homeomorphism, and that Y is homeomorphic to the Cantor set (XIII.14.4.11.).

XIII.14.5.16. Let Y be a zero-dimensional metrizable compactum, and K the Cantor set.

(a) Show that $Y \times K$ is a perfect zero-dimensional metrizable compactum, hence homeomorphic to K (XIII.14.5.15.).

(b) The first coordinate projection maps $Y \times K$ onto Y. Conclude that Y is a continuous image of K. Compare to XIII.13.6.7.(vi).

XIII.14.5.17. Prove the following theorem:

THEOREM. Every metrizable compactum is a continuous image of the Cantor set.

(a) Let (X, ρ) be a compact metric space. There is a finite open cover $\mathcal{U}_1$ of X where each $U \in \mathcal{U}_1$ has diameter < 1. Let Y_1 be the separated union of the $\{\bar{U} : U \in \mathcal{U}_1\}$. Then Y_1 is a compact metrizable space, using the restriction of ρ to each $\bar{U}$ and setting the distance between points of different $\bar{U}$ equal to 1. The $\bar{U}$ are clopen in Y_1. There is an obvious map α_1 from Y_1 onto X which is the identity on each $\bar{U}$.

(b) Similarly take a finite cover of each $\bar{U}$ by open sets of diameter $< \frac{1}{2}$ and let Y_2 be the separated union of the closures of all the sets covering the sets in Y_1. There is an obvious map ϕ_2 from Y_2 onto Y_1, and thus a map $\alpha_2 = \alpha_1 \circ \phi_2$ from Y_2 onto X. Continue inductively to get metrizable compacta Y_n, with each component of diameter $< \frac{1}{n}$, and a commutative diagram of inverse systems

$$\begin{array}{ccccccccc} Y & & \cdots & \xrightarrow{\phi_4} & Y_3 & \xrightarrow{\phi_3} & Y_2 & \xrightarrow{\phi_2} & Y_1 \\ \vdots\,\alpha & & & & \downarrow{\scriptstyle \alpha_3} & & \downarrow{\scriptstyle \alpha_2} & & \downarrow{\scriptstyle \alpha_1} \\ X & & \cdots & \xrightarrow{id} & X & \xrightarrow{id} & X & \xrightarrow{id} & X \end{array}$$

with $Y = \varprojlim(Y_n, \phi_{n+1})$ compact and metrizable and the induced map α surjective.

(c) Show that Y is zero-dimensional. [If p and q are distinct points of Y, there is an n such that the images in Y_n are distinct, and hence there is an $m > n$ such that the images are in different components.]

(d) Y need not be perfect. But by XIII.14.5.16., Y is a continuous image of the Cantor set.

XIII.14.5.18. The Hahn-Mazurkiewicz Theorem [,]. Prove the following theorem:

THEOREM. [HAHN-MAZURKIEWICZ THEOREM] A metrizable compactum X is a continuous image of $[0,1]$ if and only if X is connected and locally connected.

(a) A continuous image of $[0,1]$ is a connected, locally connected metrizable compactum ().

(b) If (X,ρ) is a locally connected compact metric space, show that X is *uniformly locally path connected*: For any $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $p, q \in X$ with $\rho(p,q) < \delta$, there is a path γ from p to q with $diam(\gamma) < \epsilon$. [Use ().]

(c) Let X be a connected, locally connected metrizable compactum. Regard the Cantor set K as a subset of $[0,1]$ in the usual way. Let ϕ be a map of K onto X (XIII.14.5.17.). Extend ϕ to a continuous function from $[0,1]$ onto X by successively defining it on each subinterval of $[0,1] \setminus K$ (note that X is path-connected by XIII.13.9.11. and XIII.13.9.17.), using (b) to obtain continuity of the extension.

XIII.14.5.19. Let $(X_i, \phi_{j,i})$ be an inverse system of compact Hausdorff spaces, X the inverse limit, and $\phi_i : X \to X_i$ the induced map for each i. Show that for each open neighborhood U of the diagonal in $X \times X$ there is an i such that $x, y \in X$ and $\phi_i(x) = \phi_i(y)$ implies $(x,y) \in U$. [If not, there is a U such that for each i there is $(x_i, y_i) \notin U$ with $\phi_i(x_i) = \phi_i(y_i)$. Let (x,y) be a cluster point of the net $((x_i, y_i))$. Show that $\phi_i(x) = \phi_i(y)$ for all i, so $x = y$, a contradiction.] In particular, if X is metrizable by a metric ρ, then for every $\epsilon > 0$ there is an i such that $x, y \in X$ and $\phi_i(x) = \phi_i(y)$ implies $\rho(x,y) < \epsilon$.

XIII.15. Uniform Spaces

The theory of uniform spaces and uniform structures fits somewhere between the theories of metric spaces and topological spaces. They are more general than metric spaces, but similarly give notions of uniform convergence and uniform continuity which are absent in general topological spaces. There is an underlying topology defined by a uniform structure, but the uniform structure is finer than the underlying topology since in general many different uniform structures can give the same topology, just as with metrics. Not every topology comes from a uniform structure; the ones that do turn out to be precisely the completely regular topologies.

There are two seemingly different standard approaches to the theory of uniform structures which turn out to be equivalent. One is based on metric spaces, the other on topological groups. The second approach seems more general, and it is remarkable that the two are actually the same (under some mild Choice assumptions).

The first approach uses a suitable family of pseudometrics. In the second approach, a uniform structure on a set X is a set of "neighborhoods" of the diagonal $\Delta = \{(x, x) : x \in X\}$ in $X \times X$, satisfying certain natural properties. Such a neighborhood W gives a uniform notion of closeness in X: x and y are "W-close" if $(x, y) \in W$. The two prototype examples of such systems are:

(i) If (X, ρ) is a metric space, then for every $\epsilon > 0$ there is a corresponding "neighborhood"

$$W_\epsilon^\rho = \{(x, y) : \rho(x, y) < \epsilon\}$$

of Δ in $X \times X$.

(ii) If X is a topological group, and U is a neighborhood of the identity element in X, then there is a corresponding "neighborhood"

$$W_U = \{(x, y) : xy^{-1} \in U\}$$

of Δ in $X \times X$.

If $X = \mathbb{R}$, these two types of neighborhoods are essentially the same if $\mathbb{R}$ is regarded as a metric space with the standard metric, or as a topological group under addition; a typical "neighborhood" of Δ is

$$\{(x, y) \in \mathbb{R}^2 : |x - y| < \epsilon\}$$

for some fixed $\epsilon > 0$, which is geometrically an open strip of uniform width around the line $y = x$.

Uniform spaces in the modern sense were first defined and studied by A. Weil [Wei37].

XIII.15.1. Topologies and Uniform Structures Defined by Pseudometrics

A straightforward but powerful generalization of a metric space is a space with a family of pseudometrics. Recall () that a *pseudometric* on a set X is a function $\rho : X \times X \to [0, \infty)$ with all the properties of a metric except that $\rho(x, y) = 0$ does not necessarily imply that $x = y$.

XIII.15.1.1. A family of pseudometrics on a set X is simply a set $\mathcal{P} = \{\rho_i : i \in I\}$ of pseudometrics on X. The family $\mathcal{P}$ is called *separating* if $\rho_i(x, y) = 0$ for all i implies that $x = y$, i.e. for any distinct $x, y \in X$ there is an i such that $\rho_i(x, y) > 0$.

Note that a single metric on a set can be regarded as a separating family of pseudometrics with one element.

We will think of a family of pseudometrics on a set X as defining a *uniform structure* on X; this will be made precise in ().

XIII.15.1.2. To each pseudometric ρ on a set X we can associate open balls $B_\epsilon^\rho(x) = \{y \in X : \rho(x,y) < \epsilon\}$, just as in the metric case. If we have a family $\mathcal{P}$ of pseudometrics, we may take the set of all open balls with respect to the pseudometrics in $\mathcal{P}$ as a subbase for a topology, called the *topology induced by* $\mathcal{P}$, denoted $\mathcal{T}_\mathcal{P}$. These open balls do not form a base for $\mathcal{T}_\mathcal{P}$ in general if $\mathcal{P}$ contains more than one pseudometric, but finite intersections do, i.e. sets of the form

$$B_\epsilon^F(x) = \{y \in X : \rho_i(x,y) < \epsilon \text{ for all } i \in F\}$$

where F is a finite subset of I and $\epsilon > 0$, form a base for $\mathcal{T}_\mathcal{P}$, i.e. a subset U of X is open if and only if, for every $x \in U$, there is a finite subset F of I and and $\epsilon > 0$ such that $B_\epsilon^F(x) \subseteq U$. (We will also use the analogous notation $B_\epsilon^\mathcal{F}(x)$, for a finite subset $\mathcal{F}$ of $\mathcal{P}$.)

Exactly as in the metric case (), using the triangle inequality every open ball has this property, hence is an open set.

XIII.15.1.3. Convergence in the topology $\mathcal{T}_\mathcal{P}$ is easily described. If (x_i) is a net in X and $x \in X$, then $x_i \to x$ in $(X, \mathcal{T}_\mathcal{P})$ if and only if $\rho(x_i, x) \to 0$ for every $\rho \in \mathcal{P}$, i.e. if and only if for every $\rho \in \mathcal{P}$ and $\epsilon > 0$ there is an i_0 such that $\rho(x_i, x) < \epsilon$ for all $i \geq i_0$.

The next observation generalizes the fact that a metrizable topology is Hausdorff:

XIII.15.1.4. PROPOSITION. Let $\mathcal{P}$ be a family of pseudometrics on a set X. Then the following are equivalent:

(i) $\mathcal{T}_\mathcal{P}$ is T_0.

(ii) $\mathcal{T}_\mathcal{P}$ is Hausdorff.

(iii) $\mathcal{P}$ is separating.

PROOF: If $\mathcal{P}$ is separating, for any $x, y \in X$, $x \neq y$, let $\rho \in \mathcal{P}$ with $\rho(x,y) > 0$, and set $\epsilon = \frac{\rho(x,y)}{2}$. Then $B_\epsilon^\rho(x)$ and $B_\epsilon^\rho(y)$ are disjoint open neighborhoods of x and y, so $\mathcal{T}_\mathcal{P}$ is Hausdorff. If $\mathcal{P}$ is not separating, let $x, y \in X$, $x \neq y$, with $\rho(x,y) = 0$ for all $\rho \in \mathcal{P}$. Then any open set containing x also contains y and vice versa, so $\mathcal{T}_\mathcal{P}$ is not T_0. 👍

Uniform Domination and Saturated Families

We need some tools to compare families of pseudometrics, generalizations of the notions of uniform domination and uniform equivalence of metrics ().

XIII.15.1.5. DEFINITION. Let $\rho_1, \ldots, \rho_n, \sigma$ be pseudometrics on a set X. Then σ is *uniformly dominated by* $\{\rho_1, \ldots, \rho_n\}$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $x, y \in X$ and $\rho_k(x,y) < \delta$ for $1 \leq k \leq n$, then $\sigma(x,y) < \epsilon$.

If $\mathcal{P}$ is a family of pseudometrics on X, then σ is *uniformly dominated by* $\mathcal{P}$ if for every $\epsilon > 0$ there is a $\delta > 0$ and a finite subset $\{\rho_1, \ldots, \rho_n\}$ of $\mathcal{P}$ such that, whenever $x, y \in X$ and $\rho_k(x,y) < \delta$ for $1 \leq k \leq n$, then $\sigma(x,y) < \epsilon$.

If $\mathcal{P}$ and $\mathcal{Q}$ are families of pseudometrics on X, then $\mathcal{Q}$ is *uniformly dominated by* $\mathcal{P}$ if every $\sigma \in \mathcal{Q}$ is uniformly dominated by $\mathcal{P}$. $\mathcal{P}$ and $\mathcal{Q}$ are *uniformly equivalent* if each is uniformly dominated by the other.

A family $\mathcal{P}$ of pseudometrics on X is *saturated* if every pseudometric on X which is uniformly dominated by $\mathcal{P}$ is in $\mathcal{P}$.

XIII.15.1.6. PROPOSITION. Let $\mathcal{P}$ be a family of pseudometrics on a set X. Then there is a smallest saturated family $\bar{\mathcal{P}}$ of pseudometrics on X containing $\mathcal{P}$. $\bar{\mathcal{P}}$ consists precisely of all pseudometrics on X which are uniformly dominated by $\mathcal{P}$.

If $\mathcal{Q}$ is another family of pseudometrics on X, then $\mathcal{Q}$ is uniformly dominated by $\mathcal{P}$ if and only if $\bar{\mathcal{Q}} \subseteq \bar{\mathcal{P}}$, and $\mathcal{P}$ and $\mathcal{Q}$ are uniformly equivalent if and only if $\bar{\mathcal{P}} = \bar{\mathcal{Q}}$. In particular, $\mathcal{P}$ and $\bar{\mathcal{P}}$ are uniformly equivalent.

The proof is left as an exercise (Exercise ()).

XIII.15.1.7. PROPOSITION. Let $\mathcal{P}$ and $\mathcal{Q}$ be families of pseudometrics on a set X. If $\mathcal{Q}$ is uniformly dominated by $\mathcal{P}$, then $\mathcal{T}_{\mathcal{P}}$ is stronger than $\mathcal{T}_{\mathcal{Q}}$. If $\mathcal{P}$ and $\mathcal{Q}$ are uniformly equivalent, then $\mathcal{T}_{\mathcal{P}} = \mathcal{T}_{\mathcal{Q}}$. In particular, $\mathcal{T}_{\bar{\mathcal{P}}} = \mathcal{T}_{\mathcal{P}}$.

PROOF: Let $\sigma \in \mathcal{Q}$, $x \in X$, and $\epsilon > 0$. Choose $\delta > 0$ and $\mathcal{F} = \{\rho_1, \ldots, \rho_n\}$ as in the definition. Then $B^{\mathcal{F}}_{\delta}(x) \subseteq B^{\sigma}_{\epsilon}(x)$. 👍

XIII.15.1.8. The converse of XIII.15.1.7. is false in general: families of pseudometrics generating the same topology need not be uniformly equivalent, even if they are singletons (metrics can be equivalent but not uniformly equivalent ()). Two families of pseudometrics generating the same topology are called *equivalent*. Equivalent but not uniformly equivalent families generate the same topology, but different uniform structures; thus uniform structures are more delicate than topologies.

XIII.15.1.9. In light of XIII.15.1.7., there is no loss of generality in considering only saturated families of pseudometrics. There is some technical advantage in working with saturated families: for example, the open balls form a base for the underlying topology instead of just a subbase. In fact, the open balls from a saturated family frequently give all, or almost all, of the open sets in the topology (they precisely give all the cozero sets, hence all open F_σ's if the space is normal). And the ϵ can be dispensed with in considering open balls from a saturated family $\mathcal{P}$: since a saturated family of pseudometrics is closed under nonnegative scalar multiples, any ball of the form $B^{\rho}_{\epsilon}(x)$ for $\rho \in \mathcal{P}$ is of the form $B^{\sigma}_{1}(x)$ for some $\sigma \in \mathcal{P}$ (specifically, $\sigma = \epsilon\rho$), so only balls of radius 1 need be considered.

Uniform Continuity

Uniform structures give a notion of uniform continuity and uniform convergence for functions between uniform spaces.

XIII.15.1.10. DEFINITION. Let X and Y be sets, and $\mathcal{P}$ and $\mathcal{Q}$ sets of pseudometrics on X and Y respectively. A function $f : X \to Y$ is $(\mathcal{P}, \mathcal{Q})$-*uniformly continuous* if, for each $\sigma \in \mathcal{Q}$ the pseudometric $\rho_{f,\sigma}$ on X is uniformly dominated by $\mathcal{P}$, where

$$\rho_{f,\sigma}(p, q) = \sigma(f(p), f(q)) \ .$$

XIII.15.1.11. In other words, f is uniformly continuous if and only if for each $\sigma \in \mathcal{Q}$ and $\epsilon > 0$, there is a $\delta > 0$ and $\rho_1, \ldots, \rho_n \in \mathcal{P}$ such that, whenever $\rho_k(x, y) < \delta$ for $1 \leq k \leq n$, then $\sigma(f(p), f(q)) < \epsilon$.

If $\mathcal{P}$ and $\mathcal{Q}$ are singletons consisting of metrics, this is precisely the usual notion of uniform continuity with respect to the metrics (). The condition in the definition works equally for a finite set $\{\sigma_1, \ldots, \sigma_m\}$ of pseudometrics in $\mathcal{Q}$.

XIII.15.1.12. PROPOSITION. Let X, Y, Z be sets, and $\mathcal{P}, \mathcal{Q}, \mathcal{R}$ sets of pseudometrics on X, Y, Z respectively. If $f : X \to Y$ is $(\mathcal{P}, \mathcal{Q})$-uniformly continuous and $g : Y \to Z$ is $(\mathcal{Q}, \mathcal{R})$-uniformly continuous, then $g \circ f$ is $(\mathcal{P}, \mathcal{R})$-uniformly continuous.

$(\mathcal{P}, \mathcal{Q})$-Uniform continuity depends on $\mathcal{P}$ and $\mathcal{Q}$ only up to uniform equivalence:

XIII.15.1.13. PROPOSITION. Let X and Y be sets, and $\mathcal{P}$ and $\mathcal{Q}$ sets of pseudometrics on X and Y respectively. A function $f : X \to Y$ is $(\mathcal{P}, \mathcal{Q})$-uniformly continuous if and only if it is $(\bar{\mathcal{P}}, \bar{\mathcal{Q}})$-uniformly continuous. Thus if $\mathcal{P}_1$ and $\mathcal{P}_2$ [resp. $\mathcal{Q}_1$ and $\mathcal{Q}_2$] are uniformly equivalent families of pseudometrics on X [resp. Y], and $f : X \to Y$, then f is $(\mathcal{P}_1, \mathcal{Q}_1)$-uniformly continuous if and only if it is $(\mathcal{P}_2, \mathcal{Q}_2)$-uniformly continuous.

More generally, we have:

XIII.15.1.14. PROPOSITION. Let X and Y be sets. If $\mathcal{P}_1$ and $\mathcal{P}_2$ [resp. $\mathcal{Q}_1$ and $\mathcal{Q}_2$] are families of pseudometrics on X [resp. Y] such that $\mathcal{P}_1$ uniformly dominates $\mathcal{P}_2$ and $\mathcal{Q}_2$ uniformly dominates $\mathcal{Q}_1$, and $f : X \to Y$ is $(\mathcal{P}_2, \mathcal{Q}_2)$-uniformly continuous, then f is $(\mathcal{P}_1, \mathcal{Q}_1)$-uniformly continuous.

Uniformly continuous functions are continuous:

XIII.15.1.15. PROPOSITION. Let X and Y be sets, and $\mathcal{P}$ and $\mathcal{Q}$ sets of pseudometrics on X and Y respectively. If a function $f : X \to Y$ is $(\mathcal{P}, \mathcal{Q})$-uniformly continuous, then f is continuous for $\mathcal{T}_{\mathcal{P}}$ and $\mathcal{T}_{\mathcal{Q}}$.

The converse is false in general, even if $\mathcal{P}$ and $\mathcal{Q}$ are singletons.

Uniform Convergence

Let X and Y be sets, and $\mathcal{Q}$ a set of pseudometrics on Y. Let (f_i) be a net of functions from X to Y, and f a function from X to Y. We do not make any continuity assumptions on the f_i or f; indeed, X need not have any uniform or topological structure.

XIII.15.1.16. DEFINITION. We say $f_i \to f$ *$\mathcal{Q}$-uniformly* if for every $\sigma \in \mathcal{Q}$ and $\epsilon > 0$, there is a i_0 such that $\sigma(f_i(x), f(x)) < \epsilon$ for all $i \geq i_0$ and all $x \in X$.

If $\mathcal{Q} = \{\sigma\}$, this definition agrees with the one in ().

The next result generalizes () and ().

XIII.15.1.17. Let X and Y be sets, and $\mathcal{P}$ and $\mathcal{Q}$ sets of pseudometrics on X and Y respectively. Let (f_i) be a net of functions from X to Y converging $(\mathcal{P}, \mathcal{Q})$-uniformly to $f : X \to Y$. Then:

(i) If each f_i is $(\mathcal{T}_\mathcal{P}, \mathcal{T}_\mathcal{Q})$-continuous, then f is $(\mathcal{T}_\mathcal{P}, \mathcal{T}_\mathcal{Q})$-continuous.

(ii) If each f_i is $(\mathcal{P}, \mathcal{Q})$-uniformly continuous, then f is $(\mathcal{P}, \mathcal{Q})$-uniformly continuous.

Bounded Pseudometrics

XIII.15.1.18. If ρ is a pseudometric on a set X, then $\tilde{\rho}$, defined by

$$\tilde{\rho}(x, y) = \min\{\rho(x, y), 1\}$$

is a bounded pseudometric on X which is uniformly equivalent to ρ. So if $\mathcal{P}$ is a family of pseudometrics on X, define

$$\tilde{\mathcal{P}} = \{\tilde{\rho} : \rho \in \mathcal{P}\} \ .$$

Then $\tilde{\mathcal{P}}$ is a family of bounded pseudometrics which is uniformly equivalent to $\mathcal{P}$. (Note, however, that $\tilde{\mathcal{P}}$ is never saturated unless $\mathcal{P} = \{0\}$, so if $\mathcal{P}$ is saturated and nonzero, then $\tilde{\mathcal{P}} \neq \mathcal{P}$; in fact, $\tilde{\mathcal{P}}$ is a proper subset of $\mathcal{P}$ in this case.) Thus we may always work with families of pseudometrics taking values in $[0, 1]$.

Countably Generated Families of Pseudometrics

XIII.15.1.19. If $\mathcal{P} = \{\rho_1, \rho_2, \dots\}$ is a countable family of pseudometrics on a set X, define a pseudometric σ by

$$\sigma(x, y) = \sum_{k=1}^{\infty} 2^{-k} \tilde{\rho}_k(x, y)$$

(it is routine and straightforward to verify that σ is a pseudometric).

XIII.15.1.20. PROPOSITION. The pseudometric σ is finitely uniformly dominated by $\mathcal{P}$.

PROOF: Let $\epsilon > 0$. Choose n so that $2^{-n+1} < \epsilon$. Set $\delta = \frac{\epsilon}{2n}$, and $\mathcal{F} = \{\rho_1, \dots, \rho_n\}$. Then, if $x, y \in X$ satisfy $\rho_k(x, y) < \delta$ for $1 \leq k \leq n$, we have

$$\sigma(x, y) = \sum_{k=1}^{n} 2^{-k} \tilde{\rho}_k(x, y) + \sum_{k=n+1}^{\infty} 2^{-k} \tilde{\rho}_k(x, y) < n\delta + 2^{-n} < \epsilon \ .$$

XIII.15.1.21. Conversely, it is obvious that each ρ_k is finitely uniformly dominated by $\mathcal{Q} = \{\sigma\}$. So $\mathcal{Q}$ is a singleton family which is uniformly equivalent to $\mathcal{P}$, so in particular $\mathcal{T}_\mathcal{Q} = \mathcal{T}_\mathcal{P}$. Thus a uniform structure generated by a countable family of pseudometrics is equivalent to the uniform structure generated by a singleton.

The pseudometric σ is a metric if and only if $\mathcal{P}$ is separating. Thus the uniform structure generated by a countable separating family of pseudometrics is also generated by a single metric. In particular, a topology defined by a countable separating family of pseudometrics is metrizable.

XIII.15.1.22. Slightly more generally, a family $\mathcal{P}$ of pseudometrics is *countably generated* if there is a countable family $\mathcal{Q}$ of pseudometrics uniformly equivalent to $\mathcal{P}$. A uniform structure generated by a countably generated family of pseudometrics is also generated by one pseudometric (one metric if the family is separating).

XIII.15.1.23. So the theory of uniform structures generated by families of pseudometrics is technically only a true generalization of the theory of metric spaces in that it allows families of pseudometrics which are not countably generated (as well as families which are not separating). However, as a practical matter the theory has merit even in the countably generated case: a metrizable space often has no *natural* metric, and a choice of a metric is arbitrary; but it may, and often does, have a natural family of pseudometrics generating the topology (cf. ()).

Pseudometrics and Real-Valued Functions

There is an intimate connection between pseudometrics and real-valued functions on a set.

XIII.15.1.24. Let X be a set. If f is a function from X to $\mathbb{R}$, then there is a naturally associated pseudometric ρ_f defined by

$$\rho_f(x, y) = |f(x) - f(y)| \ .$$

Conversely, if ρ is a pseudometric on X, and $x \in X$, there is an associated nonnegative real-valued function $f_{\rho,x}$ defined by

$$f_{\rho,x}(y) = \rho(x, y) \ .$$

XIII.15.1.25. If $f : X \to \mathbb{R}$ is a function, then f is (essentially by definition) uniformly continuous with respect to any family of pseudometrics on X containing ρ_f (and the usual metric on $\mathbb{R}$). Conversely, if ρ is a pseudometric on X, then it follows easily from the triangle inequality that, for any $x \in X$, $f_{\rho,x}$ is uniformly continuous with respect to any family of pseudometrics on X containing ρ.

XIII.15.1.26. Proposition. Let $\mathcal{P}$ be a family of pseudometrics on a set X. Then $\mathcal{T}_\mathcal{P}$ is the weakest topology on X making each function $f_{\rho,x}$ continuous, for all $\rho \in \mathcal{P}$ and $x \in X$.

XIII.15.1.27. COROLLARY. Let $\mathcal{P}$ be a separating family of pseudometrics on a set X. Then the topology $\mathcal{T}_{\mathcal{P}}$ is completely regular. Conversely, every completely regular space has a separating family of pseudometrics giving its topology.

PROOF: The proof is similar to the proof of (). Replacing $\mathcal{P}$ by $\tilde{\mathcal{P}}$ if necessary, we may assume every $\rho \in \mathcal{P}$ takes values in $[0,1]$, so $f_{\rho,x}$ takes values in $[0,1]$ for all ρ and x. Let $\mathcal{S}$ be the set of all $f_{\rho,x}$ for $\rho \in \mathcal{P}$, $x \in X$, and set $Z = [0,1]^{\mathcal{S}}$. Then Z is a compact Hausdorff space.

Define $\phi : X \to Z$ as follows. For each $y \in X$, let $\phi(y)$ be the point in Z whose $f_{\rho,x}$'th coordinate is $f_{\rho,x}(y) = \rho(x,y)$. Then ϕ is continuous for $\mathcal{T}_{\mathcal{P}}$ since its composition with each coordinate map is continuous, and ϕ is one-to-one since $\mathcal{P}$ is separating. The restriction of the topology on Z to $\phi(X)$ gives a topology on $\phi(X) \cong X$ which is weaker than $\mathcal{T}_{\mathcal{P}}$ for which each $f_{\rho,x}$ is continuous; thus this topology is exactly $\mathcal{T}_{\mathcal{P}}$, i.e. ϕ is a homeomorphism onto its image. Thus $(X, \mathcal{T}_{\mathcal{P}})$ is homeomorphic to a subspace of a compact Hausdorff space.

Conversely, if X is completely regular, for each $x \in X$ and open neighborhood U of x there is a function $f_{x,U}$ with $f_{x,U}(x) = 1$ and $f_{x,U} = 0$ on U^c. It is easily checked that the corresponding set of pseudometrics define the topology of X. 👍

Pseudometrics and Continuous Functions

XIII.15.1.28. Now let $(X, \mathcal{T})$ be a topological space. Set

$$\mathcal{C} = \{\rho_f : f \text{ is a continuous function from } X \text{ to } \mathbb{R}\}$$

$$\mathcal{C}^* = \{\rho_f : f \text{ is a continuous function from } X \text{ to } [0,1]\} .$$

Then the topologies $\mathcal{T}_{\mathcal{C}}$ and $\mathcal{T}_{\mathcal{C}^*}$ coincide and are weaker than $\mathcal{T}$. $\mathcal{T}_{\mathcal{C}}$ may be strictly weaker: if $\mathcal{C}$ (or equivalently $\mathcal{C}^*$) is separating, i.e. if continuous functions from X to $\mathbb{R}$ separate points of X, then $\mathcal{T}_{\mathcal{C}}$ is completely regular, but $\mathcal{T}$ need not be completely regular even if it is regular (XIII.7.10.21.). $\mathcal{T}_{\mathcal{C}}$ is the strongest completely regular topology weaker than $\mathcal{T}$ (if $\mathcal{C}$ is separating); in particular, $\mathcal{T} = \mathcal{T}_{\mathcal{C}}$ if $\mathcal{T}$ is completely regular. If $\mathbb{R}$ is given its uniform structure from the usual metric ρ, then every continuous function from X to $\mathbb{R}$ is $(\mathcal{C}, \rho)$-uniformly continuous, and every bounded continuous function from X to $\mathbb{R}$ is $(\mathcal{C}^*, \rho)$-uniformly continuous.

XIII.15.1.29. Even if $(X, \mathcal{T})$ is metrizable, the families $\mathcal{C}$ and $\mathcal{C}^*$ are not countably generated in general, and thus are not uniformly equivalent to a singleton family. The families $\mathcal{C}$ and $\mathcal{C}^*$ are not saturated in general.

The pseudometrics in $\mathcal{C}^*$ are all bounded, and the ones in $\mathcal{C}$ are not in general. One might think that $\mathcal{C}^*$ is essentially the same as $\tilde{\mathcal{C}}$ (XIII.15.1.18.), but this is usually not correct; in fact, $\mathcal{C}^*$ is not even uniformly equivalent to $\tilde{\mathcal{C}}$ (which is uniformly equivalent to $\mathcal{C}$) in general (). ✋

Products of Pseudometric Uniform Structures

XIII.15.1.30. Let $\{X_i : i \in I\}$ be an indexed collection of sets, and for each i let $\mathcal{P}_i$ be a set of pseudometrics on X_i. Set $X = \prod_{i \in I} X_i$, and for each i let π_i be the coordinate projection from X to X_i. Then π_i is surjective for each i (if $X \neq \emptyset$), cf. II.6.4.14.. For each i define

$$\mathcal{P}_i \circ \pi_i = \{\rho \circ (\pi_i \times \pi_i) : \rho \in \mathcal{P}_i\}$$

i.e. $\mathcal{P}_i \circ \pi_i$ consists of all pseudometrics on X of the form $\sigma((x_j),(y_j)) = \rho(x_i, y_i)$ for $\rho \in \mathcal{P}_i$, and set $\mathcal{P} = \cup_{i\in I}(\mathcal{P}_i \circ \pi_i)$. Then $\mathcal{P}$ is a set of pseudometrics on X. The uniform structure on X from $\mathcal{P}$ is called the *product uniform structure*.

XIII.15.1.31. PROPOSITION. (i) The topology $\mathcal{T}_{\mathcal{P}}$ is the product topology of the $\mathcal{T}_{\mathcal{P}_i}$.
(ii) If Y is a set and $\mathcal{Q}$ a set of pseudometrics on Y, then a function $f : Y \to X$ is $(\mathcal{Q}, \mathcal{P})$-uniformly continuous if and only if $\pi_i \circ f$ is $(\mathcal{Q}, \mathcal{P}_i)$-uniformly continuous for each $i \in I$.

XIII.15.2. General Uniform Structures

We now turn to the second approach which seems to be more general, although it turns out to actually be equivalent (under a mild Choice assumption). This approach is due to BOURBAKI [Bou40]; our exposition is adapted from [Kel75].

XIII.15.2.1. DEFINITION. Let X be a set. A *uniform structure*, or *uniformity*, on X is a nonempty set $\mathfrak{U}$ of subsets of $X \times X$, called *entourages*, such that

(i) Every $U \in \mathfrak{U}$ contains the diagonal $\Delta = \{(x, x) : x \in X\}$.

(ii) If $U \in \mathfrak{U}$, then $U^{-1} = \{(y, x) : (x, y) \in U\} \in \mathfrak{U}$.

(iii) If $U, V \in \mathfrak{U}$, then $U \cap V \in \mathfrak{U}$.

(iv) For every $U \in \mathfrak{U}$, there is a $V \in \mathfrak{U}$ with $V \circ V \subseteq U$.

(v) If $U \in \mathfrak{U}$ and $U \subseteq V \subseteq X \times X$, then $V \in \mathfrak{U}$.

The uniform structure $\mathfrak{U}$ is *separating* if $\cap_{U\in\mathfrak{U}} U = \Delta$.

A pair $(X, \mathfrak{U})$, where X is a set and $\mathfrak{U}$ a uniform structure on X, is called a *uniform space*.

XIII.15.2.2. The entourages (elements of $\mathfrak{U}$) are relations on X, and by (i) the relations are reflexive. The composition in (iv) is in the sense of relations, i.e.

$$V \circ V = \{(x, z) : x, z \in X, \exists y \in X \text{ with } (x, y), (y, z) \in V\} .$$

Such composition is associative. Note that if $U, V \in \mathfrak{U}$, then $U, V \subseteq U \circ V$ since $\Delta \subseteq U, V$. So if $U \in \mathfrak{U}$, there is a $V \in \mathfrak{U}$ with $V \circ V \subseteq U$, and a $W \in \mathfrak{U}$ with $W \circ W \subseteq V$. Thus

$$W \circ W \circ W \subseteq W \circ W \circ W \circ W \subseteq V \circ V \subseteq U$$

and similarly for each n there is a $Z \in \mathfrak{U}$ with $Z^n = Z \circ \cdots \circ Z$ (n times) $\subseteq U$.

By (i), (iii), and (v), a uniform structure is a filter () on $X \times X$. The elements of $\mathfrak{U}$ should be thought of as "neighborhoods of Δ" in $X \times X$; condition (iv) is a weak version of the triangle inequality. The sets in $\mathfrak{U}$ should not necessarily be considered "open" sets: indeed, in general, there is no topology in the picture (yet!)

XIII.15.2.3. Any set X has a largest uniformity, namely the collection of all subsets of $X \times X$ containing Δ, called the *discrete uniformity*, and a smallest uniformity, $\{X \times X\}$, called the *indiscrete uniformity*. The discrete uniformity is separating, but the indiscrete one not unless X is a singleton.

Bases and Subbases

XIII.15.2.4. If X is a set, then a union or intersection of uniformities is not necessarily a uniformity (the problem with unions is (iii), and the problem with intersections is (iv)), and if $\mathcal{S}$ is a collection of subsets of $X \times X$ containing Δ, there is not in general a smallest uniformity on X containing $\mathcal{S}$. But we can still define bases and subbases for uniformities in a way analogous to the definitions for topologies, but somewhat simpler since there is no notion of a local base (in fact, there is no notion of "local" in uniform spaces – everything is "uniform"!)

XIII.15.2.5. Definition. Let $\mathfrak{U}$ be a uniform structure on a set X.

(i) A subset $\mathfrak{B}$ of $\mathfrak{U}$ is a *base* for $\mathfrak{U}$ if for every $U \in \mathfrak{U}$ there is a $B \in \mathfrak{B}$ with $B \subseteq U$ (i.e. $\mathfrak{B}$ is a filterbase for $\mathfrak{U}$).

(ii) A subset $\mathfrak{S}$ of $\mathfrak{U}$ is a *subbase* for $\mathfrak{U}$ if the set of finite intersections of elements of $\mathfrak{S}$ is a base for $\mathfrak{U}$.

XIII.15.2.6. If $\mathfrak{B}$ is a base for a uniformity $\mathfrak{U}$, the uniformity $\mathfrak{U}$ is uniquely determined by $\mathfrak{B}$ as the set of all subsets of $X \times X$ containing an element of $\mathfrak{B}$, and $\mathfrak{U}$ is the smallest uniformity containing $\mathfrak{B}$. Similarly, if $\mathfrak{S}$ is a subbase for $\mathfrak{U}$, then $\mathfrak{U}$ is uniquely determined as the smallest uniformity containing $\mathfrak{S}$.

There is a clean description of which subsets of $X \times X$ are bases for uniformities:

XIII.15.2.7. Proposition. Let X be a set and $\mathfrak{B}$ a collection of subsets of $X \times X$. Then $\mathfrak{B}$ is a base for a uniformity on X (which is unique) if and only if $\mathfrak{B}$ has the following properties:

(i) $\Delta \subseteq B$ for all $B \in \mathfrak{B}$.

(ii) If $B \in \mathfrak{B}$, then there is a $C \in \mathfrak{B}$ with $C \subseteq B^{-1}$.

(iii) If $B_1, B_2 \in \mathfrak{B}$, there is a $C \in \mathfrak{B}$ with $C \subseteq B_1 \cap B_2$.

(iv) For any $B \in \mathfrak{B}$, there is a $C \in \mathfrak{B}$ with $C \circ C \subseteq B$.

Conditions (ii)–(iv) can be combined into one condition:

(v) If $B_1, B_2 \in \mathfrak{B}$, there is a $C \in \mathfrak{B}$ with $C \circ C \subseteq (B_1 \cap B_1^{-1}) \cap (B_2 \cap B_2^{-1})$.

There is not a clean characterization for which collections of subsets of $X \times X$ are subbases for a uniformity (XIII.15.2.1.(iv) makes it clumsy to give a characterization). However, there is a useful set of sufficient conditions:

XIII.15.2.8. PROPOSITION. Let X be a set and $\mathfrak{S}$ a collection of subsets of $X \times X$. Then $\mathfrak{S}$ is a subbase for a uniformity on X (which is uniquely defined) if $\mathfrak{S}$ has the following properties:

(i) $\Delta \subseteq S$ for all $S \in \mathfrak{S}$.

(ii) If $S \in \mathfrak{S}$, then there is a $T \in \mathfrak{S}$ with $T \subseteq S^{-1}$.

(iii) For any $S \in \mathfrak{S}$, there is a $T \in \mathfrak{S}$ with $T \circ T \subseteq S$.

The union of any collection of uniformities (or subbases for uniformities) is a subbase for a uniformity.

XIII.15.2.9. A subset U of $X \times X$ is *symmetric* if $U = U^{-1}$ (i.e. the relation is symmetric). If $V \circ V \subseteq U$, then $V^{-1} \circ V^{-1} \subseteq U^{-1}$, and so $(V \cap V^{-1}) \circ (V \cap V^{-1}) \subseteq U \cap U^{-1}$. It follows that if $\mathfrak{B}$ is a base for a uniformity $\mathfrak{U}$, and we set

$$\mathfrak{B}_s = \{B \cap B^{-1} : B \in \mathfrak{B}\}$$

then $\mathfrak{B}_s$ is a base for $\mathfrak{U}$ consisting of symmetric sets. A similar statement holds for subbases. Thus we can almost entirely restrict attention to symmetric sets in the theory of uniform spaces.

The Uniformity of a Set of Pseudometrics

One of the most fundamental examples of a uniform structure is generated by a family of pseudometrics. In fact, it turns out (under a mild Choice hypothesis) that every uniformity is of this form.

XIII.15.2.10. Let X be a set and $\mathcal{P}$ a set of pseudometrics on X. For each $\rho \in \mathcal{P}$ and $r > 0$, set

$$V_r^\rho = \{(x, y) \in X \times X : \rho(x, y) < r\} .$$

If $\rho \in \mathcal{P}$ and $r, s > 0$, then $V_r^\rho \circ V_s^\rho \subseteq V_{r+s}^\rho$ by the triangle inequality, and in particular $V_{r/2}^\rho \circ V_{r/2}^\rho \subseteq V_r^\rho$ (this is motivation for including (iv) in definition XIII.15.2.1. and for calling it a "weak triangle inequality"). Thus

$$\mathfrak{S}_\mathcal{P} = \{V_r^\rho : \rho \in \mathcal{P}, r > 0\}$$

satisfies XIII.15.2.8.. It obviously satisfies (i) and (ii) (in fact, $[V_r^\rho]^{-1} = V_r^\rho$), so $\mathfrak{S}$ is a subbase for a unique uniformity $\mathfrak{U}_\mathcal{P}$ on X, called the *uniformity of* $\mathcal{P}$ or the *uniformity generated by* $\mathcal{P}$. If $\mathcal{P} = \{\rho\}$, write $\mathfrak{U}_\rho$ for $\mathfrak{U}_{\{\rho\}}$. In this case, $\mathfrak{S}_\rho$ is a base for $\mathfrak{U}_\rho$.

XIII.15.2.11. If $\mathcal{P}$ is a set of pseudometrics on a set X, σ is a pseudometric on X, and $\epsilon > 0$, if σ is uniformly dominated by $\mathcal{P}$ there are $\rho_1, \ldots, \rho_n \in \mathcal{P}$ and $\delta > 0$ such that $V_\delta^{\rho_1} \cap \cdots \cap V_\delta^{\rho_n} \subseteq V_\epsilon^\sigma$, and hence $V_\epsilon^\sigma \in \mathfrak{U}_\mathcal{P}$. So $\mathfrak{U}_\sigma \subseteq \mathfrak{U}_\mathcal{P}$. The converse is true by the same argument. It follows that if $\mathcal{Q}$ is another set of pseudometrics on X, then $\mathfrak{U}_\mathcal{Q} \subseteq \mathfrak{U}_\mathcal{P}$ if and only if $\mathcal{Q}$ is uniformly dominated by $\mathcal{P}$. Thus $\mathfrak{U}_\mathcal{Q} = \mathfrak{U}_\mathcal{P}$ if and only if $\mathcal{P}$ and $\mathcal{Q}$ are uniformly equivalent. In particular, $\mathfrak{U}_\mathcal{P} = \mathfrak{U}_{\bar{\mathcal{P}}}$, and $\bar{\mathcal{P}}$ is the unique saturated set of pseudometrics generating $\mathfrak{U}_\mathcal{P}$.

XIII.15.2.12. DEFINITION. If $\mathfrak{U}$ is a uniform structure on a set X, and $\bar{\mathcal{P}}$ is a saturated set of pseudometrics on X for which $\mathfrak{U} = \mathfrak{U}_{\bar{\mathcal{P}}}$, then $\bar{\mathcal{P}}$ is the unique saturated set with this property and is called the *gauge* of $\mathfrak{U}$ (some references spell this *gage*).

We will see later () that (under a mild Choice assumption) every uniform structure has a gauge.

The Topology of a Uniformity

If $\mathfrak{U}$ is a uniform structure on a set X, $\mathfrak{U}$ induces a compatible topology on X. If $S \subseteq X \times X$, for $x \in X$ we will use the notation $S[x]$ to denote $S(\{x\}) = \{y \in X : (x, y) \in S\}$.

XIII.15.2.13. DEFINITION. Let $\mathfrak{U}$ be a uniform structure on X. Let $\mathcal{T}_{\mathfrak{U}}$ be the collection of all sets V with the property that for every $x \in V$, there is a $U \in \mathfrak{U}$ with $U[x] \subseteq V$.

This definition is the analog of the definition of the topology of a metric space: a set is open if and only if it contains an open ball around each of its points. In fact, it is easily seen that if $\mathcal{P}$ is a set of pseudometrics on X, then $\mathcal{T}_{\mathfrak{U}_{\mathcal{P}}} = \mathcal{T}_{\mathcal{P}}$ as defined in XIII.15.1.2..

The condition can be equivalently expressed using a base $\mathfrak{B}$ for $\mathfrak{U}$: $V \in \mathcal{T}_{\mathfrak{U}}$ if and only if, for each $x \in V$, there is a $B \in \mathfrak{B}$ with $B[x] \subseteq V$.

XIII.15.2.14. PROPOSITION. Let $(X, \mathfrak{U})$ be a uniform space. Then

(i) $\mathcal{T}_{\mathfrak{U}}$ is a topology on X.

(ii) Every set in $\mathfrak{U}$ is a neighborhood of Δ with respect to the product topology on $X \times X$, and $\mathcal{T}_{\mathfrak{U}}$ is the smallest topology on X for which this holds.

XIII.15.2.15. Note though that the sets in $\mathfrak{U}$ are not generally open in the product topology, because of XIII.15.2.1.(v) (we use the convention from () that neighborhoods are not necessarily open).

XIII.15.2.16. PROPOSITION. Let $(X, \mathfrak{U})$ be a uniform space, and $\mathcal{T}_{\mathfrak{U}}$ the corresponding uniform topology. The following are equivalent:

(i) $\mathcal{T}_{\mathfrak{U}}$ is T_0.

(ii) $\mathcal{T}_{\mathfrak{U}}$ is Hausdorff.

(iii) $\mathfrak{U}$ is separating.

XIII.15.2.17. Let $\mathcal{T}$ be a topology on X. If there is a uniformity $\mathfrak{U}$ such that $\mathcal{T}_{\mathfrak{U}} = \mathcal{T}$, we say $\mathfrak{U}$ is an *admissible* uniform structure for $\mathcal{T}$ or that $\mathcal{T}$ *admits* the structure $\mathfrak{U}$. A topology which admits a uniform structure is called *uniformizable*. We will show () that a Hausdorff topology is uniformizable if and only if it is completely regular. Thus a fourth equivalent condition can be added to XIII.15.2.16.:

(iv) $\mathcal{T}_{\mathfrak{U}}$ is completely regular.

XIII.15.2.18. EXAMPLE. A (completely regular) topology admits many uniform structures in general. Here is a simple example. Let ρ be the discrete metric on $\mathbb{N}$ (), and $\mathfrak{U}_\rho$ the corresponding uniformity. Let σ be the metric on $\mathbb{N}$ with $\sigma(m,n) = \left|\frac{1}{m} - \frac{1}{n}\right|$. Then $\mathcal{T}_{\mathfrak{U}_\rho}$ and $\mathcal{T}_{\mathfrak{U}_\sigma}$ are both the discrete topology, i.e. the discrete topology admits both $\mathfrak{U}_\rho$ and $\mathfrak{U}_\sigma$ (in fact, the discrete topology on any set admits the discrete uniformity). But $\mathfrak{U}_\rho$ is the discrete uniformity and $\mathfrak{U}_\sigma$ is not, e.g. $\Delta \notin \mathfrak{U}_\sigma$. In this case, the discrete uniformity is obviously the largest admissible uniformity for the discrete topology, and $\mathfrak{U}_\sigma$ turns out to be the smallest admissible uniformity (XIII.15.10.20.); there are many other admissible uniformities for the discrete topology on $\mathbb{N}$, in fact $2^{2^{\aleph_0}}$ such uniformities (XIII.15.12.3.).

Thus uniform structures are more delicate than topologies in general. Every completely regular topology admits a largest uniformity (XIII.15.5.4.), but not generally a smallest uniformity (XIII.15.12.4.).

Open and Closed Entourages

Let $\mathfrak{U}$ be a uniform structure on a set X. We characterize entourages which are either open or closed for the product topology, and show there are enough of each to form a base for $\mathfrak{U}$.

XIII.15.2.19. LEMMA. Let X be a set, and $A, B \subseteq X \times X$. If A is symmetric, then

$$A \circ B \circ A = \bigcup \{A[x] \times A[y] : (x,y) \in B\} .$$

PROOF:

$$\begin{aligned} A \circ B \circ A &= \{(u,v) \in X \times X : \exists (x,y) \in B \text{ with } (u,x) \in A \text{ and } (y,v) \in A\} \\ &= \{(u,v) \in X \times X : \exists (x,y) \in B \text{ with } u \in A[x] \text{ and } v \in A[y]\} \end{aligned}$$

since A is symmetric. 👍

XIII.15.2.20. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space and $\mathfrak{U}$ an admissible uniformity for $\mathcal{T}$. If $A \subseteq X$, then

$$A^\circ = B := \{x \in X : \exists U \in \mathfrak{U} \text{ with } U[x] \subseteq A\} .$$

PROOF: We have $B \subseteq A$, and every open subset of A is contained in B, so we need only show that B is open. Let $x \in B$. Then there is a $U \in \mathfrak{U}$ with $U[x] \subseteq A$, and $V \in \mathfrak{U}$ with $V \circ V \subseteq U$. If $y \in V[x]$, then

$$V[y] \subseteq (V \circ V)[x] \subseteq U[x] \subseteq A$$

so $y \in B$. 👍

XIII.15.2.21. COROLLARY. Let $(X, \mathcal{T})$ be a topological space and $\mathfrak{U}$ an admissible uniformity for $\mathcal{T}$. If $\mathfrak{B}$ is a base for $\mathfrak{U}$, then for each $x \in X$ the set

$$\{B[x] : B \in \mathfrak{B}\}$$

is a local base for $\mathcal{T}$ at x.

XIII.15.2.22. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space and $\mathfrak{U}$ an admissible uniformity for $\mathcal{T}$. Give $X \times X$ the product topology. Then for each $U \in \mathfrak{U}$ we have $U^\circ \in \mathfrak{U}$. Thus the symmetric open sets in $\mathfrak{U}$ form a base for $\mathfrak{U}$.

PROOF:

$$U^\circ = \{(x, y) \in X \times X : \exists V, W \in \mathfrak{U} \text{ with } V[x] \times W[y] \subseteq U\} .$$

Since $V \cap W \in \mathfrak{U}$ for any such V and W, we have

$$U^\circ = \{(x, y) \in X \times X : \exists V \in \mathfrak{U} \text{ with } V[x] \times V[y] \subseteq U\} .$$

There is a symmetric $V \in \mathfrak{U}$ with $V \circ V \circ V \subseteq U$, and

$$V \circ V \circ V = \bigcup \{V[x] \times V[y] : (x, y) \in V\}$$

by Lemma XIII.15.2.19.. Thus $V \subseteq U^\circ$, and since $V \in \mathfrak{U}$ we have $U^\circ \in \mathfrak{U}$. 👍

XIII.15.2.23. PROPOSITION. Let $(X, \mathcal{T})$ be a topological space and $\mathfrak{U}$ an admissible uniformity for $\mathcal{T}$.
(i) If $A \subseteq X$, then

$$\bar{A} = \bigcap \{U(A) : U \in \mathfrak{U}\} .$$

(ii) If $X \times X$ has the product topology, and $E \subseteq X \times X$, then

$$\bar{E} = \bigcap \{U \circ E \circ U : U \in \mathfrak{U}\} .$$

PROOF: (i) If $x \in X$, then $x \in \bar{A}$ if and only if $U[x] \cap A \neq \emptyset$ for every symmetric $U \in \mathfrak{U}$. But $U[x] \cap A \neq \emptyset$ if and only if $x \in U^{-1}(A) = U(A)$.
(ii) If $(x, y) \in X \times X$, then $(x, y) \in \bar{E}$ if and only if $(U[x] \times U[y]) \cap E \neq \emptyset$ for every symmetric $U \in \mathfrak{U}$. If $U \in \mathfrak{U}$ is symmetric, then

$$(U[x] \times U[y]) \cap E \neq \emptyset \iff (x, y) \in U[z] \times U[w] \text{ for some } (z, w) \in E$$

$$\iff (x, y) \in \bigcup \{U[z] \times U[w] : (z, w) \in E\} \iff (x, y) \in U \circ E \circ U$$

by Lemma XIII.15.2.19.. Thus $\bar{E} = \bigcap \{U \circ E \circ U : U \in \mathfrak{U}\}$. 👍

XIII.15.2.24. Corollary. Let $(X, \mathcal{T})$ be a topological space and $\mathfrak{U}$ an admissible uniformity for $\mathcal{T}$. Give $X \times X$ the product topology. Then the symmetric closed sets in $\mathfrak{U}$ form a base for $\mathfrak{U}$.

Proof: Let $U \in \mathfrak{U}$. Then there is a symmetric $V \in \mathfrak{U}$ with $V \circ V \circ V \subseteq U$. Then $\bar{V} \subseteq V \circ V \circ V \subseteq U$, and $\bar{V} \in \mathfrak{U}$ since $V \in \mathfrak{U}$ and $V \subseteq \bar{V}$. 👍

XIII.15.3. The Covering Approach

A third approach to uniform spaces was developed by J. Tukey [Tuk40], using sets of "uniform" open covers of X instead of entourages or pseudometrics.

XIII.15.3.1. Definition. Let X be a set. A *covering uniformity* of X is a family $\mathscr{U}$ of coverings of X, with the properties that

(i) If $\mathcal{U}, \mathcal{V} \in \mathscr{U}$, then $\mathcal{U} \vee \mathcal{V} = \{U \cap V : U \in \mathcal{U}, V \in \mathcal{V}\} \in \mathscr{U}$.

(ii) If $\mathcal{U} \in \mathscr{U}$, there is a $\mathcal{V} \in \mathscr{U}$ which star-refines $\mathcal{U}$ ().

(iii) If $\mathcal{U}$ is a cover of X and there is a $\mathcal{V} \in \mathscr{U}$ which refines $\mathcal{U}$, then $\mathcal{U} \in \mathscr{U}$.

A covering in $\mathscr{U}$ is called a *uniform covering* or *even covering* of X. Conditions (i) and (ii) can be combined into

(iv) If $\mathcal{U}, \mathcal{V} \in \mathscr{U}$, there is a $\mathcal{W} \in \mathscr{U}$ which star-refines both $\mathcal{U}$ and $\mathcal{V}$.

Note that X is not assumed to have a topology, so the coverings are not "open coverings" (in fact this does not make sense in general); but see ().

XIII.15.3.2. Example. Here is the prototype example. Let (X, ρ) be a metric space, and let $\mathscr{U}_\rho$ be the collection of all coverings of X with positive Lebesgue number (), i.e. all coverings $\mathcal{U}$ such that there is an $\epsilon > 0$ for which, for every $x \in X$, there is a $U \in \mathcal{U}$ with $B^\rho_\epsilon(x) \subseteq U$. The covering of X by all open balls of radius ϵ is such a cover. If $\mathcal{U}$ is a cover with Lebesgue number ϵ, then the covering of X by all open balls of radius $\epsilon/2$ star-refines $\mathcal{U}$. Thus $\mathscr{U}_\rho$ is a uniform covering of X.

Bases

XIII.15.3.3. A collection $\mathscr{B}$ of covers of a set X is a *base* for a uniform covering $\mathscr{U}$ of X if $\mathscr{B} \subseteq \mathscr{U}$ and every $\mathcal{U} \in \mathscr{U}$ is refined by some $\mathcal{B} \in \mathscr{B}$. Subbases can be defined similarly. A collection $\mathscr{B}$ of covers of X is a base for a uniform cover if and only if

For every $\mathcal{B}_1, \mathcal{B}_2 \in \mathscr{B}$ there is a $\mathcal{C} \in \mathscr{B}$ which star-refines both $\mathcal{B}_1$ and $\mathcal{B}_2$.

The Topology of a Uniform Covering

XIII.15.3.4. If $\mathscr{B}$ is a base for a uniform covering $\mathcal{U}$ of a set X, then

$$\bigcup_{\mathcal{B}\in\mathscr{B}} \mathcal{B}$$

is a base for a topology $\mathcal{T}_{\mathscr{B}}$ on X. If $\mathscr{C}$ is another base for $\mathscr{U}$, then $\mathcal{T}_{\mathscr{C}} = \mathcal{T}_{\mathscr{B}}$; in particular, $\mathcal{T}_{\mathscr{B}} = \mathcal{T}_{\mathscr{U}}$ for any base $\mathscr{B}$ for $\mathscr{U}$. This is called the *uniform topology* for $\mathscr{U}$.

XIII.15.3.5. Example. If (X, ρ) is a metric space and $\mathscr{U}_\rho$ is the uniform cover of XIII.15.3.2., then $\mathcal{T}_{\mathscr{U}_\rho} = \mathcal{T}_\rho$ is the metric topology.

Covering Uniformities vs. Entourage Uniformities

There is a natural one-one correspondence between covering uniformities on a set X and uniformities on X in the sense of ().

XIII.15.3.6. If $\mathfrak{U}$ is a uniformity on a set X, then for each $U \in \mathfrak{U}$ the set

$$\mathcal{U} = \{U[x] : x \in X\}$$

is a covering of X. If $V \in \mathfrak{U}$, then the condition $V \circ V \subseteq U$ exactly corresponds to the condition that $\mathcal{V}$ star-refines $\mathcal{U}$. Thus if $\mathscr{B}_{\mathfrak{U}} = \{\mathcal{U} : U \in \mathfrak{U}\}$, then $\mathscr{B}_{\mathfrak{U}}$ is a base for a covering uniformity $\mathscr{U}(\mathfrak{U})$ on X. (The overuse of the letter U here in various fonts is somewhat tongue-in-cheek but serves a serious purpose, and thus does not violate the admonition of (). We write $\mathscr{U}(\mathfrak{U})$ instead of $\mathscr{U}_{\mathfrak{U}}$ to avoid multiple subscripts below.)

XIII.15.3.7. Conversely, let $\mathscr{U}$ be a covering uniformity on X. For each $\mathcal{U} \in \mathscr{U}$, set

$$B_{\mathcal{U}} = \bigcup_{U\in\mathcal{U}} (U \times U) \subseteq X \times X \ .$$

Then $B_{\mathcal{U}}$ is a symmetric subset of $X \times X$ containing Δ. If $\mathfrak{B}_{\mathscr{U}} = \{B_{\mathcal{U}} : \mathcal{U} \in \mathscr{U}\}$, then $\mathfrak{B}_{\mathscr{U}}$ is a base for a uniformity $\mathfrak{U}(\mathscr{U})$.

XIII.15.3.8. Proposition. (i) If $\mathfrak{U}$ is a uniformity on a set X and $\mathscr{U}(\mathfrak{U})$ the corresponding covering uniformity, then $\mathfrak{U}(\mathscr{U}(\mathfrak{U})) = \mathfrak{U}$.

(ii) If $\mathscr{U}$ is a covering uniformity on a set X and $\mathfrak{U}(\mathscr{U})$ the corresponding uniformity, then $\mathscr{U}(\mathfrak{U}(\mathscr{U})) = \mathscr{U}$.

Thus there is a natural one-one correpondence between uniformities and covering uniformities.

(iii) If $\mathfrak{U}$ and $\mathscr{U}$ are corresponding (i.e. $\mathfrak{U} = \mathfrak{U}(\mathscr{U})$ and hence $\mathscr{U} = \mathscr{U}(\mathfrak{U})$), then $\mathcal{T}_{\mathfrak{U}} = \mathcal{T}_{\mathscr{U}}$.

XIII.15.3.9. Thus all the notions discussed in the previous section have exact translations into covering uniformity language. For example, a topological space $(X, \mathcal{T})$ is uniformizable if and only if there is a covering uniformity $\mathscr{U}$ on X with $\mathcal{T}_{\mathscr{U}} = \mathcal{T}$.

XIII.15.4. Comparing the Definitions

We have seen (XIII.15.3.8.) that the entourage and covering approaches are equivalent. In this section, under a mild Choice assumption, we show they are also equivalent to the pseudometric approach. Thus one is free to use whichever approach is most convenient or more to the author's taste: many authors including Bourbaki [**?**] and Kelley [Kel75] use the entourage approach, Gillman and Jerison [GJ76] the pseudometric approach, and Isbell [Isb64] the covering approach (Isbell calls Tukey's covering approach "heretical" and then proceeds to use it).

In comparing the entourage and pseudometric approaches, we must show that every uniform structure is induced by a family of pseudometrics (i.e. has a gauge). The key technical fact is the following lemma, due to Weil [**?**], which is reminiscent of Urysohn's Lemma (). See [Kel75, 6.12] for a slightly more general version.

XIII.15.4.1. Lemma. [Metrization Lemma] Let X be a set, and $\{U_n : n \geq 0\}$ a sequence of subsets of $X \times X$ such that

$U_0 = X \times X$.

Each U_n is symmetric and contains Δ.

$U_{n+1} \circ U_{n+1} \circ U_{n+1} \subseteq U_n$ for each n.

Then there is a pseudometric ρ on X such that

$$U_n \subseteq \{(x, y) \in X \times X : \rho(x, y) < 2^{-n}\} \subseteq U_{n-1} \text{ for all } n \in \mathbb{N} .$$

Proof: Note that $U_n \subseteq U_{n-1}$ for all $n \in \mathbb{N}$. As a first approximation to ρ, define $f : X \times X \to [0, 1]$ by setting $f(x, y) = 2^{-n}$ if $(x, y) \in U_{n-1} \setminus U_n$ $(n \in \mathbb{N})$, and $f(x, y) = 0$ if $(x, y) \in \cap_n U_n$. We have $f(x, y) = f(y, x)$ for all (x, y).

Now for $(x, y) \in X \times X$ set

$$\rho(x, y) = \inf \left\{ \sum_{j=0}^{n} f(x_j, x_{j+1}) \right\}$$

where the infimum is over all n and finite sequences $(x_0, x_1, \ldots, x_{n+1})$ of points of X with $x_0 = x$ and $x_{n+1} = y$. Then $\rho(x, y) = \rho(y, x)$ for all (x, y) by reversing sequences, and ρ satisfies the triangle inequality by concatenating sequences. Thus ρ is a pseudometric. Since $\rho(x, y) \leq f(x, y)$ for all (x, y), $\rho(x, y) < 2^{-n}$ for $(x, y) \in U_n$, i.e. $U_n \subseteq \{(x, y) \in X \times X : \rho(x, y) < 2^{-n}\}$ for each n.

To complete the proof, we show by complete induction on n that for any sequence $(x_0, x_1, \ldots, x_{n+1})$ we have

$$f(x_0, x_{n+1}) \leq 2 \sum_{j=0}^{n} f(x_j, x_{j+1})$$

from which it follows that if $\rho(x, y) < 2^{-n}$ $(n \in \mathbb{N})$, then $f(x, y) < 2^{-n+1}$ and $(x, y) \in U_{n-1}$. So let $(x_0, x_1, \ldots, x_{n+1})$ be such a sequence, and set

$$a = \sum_{j=0}^{n} f(x_j, x_{j+1}) .$$

Let k be the largest index such that

$$f(x_0, x_1) + \cdots + f(x_{k-1}, x_k) \leq \frac{a}{2}$$

(we could have $k = 0$ if $f(x_0, x_1) > \frac{a}{2}$). Then $k \leq n$. If $k < n$, we have

$$f(x_{k+1}, x_{k+2}) + \cdots + f(x_n, x_{n+1}) \leq \frac{a}{2}$$

and in any event we have $f(x_k, x_{k+1}) \leq a$. By the inductive hypothesis we have $f(x_0, x_k) \leq 2\left(\frac{a}{2}\right) = a$ and $f(k+1, n+1) \leq a$ (these hold even if $k = 0$ or $k = n$).
Let m be the smallest natural number such that $2^{-m} \leq a$. Then $f(x_0, x_k)$, $f(x_k, x_{k+1})$, $f(x_{k+1}, x_{n+1})$ are all $\leq 2^{-m}$, so (x_0, x_k), (x_k, x_{k+1}), (x_{k+1}, x_{n+1}) are all in U_m. Since $U_m \circ U_m \circ U_m \subseteq U_{m-1}$, $(x_0, x_{n+1}) \in U_{m-1}$ and $f(x_0, x_{n+1}) \leq 2^{-m+1} \leq 2a$. 👍

XIII.15.4.2. THEOREM. Let $(X, \mathfrak{U})$ be a uniform space. Then there is a set $\mathcal{P}$ of pseudometrics on X such that $\mathfrak{U} = \mathfrak{U}_{\mathcal{P}}$, i.e. $\mathfrak{U}$ has a gauge.

PROOF: Let $\mathcal{P}$ be the set of all pseudometrics ρ on X with $\mathfrak{U}_\rho \subseteq \mathfrak{U}$. Then $\mathfrak{U}_{\mathcal{P}} \subseteq \mathfrak{U}$. Conversely, if $U = U^{-1} \in \mathfrak{U}$, set $U_0 = X \times X$, $U_1 = U$, and inductively choose $U_n \in \mathfrak{U}$ with $U_n = U_n^{-1}$ and $U_n \circ U_n \circ U_n \subseteq U_{n-1}$ for each n. Apply XIII.15.4.1. to get ρ. Then $\rho \in \mathcal{P}$, and $U \in \mathfrak{U}_\rho \subseteq \mathfrak{U}_{\mathcal{P}}$. Thus $\mathfrak{U} \subseteq \mathfrak{U}_{\mathcal{P}}$. 👍

XIII.15.4.3. Although no Choice is needed for the Metrization Lemma XIII.15.4.1., the Axiom of Dependent Choice is used in the proof of XIII.15.4.2.. Some form of Choice is necessary for this result: if $\mathfrak{U}$ is a uniformity on X and $U \in \mathfrak{U}$, without at least the Countable AC there may be no sequence (U_n) in $\mathfrak{U}$ satisfying the hypotheses of XIII.15.4.1. with $U_1 = U \cap U^{-1}$. In fact, there is a counterexample (XIII.15.10.2.). Few if any of the standard references on uniform spaces acknowledge or recognize this point.

However, no Choice is needed to prove the following crucial special case. This result was essentially proved in [AU23], with the proof simplified in [Fri37], before uniform spaces were defined, and was explicitly stated in [Wei37].

XIII.15.4.4. THEOREM. Let $(X, \mathcal{T})$ be a topological space. Then $(X, \mathcal{T})$ is metrizable if and only if there is a separated uniformity $\mathfrak{U}$ with a countable base which is admissible for $\mathcal{T}$.

PROOF: If $(X, \mathcal{T})$ is metrizable by ρ, then $\mathfrak{U}_\rho$ is a separated uniformity with countable base admissible for $\mathcal{T}$. Conversely, suppose $\mathfrak{U}$ is a separated uniformity admissible for $\mathcal{T}$ with a countable base $\mathfrak{B} = \{B_k : k \in \mathbb{N}\}$. For each k set $V_k = B_1 \cap B_1^{-1} \cap \cdots \cap B_k \cap B_k^{-1}$. Then $\{V_k : k \in \mathbb{N}\}$ is a base for $\mathfrak{U}$ consisting of a decreasing sequence of symmetric sets. Set $U_0 = X \times X$, $U_1 = V_1$, and inductively let U_{n+1} be the first V_k such that $V_k \circ V_k \circ V_k \subseteq U_n \cap V_n$. The U_n are a base for $\mathfrak{U}$ and satisfy the hypotheses of XIII.15.4.1., so there is a corresponding pseudometric ρ. Then $\mathfrak{U}_\rho \subseteq \mathfrak{U}$, and the U_n are all in $\mathfrak{U}_\rho$, so $\mathfrak{U} = \mathfrak{U}_\rho$. Since $\mathfrak{U}$ is separated, we have that $\cap_n U_n = \cap_k V_k = \Delta$, so ρ is a metric. 👍

XIII.15.5. Pullbacks and Fine Uniformities

Pullbacks of Uniformities

XIII.15.5.1. Let $\{(X_i, \mathfrak{U}_i) : i \in I\}$ be a set of uniform spaces, Y a set, and $f_i : Y \to X_i$ a function for all i. For each i set

$$\mathfrak{S}_i = f_i^{-1}(\mathfrak{U}_i) := \{(f_i \times f_i)^{-1}(U) : U \in \mathfrak{U}_i\} \subseteq Y \times Y$$

and set $\mathfrak{S} = \cup_i \mathfrak{S}_i$. Then $\mathfrak{S}$ is a subbase for a uniformity $\mathfrak{V}$ on Y, which is the smallest uniformity on Y making all the f_i uniformly continuous. The uniformity $\mathfrak{V}$ is called the *pullback* of the $\mathfrak{U}_i$ under the f_i (it is not in general a pullback in the category theory sense ()).

Suprema of Uniformities

Unions and intersections of uniformities are not uniformities in general. However, unions behave reasonably and as a consequence, if $(X, \mathcal{T})$ is a uniformizable topological space, there is always a largest uniformity on X admissible for $\mathcal{T}$.

XIII.15.5.2. If $\{\mathfrak{U}_i : i \in I\}$ is a nonempty set of uniformities on a set X, then $\mathfrak{B} = \bigcup_{i\in I} \mathfrak{U}_i$ is a subbase for a uniformity $\mathfrak{U} = \bigvee_{i\in I} \mathfrak{U}_i$ on X. This is the smallest uniformity on X finer than all the $\mathfrak{U}_i$ and is the *supremum* or *join* of the $\mathfrak{U}_i$. The uniformity $\bigvee_{i\in I} \mathfrak{U}_i$ is the pullback of the $\mathfrak{U}_i$ where Y and all the X_i are the same set X and the f_i are all the identity map.

XIII.15.5.3. Proposition. If $\{\mathfrak{U}_i : i \in I\}$ is a nonempty set of uniformities on a set X, and $\mathfrak{U} = \bigvee_{i\in I} \mathfrak{U}_i$, then $\mathcal{T}_{\mathfrak{U}}$ is the weakest topology $\bigvee_{i\in I} \mathcal{T}_{\mathfrak{U}_i}$ on X stronger than all the $\mathcal{T}_{\mathfrak{U}_i}$. In particular, if all the $\mathfrak{U}_i$ are admissible for a topology $\mathcal{T}$, then $\mathfrak{U}$ is also admissible for $\mathcal{T}$.

Fine Uniformities

XIII.15.5.4. Corollary. Let $(X, \mathcal{T})$ be a topological space. If $(X, \mathcal{T})$ is uniformizable, then there is a largest uniformity $\mathfrak{U}_f(\mathcal{T})$ admissible for $\mathcal{T}$, called the *fine uniformity* for $\mathcal{T}$.

Proof: Apply XIII.15.5.3. to the set of all uniformities on X admissible for $\mathcal{T}$ (note this is indeed a set, a subset of $\mathcal{P}(\mathcal{P}(X \times X))$).

XIII.15.5.5. In particular, if $(X, \mathcal{T})$ is completely regular, there is a fine uniformity for $\mathcal{T}$ (the converse is also true under a mild Choice assumption). Note, however, that the fine uniformity for $\mathcal{T}$ does not consist of all neighborhoods of Δ for the product topology in general (XIII.15.12.6.). It also is also not generally the same as the uniformity $\mathfrak{C}$ correspond to the set $\mathcal{C}$ of XIII.15.1.28.. 

Infima of Uniformities

XIII.15.5.6. In contrast, intersections of uniformities do not work reasonably. A uniformizable topological space $(X, \mathcal{T})$ does not have a smallest (coarsest) uniformity admissible for $\mathcal{T}$ in general; in fact, there is a smallest admissible uniformity if and only if $(X, \mathcal{T})$ is locally compact (XIII.15.10.20.).

However, if $\{\mathfrak{U}_i : i \in I\}$ is a set of uniformities on X, there is a largest uniformity $\bigwedge_{i\in I} \mathfrak{U}_i$ contained in $\bigcap_{i\in I} \mathfrak{U}_i$, namely the supremum (join) of all uniformities contained in this intersection. But if $\mathfrak{U}$ and $\mathfrak{V}$ are uniformities on X, $\mathcal{T}_{\mathfrak{U}\wedge\mathfrak{V}}$ is not necessarily $\mathcal{T}_{\mathfrak{U}} \wedge \mathcal{T}_{\mathfrak{V}}$; if $\mathfrak{U}$ and $\mathfrak{V}$ are admissible for a completely regular $\mathcal{T}$, $\mathfrak{U} \wedge \mathfrak{V}$ can even be the indiscrete uniformity (XIII.15.12.4.).

Products of Uniform Spaces

Another important special case of the pullback is product uniformities.

XIII.15.5.7. Let $(X_i, \mathfrak{U}_i)$ be a set of uniform spaces. Set $X = \prod_i X_i$. Assume $X \neq \emptyset$ (this is automatic under AC, but we need no Choice for this construction). For each i let $\pi_i : X \to X_i$ be the i'th coordinate projection. The pullback of the $\mathfrak{U}_i$ to X via π_i is called the *product uniformity* on X. The construction in () is a special case (the general case under a mild Choice assumption).

XIII.15.5.8. Proposition. Let $(X_i, \mathfrak{U}_i)$ be a set of uniform spaces, $X = \prod_i X_i \neq \emptyset$, and $\mathfrak{U}$ the product uniformity on X. Then $\mathcal{T}_{\mathfrak{U}}$ is the product topology on X.

XIII.15.5.9. Corollary. An arbitrary (nonempty) product of uniformizable spaces is uniformizable.

XIII.15.6. Completeness

Completeness is an important notion in metric spaces. The property extends naturally to uniform spaces and can be characterized in various ways; two of the most important is via Cauchy nets and Cauchy filters.

Cauchy Nets

In analogy with Cauchy sequences, a Cauchy net in a uniform space is a net in which the terms are all "close" far enough out in the net. Formally:

XIII.15.6.1. Definition. Let $(X, \mathfrak{U})$ be a uniform space. A net $(x_i)_{i\in I}$ in X is a $\mathfrak{U}$-*Cauchy net* if for every $U \in \mathfrak{U}$ there is an i_0 such that $(x_i, x_j) \in U$ for all $i, j > i_0$.

XIII.15.6.2. Note that if $(X, \mathfrak{U})$ is a uniform space, a net (x_i) converges to x in $(X, \mathcal{T}_{\mathfrak{U}})$ if and only if for every $U \in \mathfrak{U}$ there is an i_0 such that $(x_i, x) \in U$ for all $i > i_0$. Thus every convergent net is a $\mathfrak{U}$-Cauchy net. In both this convergence criterion and XIII.15.6.1. it suffices to let U range over a base or even a subbase for $\mathfrak{U}$.

XIII.15.6.3. Definition. A uniform space $(X, \mathfrak{U})$ is *complete* if every $\mathfrak{U}$-Cauchy net converges (in $(X, \mathcal{T}_{\mathfrak{U}})$) to a point of X.

XIII.15.6.4. While convergence of a net is a topological notion depending (in this context) only on $\mathcal{T}_{\mathfrak{U}}$, the property of being a Cauchy net is a uniform property: if $\mathfrak{U}$ and $\mathfrak{V}$ are uniformities admissible for a topology $\mathcal{T}$, a $\mathfrak{U}$-Cauchy net is not necessarily a $\mathfrak{V}$-Cauchy net (). However, if $\mathfrak{U} \subseteq \mathfrak{V}$, then every $\mathfrak{V}$-Cauchy net is a $\mathfrak{U}$-Cauchy net. In particular, every net in X which is Cauchy for the fine uniformity for $\mathcal{T}$ is $\mathfrak{U}$-Cauchy for every uniformity $\mathfrak{U}$ admissible for $\mathcal{T}$.

XIII.15.6.5. Definition. A uniformizable space $(X, \mathcal{T})$ is *topologically complete* if there is a uniformity $\mathfrak{U}$ on X admissible for $\mathcal{T}$ such that $(X, \mathfrak{U})$ is complete.

XIII.15.6.6. If $(X, \mathcal{T})$ is a topological space and $\mathfrak{U}$ and $\mathfrak{V}$ are uniformities on X admissible for $\mathcal{T}$, with $\mathfrak{U} \subseteq \mathfrak{V}$, and $(X, \mathfrak{U})$ is complete, then by XIII.15.6.4. $(X, \mathfrak{V})$ is complete. In particular, $(X, \mathcal{T})$ is topologically complete if and only if $(X, \mathfrak{U}_f(\mathcal{T}))$ is complete.

Topologically complete spaces are discussed further in ().

The definition of a Cauchy net can be cleanly rephrased in terms of covering uniformities or pseudometrics:

XIII.15.6.7. Proposition. Let $(X, \mathfrak{U})$ be a uniform space, and $\mathscr{U}$ the corresponding covering uniformity. A net $(x_i)_{i \in I}$ in X is a $\mathfrak{U}$-Cauchy net if and only if, for every $\mathcal{U} \in \mathscr{U}$ there is an i_0 and a $U \in \mathcal{U}$ such that x_i and x_j are both in U for all $i, j > i_0$. We will call a net with this property a *$\mathscr{U}$-Cauchy net*.

XIII.15.6.8. Proposition. Let X be a set, $\mathcal{P}$ a set of pseudometrics on X, and $\mathfrak{U}_{\mathcal{P}}$ the corresponding uniformity. A net $(x_i)_{i \in I}$ in X is a $\mathfrak{U}_{\mathcal{P}}$-Cauchy net if and only if, for every $\rho \in \mathcal{P}$ and $\epsilon > 0$ there is an i_0 such that $\rho(x_i, x_j) < \epsilon$ for all $i, j > i_0$. We will call a net with this property a *$\mathcal{P}$-Cauchy net*.

The following generalization of () holds:

XIII.15.6.9. Proposition. Let $(X, \mathcal{T})$ be a (uniformizable) topological space, and $\mathfrak{U}$ a uniformity on X admissible for $\mathcal{T}$. If (x_i) is a $\mathfrak{U}$-Cauchy net and (x_i) has a convergent subnet, then (x_i) converges.

XIII.15.6.10. Corollary. Let $(X, \mathcal{T})$ be a compact Hausdorff space, and $\mathfrak{U}$ a uniformity on X admissible for $\mathcal{T}$. Then $(X, \mathfrak{U})$ is complete.

In fact, if $(X, \mathcal{T})$ is a compact Hausdorff space, then $(X, \mathcal{T})$ is uniformizable (**??**) and there is a unique uniformity admissible for $\mathcal{T}$ (which is the fine uniformity and consists of all neighborhoods of Δ for the product topology). Thus $(X, \mathcal{T})$ is topologically complete.

Cauchy Filters

Because of the correspondence between nets and filters (), things can be rephrased in filter language.

XIII.15.6.11. DEFINITION. Let $(X, \mathfrak{U})$ be a uniform space. A family $\{A_i : i \in I\}$ of subsets of X, indexed by a directed set I, is *eventually small* (in $(X, \mathfrak{U})$) if for every $U \in \mathfrak{U}$ there is an i_0 such that $F_i \times F_i \subseteq U$ for all $i > i_0$.

XIII.15.6.12. DEFINITION. Let $(X, \mathfrak{U})$ be a uniform space. A $\mathfrak{U}$-*Cauchy filter* on $(X, \mathfrak{U})$ is a filter on X which is eventually small when indexed by reverse inclusion. A $\mathfrak{U}$-*Cauchy filterbase* is a filterbase for a Cauchy filter, i.e. an eventually small filterbase.

XIII.15.6.13. If (x_i) is a net in X, then the tail filter () of (x_i) is a $\mathfrak{U}$-Cauchy filter if and only if (x_i) is a $\mathfrak{U}$-Cauchy net.

XIII.15.6.14. PROPOSITION. (i) Let $\mathscr{U}$ be a covering uniformity on X, and $\mathfrak{U}$ the corresponding uniformity. A filter $\mathcal{F}$ on X is a $\mathfrak{U}$-Cauchy filter if and only if for every $\mathcal{U} \in \mathscr{U}$ there is a $U \in \mathcal{U}$ and $F \in \mathcal{F}$ with $F \subseteq U$. The same holds for filterbases. Thus a filter $\mathcal{F}$ on X is a Cauchy filter if and only if for every $\mathcal{U} \in \mathscr{U}$ there is a $U \in \mathcal{U}$ with $U \in \mathcal{F}$.

(ii) Let $\mathcal{P}$ be a set of pseudometrics on X. A filter $\mathcal{F}$ on X is a $\mathfrak{U}_{\mathcal{P}}$-Cauchy filter if and only if, for every $\rho \in \mathcal{P}$ and $\epsilon > 0$, there is a $U \in \mathfrak{U}$ such that $\rho(x, y) < \epsilon$ for all $(x, y) \in U$.

Cauchy Nests

XIII.15.6.15. DEFINITION. Let $(X, \mathfrak{U})$ be a uniform space. A $\mathfrak{U}$-*Cauchy nest* is an eventually small family $\{F_i : i \in I\}$ of nonempty closed subsets of X, directed by reverse inclusion, i.e. $F_j \subseteq F_i$ if $i < j$. A $\mathfrak{U}$-Cauchy nest is *regular* if each F_i is a regular closed set, i.e. the closure of its interior.

XIII.15.6.16. THEOREM. Let $(X, \mathcal{T})$ be a topological space and $\mathfrak{U}$ a uniformity admissible for $\mathcal{T}$. Then the following are equivalent:

(i) $(X, \mathfrak{U})$ is complete.

(ii) Whenever (F_i) is a $\mathfrak{U}$-Cauchy nest in X, then $\cap_{i \in I} F_i$ is nonempty.

(iii) Whenever (F_i) is a regular $\mathfrak{U}$-Cauchy nest in X, then $\cap_{i \in I} F_i$ is nonempty.

PROOF: (i) $\Rightarrow$ (ii): Let (F_i) be a $\mathfrak{U}$-Cauchy nest in X. Let $\mathcal{F}$ be the collection of all subsets of X containing some F_i. Then $\mathcal{F}$ is a filter since the F_i are nonempty and directed by reverse inclusion, and since the F_i are in $\mathcal{F}$ it is eventually small, i.e. a $\mathfrak{U}$-Cauchy filter. Thus $\mathcal{F}$ converges to an $x \in X$, and $x \in F_i$ for all i since F_i is closed.

(ii) $\Rightarrow$ (iii) is trivial.

(iii) $\Rightarrow$ (i): Let $\mathscr{U}$ be the covering uniformity corresponding to $\mathfrak{U}$. Suppose $\mathcal{F}$ is a $\mathfrak{U}$-Cauchy filter on X. For each $\mathcal{U} \in \mathscr{U}$ let $F_{\mathcal{U}}$ be the closure of the union of all $U \in \mathcal{U} \cap \mathcal{F}$ (there is such a U in each $\mathcal{U} \in \mathscr{U}$ since $\mathcal{F}$ is a Cauchy filter). Then $F_{\mathcal{U}}$ is a nonempty regular closed set in X. If $U_0 \in \mathcal{U} \cap \mathcal{F}$, then every $U \in \mathcal{U} \cap \mathcal{F}$ is in $St(U_0)$ since $\mathcal{F}$ is a filter. If $\mathcal{U}, \mathcal{V} \in \mathscr{U}$ and $\mathcal{W} \in \mathscr{U}$ star-star-refines both $\mathcal{U}$ and $\mathcal{V}$, then $F_{\mathcal{W}}$ is contained in some $U \in \mathcal{U}$ and in some $V \in \mathcal{V}$, and in particular $F_{\mathcal{W}}$ is contained in $F_{\mathcal{U}} \cap F_{\mathcal{V}}$. So the $F_{\mathcal{U}}$ are eventually

small and directed under reverse inclusion. Thus, by (iii), there is some $x \in \cap_{\mathcal{U}\in\mathscr{U}} F_{\mathcal{U}}$, and $\mathcal{F}$ must converge to x. 👍

XIII.15.7. Subspace Uniformities and Completeness

XIII.15.7.1. If $(X, \mathfrak{U})$ is a uniform space and $Y \subseteq X$, then

$$\mathfrak{U}_Y = \{U \cap (Y \times Y) : U \in \mathfrak{U}\}$$

is easily seen to be a uniformity on Y, which is separated if $\mathfrak{U}$ is separated. $\mathfrak{U}_Y$ is called the *relative uniformity* or *subspace uniformity* on Y. It is the smallest uniformity $\mathfrak{V}$ on Y such that the inclusion map from Y to X is $(\mathfrak{V}, \mathfrak{U})$-uniformly continuous.

XIII.15.7.2. Proposition. If $(X, \mathfrak{U})$ is a uniform space and $Y \subseteq X$, then $\mathcal{T}_{\mathfrak{U}_Y}$ is the relative topology on Y from $\mathcal{T}_{\mathfrak{U}}$. In particular, every subspace of a uniformizable space is uniformizable.

XIII.15.7.3. More generally, if $(X, \mathfrak{U})$ is a uniform space, Y a set, and $f : Y \to X$ is a function, then

$$\{(f \times f)^{-1}(U) : U \in \mathfrak{U}\}$$

is a base for a uniformity on Y (it is not a uniformity in general), which is the smallest uniformity on Y making f uniformly continuous. (This is the special case of a pullback (XIII.15.5.1.) where I is a singleton.)

There is a straightforward translation into the covering and pseudometric pictures:

XIII.15.7.4. Proposition. (i) Let $\mathscr{U}$ be a covering uniformity on X and $\mathfrak{U}$ the corresponding uniformity. Let $Y \subseteq X$. For each $\mathcal{U} \in \mathscr{U}$ set $\mathcal{U}_Y = \{U \cap Y : U \in \mathcal{U}\}$, and set $\mathscr{U}_Y = \{\mathcal{U}_Y : \mathcal{U} \in \mathscr{U}\}$. Then $\mathscr{U}_Y$ is a covering uniformity on Y, and is the covering uniformity on Y corresponding to $\mathfrak{U}_Y$.
(ii) Let $\mathcal{P}$ be a set of pseudometrics on a set X, and $Y \subseteq X$. Set $\mathcal{P}_Y = \{\rho|_{Y\times Y} : \rho \in \mathcal{P}\}$. Then $\mathcal{P}_Y$ is a set of pseudometrics on Y, and $\mathfrak{U}_{\mathcal{P}_Y} = (\mathfrak{U}_{\mathcal{P}})_Y$.

XIII.15.7.5. Proposition. Let $(X, \mathfrak{U})$ be a uniform space, and $Y \subseteq X$. Then a net (y_i) in Y is $\mathfrak{U}_Y$-Cauchy if and only if it is $\mathfrak{U}$-Cauchy in X.

Completeness and Subspaces

XIII.15.7.6. Proposition. Let $(X, \mathfrak{U})$ be a complete uniform space, and $Y \subseteq X$. Then
(i) If Y is $\mathcal{T}_{\mathfrak{U}}$-closed, then $(Y, \mathfrak{U}_Y)$ is complete.
(ii) If $(X, \mathfrak{U})$ is separated and $(Y, \mathfrak{U}_Y)$ is complete, then Y is $\mathcal{T}_{\mathfrak{U}}$-closed in X.

Proof: (i) Let (y_i) be a $\mathfrak{U}_Y$-Cauchy net. Then (y_i) is $\mathfrak{U}$-Cauchy by XIII.15.7.5., so $y_i \to y$ for some $y \in X$ since $(X, \mathfrak{U})$ is complete. Then $y \in Y$ since Y is $\mathcal{T}_{\mathfrak{U}}$-closed in X, i.e. (y_i) converges in Y.

(ii) Let $x \in \bar{Y}$, and let (y_i) be a net in Y with $y_i \to x$. Then (y_i) is a $\mathfrak{U}$-Cauchy net, so it is $\mathfrak{U}_Y$-Cauchy by XIII.15.7.5.. Since $(Y, \mathfrak{U}_Y)$ is complete, $y_i \to y$ for some $y \in Y$. We have $y = x$ since $\mathcal{T}_{\mathfrak{U}}$ is Hausdorff, so $x \in Y$ and Y is $\mathcal{T}_{\mathfrak{U}}$-closed. 👍

XIII.15.7.7. Example. Part (ii) fails if $(X, \mathfrak{U})$ is not separated.

(i) The simplest example is to take $\mathfrak{U}$ to be the indiscrete uniformity on a set X. Then $(X, \mathfrak{U})$ is trivially complete since every net in X converges. If Y is any subset of X, then $\mathfrak{U}_Y$ is the indiscrete uniformity on Y, hence $(Y, \mathfrak{U}_Y)$ is complete, but Y is not closed in X if it is a proper subset.

(ii) Let $(X, \mathfrak{U})$ be any complete uniform space which is not separated, and let $(x, y) \in \cap_{U \in \mathfrak{U}} U$, $x \neq y$. Then x and y have the same $\mathcal{T}_{\mathfrak{U}}$-neighborhoods in X, so a net in X converges to x if and only if it converges to y. If $Y = X \setminus \{x\}$, then $(Y, \mathfrak{U}_Y)$ is complete, but Y is not closed in X (it is dense).

Completeness of Products

XIII.15.7.8. Proposition. Let $\{(X_i, \mathfrak{U}_i) : i \in I\}$ be a set of separated uniform spaces, with $X = \prod_{i \in I} X_i \neq \emptyset$, and let $\mathfrak{U}$ be the product uniformity () on X. Then $(X, \mathfrak{U})$ is complete if and only if each $(X_i, \mathfrak{U}_i)$ is complete.

Proof: Let $\pi_i : X \to X_i$ be the i'th coordinate projection for each i. Then it is easily checked that (x_λ) is a $\mathfrak{U}$-Cauchy net if and only if $(\pi_i(x_\lambda))$ is a $\mathfrak{U}_i$-Cauchy net for all i. If (x_λ) is a $\mathfrak{U}$-Cauchy net in X, and each $(X_i, \mathfrak{U}_i)$ is complete, then for each i $\pi_i(x_\lambda) \to y_i$ for a unique $y_i \in X_i$. The function $i \mapsto y_i$ defines an element $y \in X$, and $x_\lambda \to y$. Thus $(X, \mathfrak{U})$ is complete. Conversely, if X is nonempty and $(X, \mathfrak{U})$ is complete, fix i and let (y_λ) be a $\mathfrak{U}_i$-Cauchy net in X_i. Let $z = (z_j) \in X$, and define a net (x_λ) in X by

$$(x_\lambda)_j = \begin{cases} y_\lambda & \text{if} \quad j = i \\ z_j & \text{if} \quad j \neq i \end{cases} .$$

Then (x_λ) is a $\mathfrak{U}$-Cauchy net in X, so $x_\lambda \to x$ for some x, and $y_\lambda \to \pi_i(x)$. 👍

The assumption $X \neq \emptyset$, automatic under AC, emphasizes that this is a theorem of ZF. The result is also true if the $(X_i, \mathfrak{U}_i)$ are not separated, but the proof requires the AC since the y_i are not necessarily unique (in fact the statement is equivalent to the AC in this case).

XIII.15.7.9. Corollary. An arbitrary (nonempty) product of topologically complete Hausdorff spaces is topologically complete. (If the AC is assumed, "Hausdorff" can be eliminated.)

XIII.15.7.10. Corollary. Let $\{\mathfrak{U}_i : i \in I\}$ be a family of separated uniformities on a set X, and $\mathfrak{U} = \bigvee_i \mathfrak{U}_i$. Then $(X, \mathfrak{U})$ is complete if and only if each $(X, \mathfrak{U}_i)$ is complete.

Proof: The diagonal subspace X is a closed subspace of $\prod_i (X, \mathfrak{U}_i)$, and $\mathfrak{U}$ is the relative uniformity on this subspace, so the result follows from XIII.15.7.8. and XIII.15.7.6.. 👍

XIII.15.7.11. This result is unclear if "separated" is remoived, even under the AC. Indeed, it is even unclear if $\mathfrak{U}$ and $\mathfrak{V}$ are uniformities on a set X with $\mathcal{T}_{\mathfrak{U}} = \mathcal{T}_{\mathfrak{V}}$, whether $(X, \mathfrak{U} \vee \mathfrak{V})$ is complete if $(X, \mathfrak{U})$ and $(X, \mathfrak{V})$ are complete if they are not separated.

XIII.15.8. Completions of Uniform Spaces

In this section we show that every separated uniform space has an essentially unique completion (non-separated ones have a completion also, but there is no uniqueness, although there is a "minimal" one, cf. XIII.15.12.7.). There are various ways to construct the completion; we will use a method based on Cantor's construction of $\mathbb{R}$ (III.5.2.) by taking points of the completion to be equivalence classes of Cauchy filters. This construction has the advantage, among other things, of avoiding any use of Choice.

XIII.15.8.1. Definition. Let $(X, \mathfrak{U})$ be a separated uniform space. A *completion* of $(X, \mathfrak{U})$ is a separated complete uniform space $(\bar{X}, \bar{\mathfrak{U}})$ and an embedding of X as a dense subspace of $\bar{X}$ such that $\mathfrak{U} = \bar{\mathfrak{U}}_X$.

Construction

We now fix a separated uniform space $(X, \mathfrak{U})$, and let $\mathscr{U}$ be the corresponding covering uniformity on X.

XIII.15.8.2. We first put an equivalence relation on the set $\mathscr{C}$ of $\mathfrak{U}$-Cauchy filters on X. If $\mathcal{F}, \mathcal{G} \in \mathscr{C}$, we say $\mathcal{F} \sim \mathcal{G}$ if for every $\mathcal{U} \in \mathscr{U}$ and $U \in \mathcal{U}$, $U \in \mathcal{F}$ if and only if $U \in \mathcal{G}$ (i.e. U contains an $F \in \mathcal{F}$ if and only if it contains a $G \in \mathcal{G}$). Note that if $\mathcal{F}$ converges to $x \in X$, then $\mathcal{G} \in \mathscr{C}$ satisfies $\mathcal{G} \sim \mathcal{F}$ if and only if $\mathcal{G}$ also converges to x, i.e. a Cauchy filter converges to x if and only if it is equivalent to the neighborhood filter of x.

XIII.15.8.3. Proposition. The relation $\sim$ is an equivalence relation on $\mathscr{C}$.

Proof: The relation $\sim$ is obviously symmetric and transitive, and is reflexive since the filters in $\mathscr{C}$ are Cauchy filters. 👍

XIII.15.8.4. We now take $\bar{X}$ to be the set $\mathscr{C}/\sim = \{[\mathcal{F}] : \mathcal{F} \in \mathscr{C}\}$, where $[\mathcal{F}]$ denotes the equivalence class of $\mathcal{F}$. There is an obvious embedding of X into $\bar{X}$ sending x to the equivalence class of filters converging to x (this map is injective since $(X, \mathfrak{U})$ is separated). We identify X with its image.

XIII.15.8.5. Put a uniform structure on $\bar{X}$ as follows. For each $\mathcal{U} \in \mathscr{U}$ and $U \in \mathcal{U}$, let

$$\bar{U} = \{[\mathcal{F}] : \mathcal{F} \in \mathscr{C} \text{ and } U \in \mathcal{F}\} \subseteq \bar{X}$$

(this is well defined since if $\mathcal{F} \sim \mathcal{G}$, then $U \in \mathcal{F}$ if and only if $U \in \mathcal{G}$). If $x \in X$ and $\mathcal{F}$ is the neighborhood filter of x, then $x = [\mathcal{F}] \in \bar{U}$ if and only if $x \in U$. Thus $\bar{U} \cap X = U$.

Set $\bar{\mathcal{U}} = \{\bar{U} : U \in \mathcal{U}\}$ and $\bar{\mathscr{U}} = \{\bar{\mathcal{U}} : \mathcal{U} \in \mathscr{U}\}$.

XIII.15.8.6. PROPOSITION. $\bar{\mathscr{U}}$ is a covering uniformity on $\bar{X}$, and $\bar{\mathscr{U}}_X = \mathscr{U}$. In particular, $\mathcal{T}_{\mathscr{U}}$ is the relative topology on X from $\mathcal{T}_{\bar{\mathscr{U}}}$, and X is dense in $\bar{X}$.

Let $\bar{\mathfrak{U}}$ be the corresponding uniformity on $\bar{X}$.

XIII.15.8.7. PROPOSITION. $(\bar{X}, \bar{\mathfrak{U}})$ is complete and separated, and thus is a completion of $(X, \mathfrak{U})$.

Uniqueness

We give a general result which implies uniqueness of completions in the separated case:

XIII.15.8.8. PROPOSITION. Let X be a set, and $\mathfrak{U}$ and $\mathfrak{V}$ uniformities on X with $\mathfrak{U} \subseteq \mathfrak{V}$. Suppose $(X, \mathfrak{U})$ is separated (so $(X, \mathfrak{V})$ is also separated). Let $(Y, \tilde{\mathfrak{U}})$ and $(Z, \tilde{\mathfrak{V}})$ be (separated) completions of $(X, \mathfrak{U})$ and $(X, \mathfrak{V})$ respectively. Then there is a unique map $\phi : Z \to Y$ which is continuous for $\mathcal{T}_{\tilde{\mathfrak{V}}}$ and $\mathcal{T}_{\tilde{\mathfrak{U}}}$ and is the identity on X, under the canonical identifications of X with (dense) subspaces of Y and Z. In addition, ϕ is $(\tilde{\mathfrak{V}}, \tilde{\mathfrak{U}})$-uniformly continuous.

PROOF: Uniqueness of ϕ is clear, since Y is Hausdorff and two such continuous functions agree on the dense subset X of Z. We show existence.
Let $z \in Z$, and let $\mathcal{N}_z$ be the $\mathcal{T}_{\tilde{\mathfrak{V}}}$-neighborhood filter of z in Z. Set

$$\mathcal{F}_z = \{V \cap X : V \in \mathcal{N}_z\} .$$

Since X is dense in Z, $\mathcal{F}_z$ is a filter on X. Because $\mathcal{N}_z$ is a $\tilde{\mathfrak{V}}$-Cauchy filter on Z and $\mathfrak{V} = \tilde{\mathfrak{V}}_X$, $\mathcal{F}_z$ is a $\mathfrak{V}$-Cauchy filter on X, hence a $\mathfrak{U}$-Cauchy filter on X. Since $\mathfrak{U} = \tilde{\mathfrak{U}}_X$, $\mathcal{F}_z$ is a $\tilde{\mathfrak{U}}$-Cauchy filterbase on Y. If $\mathcal{G}_z$ is the corresponding $\tilde{\mathfrak{U}}$-Cauchy filter on Y, then $\mathcal{G}_z$ converges to $y \in Y$, which is unique since Y is Hausdorff. Set $\phi(z) = y$. It is routine to check that ϕ has the desired properties. 👍

XIII.15.8.9. COROLLARY. Let $(X, \mathfrak{U})$ be a separated uniform space. Then the completion $(\bar{X}, \bar{\mathfrak{U}})$ of XIII.15.8.7. is the unique (separated) completion of $(X, \mathfrak{U})$, i.e. any completion is uniformly isomorphic to $(\bar{X}, \bar{\mathfrak{U}})$.

PROOF: Let $(Y, \tilde{\mathfrak{U}})$ and $(Z, \hat{\mathfrak{U}})$ be completions of $(X, \mathfrak{U})$. Apply XIII.15.8.8. twice with $\mathfrak{V} = \mathfrak{U}$ to obtain $\phi : Z \to Y$ and $\psi : Y \to Z$. Then $\phi \circ \psi$ and $\psi \circ \phi$ are the identity on X, hence the identity, so ϕ and ψ are uniform isomorphisms. 👍

Absolutely Closed Uniform Spaces

XIII.15.8.10. DEFINITION. Let $(X, \mathfrak{U})$ be a separated uniform space. Then $(X, \mathfrak{U})$ is *absolutely closed* if, whenever $(X, \mathfrak{U})$ is uniformly embedded in a separated uniform space $(Y, \mathfrak{V})$ (i.e. $X \subseteq Y$ and $\mathfrak{U} = \mathfrak{V}_X$), then X is closed in Y.

XIII.15.8.11. PROPOSITION. Let $(X, \mathfrak{U})$ be a separated uniform space. Then $(X, \mathfrak{U})$ is absolutely closed if and only if it is complete.

PROOF: $(X, \mathfrak{U})$ is uniformly embedded in $(\bar{X}, \bar{\mathfrak{U}})$, and if $(X, \mathfrak{U})$ is absolutely closed, then X is dense and closed in $\bar{X}$, hence $X = \bar{X}$ and $(X, \mathfrak{U})$ is complete. Conversely, suppose $(X, \mathfrak{U})$ is complete and uniformly embedded in $(Y, \mathfrak{V})$. Then $(X, \mathfrak{U})$ is also uniformly embedded in $(\bar{Y}, \bar{\mathfrak{V}})$, so X is closed in $\bar{Y}$ by XIII.15.7.6. and hence also closed in Y. Thus $(X, \mathfrak{U})$ is absolutely closed. 👍

XIII.15.9. Metric Spaces

In this section, we specialize to the case of metric spaces (or metrizable uniformizable spaces). We give some additional equivalent conditions for completeness in the sense of XIII.15.6.3..

XIII.15.9.1. We will not assume any form of Choice in this subsection, i.e. we will work in ZF set theory. It is useful to see that almost everything one would want to show, with a single notable exception, can be done without any Choice assumed.

The exception is that without the Countable AC we cannot simply work with Cauchy sequences.

XIII.15.9.2. DEFINITION. Let (X, ρ) be a metric space. then (X, ρ) is *uniformly complete* if $(X, \mathfrak{U}_\rho)$ is a complete uniform space (i.e. if every $\mathfrak{U}_\rho$-Cauchy net converges), and (X, ρ) is *Cauchy complete* if every ρ-Cauchy sequence in X converges (i.e. (X, ρ) is complete in the sense of ()).

There is one additional notion of completeness we will use:

XIII.15.9.3. DEFINITION. Let (X, ρ) be a metric space. Then (X, ρ) is *σ-complete* if, whenever (F_n) is a decreasing sequence of nonempty closed sets in X whose diameters approach zero, then $\cap_n F_n$ is nonempty (it is then necessarily a singleton). (X, ρ) is *regularly σ-complete* if, whenever (F_n) is a decreasing sequence of nonempty regular (XIII.15.6.15.) closed sets in X whose diameters approach zero, then $\cap_n F_n$ is nonempty.

One might think σ-completeness is the same as Cauchy completeness, and it is if the Countable AC is assumed; but they can be different otherwise (the difficulty is that without the Countable AC, a metric space may have "too few" Cauchy sequences, cf. XIII.15.12.8.). The completeness conditions are equivalent in the separable case without any Choice (XIII.15.9.7.).

XIII.15.9.4. Any closed subspace of a σ-complete metric space is obviously σ-complete. Conversely, a σ-complete subspace of a metric space is closed (consider a decreasing sequence of closed balls around a point of the closure).

XIII.15.9.5. PROPOSITION. Let (X, ρ) be a metric space.

(i) If (X, ρ) is σ-complete, then it is Cauchy complete.

(ii) If the Countable AC is assumed and (X, ρ) is Cauchy complete, then it is σ-complete.

PROOF: (i) Let (x_n) be a ρ-Cauchy sequence in X. For each n let F_n be the closure of the set $\{x_k : k \geq n\}$. Then (F_n) is a decreasing sequence of nonempty closed sets in X, whose diameters approach 0 since (x_n) is a Cauchy sequence. Thus $\cap_n F_n = \{x\}$ for some $x \in X$, and $x_n \to x$.

(ii) Let (F_n) be a decreasing sequence of nonempty closed sets whose diameters approach 0. Using Countable AC choose $x_n \in F_n$ for each n. Then (x_n) is a ρ-Cauchy sequence since $x_k \in F_n$ for all n and all $k \geq n$. Thus $x_n \to x$ for some $x \in X$, and $x \in \cap_n F_n$ since the F_n are closed. 👍

The following is a theorem of ZF, i.e. no Choice is assumed. If the Countable AC is assumed, then (vii) is equivalent to the other conditions (XIII.15.9.5.).

XIII.15.9.6. THEOREM. Let (X, ρ) be a metric space. The following are equivalent:

(i) (X, ρ) is absolutely closed.

(ii) (X, ρ) is σ-complete.

(iii) (X, ρ) is regularly σ-complete.

(iv) Whenever (F_i) is a $\mathfrak{U}_\rho$-Cauchy nest in X, then $\cap_{i \in I} F_i$ is nonempty.

(v) Whenever (F_i) is a regular $\mathfrak{U}_\rho$-Cauchy nest in X, then $\cap_{i \in I} F_i$ is nonempty.

(vi) (X, ρ) is uniformly complete.

These conditions imply

(vii) (X, ρ) is Cauchy complete.

PROOF: (iv) $\Rightarrow$ (ii) $\Rightarrow$ (iii) and (iv) $\Rightarrow$ (v) $\Rightarrow$ (iii) are trivial, (iv) $\iff$ (vi) is XIII.15.6.16., and (ii) $\Rightarrow$ (vii) is XIII.15.9.5.. (i) $\Rightarrow$ (vi) since (X, ρ) is then closed and dense in its uniform completion $(\bar{X}, \bar{\rho})$.

(iii) $\Rightarrow$ (i): Suppose (X, ρ) is not absolutely closed. Identify (X, ρ) with a nonclosed subspace of a metric space (Y, ρ'), and let $p \in \bar{X} \setminus X$. For each n let U_n be the intersection of X with the open ball $B_{1/n}$ (in Y) of radius $1/n$ around p, and F_n be the closure of U_n in X. Since p is a limit point of X, each F_n is nonempty, and is a regular closed set in X, and $F_{n+1} \subseteq F_n$, $F_{n+1} \subseteq B_{1/n}$. But

$$\bigcap_n F_n \subseteq X \cap \bigcap_n B_{1/n} = X \cap \{p\} = \emptyset$$

so (X, ρ) is not regularly σ-complete. 👍

XIII.15.9.7. PROPOSITION. Let (X, ρ) be a separable metric space. Then (X, ρ) is σ-complete if and only if it is Cauchy complete.

PROOF: One direction is XIII.15.9.5.. Conversely, suppose (X, ρ) is Cauchy complete. Fix a dense sequence (x_k) in X. Let (F_n) be a regular Cauchy nest in X. For each n let y_n be the first x_k with $x_k \in F_n^\circ$. Since $diam(F_n) \to 0$, (y_n) is a Cauchy sequence in X. If y is the limit, then $y \in \cap_n F_n$. 👍

The proof of the next result is adapted from [KT01]. We describe an algorithm for choosing a point from any nonempty open subset; the algorithm requires one choice of a countable base, but no further choice.

XIII.15.9.8. LEMMA. Let (X, ρ) be a second countable σ-complete metric space. Fix a countable base $\mathcal{U} = \{U_k : k \in \mathbb{N}\}$ for the topology of X consisting of nonempty sets. Then there is a recursive algorithm (depending on $\mathcal{U}$) for choosing an element from each nonempty open set in X.

PROOF: Let W be a nonempty open set in X. Recursively generate a decreasing sequence (F_n) of closed sets as follows. Let V_1 be the the first U_k for which $F_1 := \bar{U}_k \subseteq W$ and $diam(U_k) < 1$ (it is easily seen that there is such a U_k). If V_n and $F_n = \bar{V}_n$ have been defined, let V_{n+1} be the first U_k for which $F_{n+1} := \bar{U}_k \subseteq V_n$ and $diam(U_k) < \frac{1}{n+1}$. Then the F_n form a (regular) Cauchy nest in X. Let x be the unique point in the intersection. We have $x \in F_1 \subseteq W$. 👍

XIII.15.9.9. THEOREM. Let (X, ρ) be a σ-complete metric space. Then (X, ρ) is separable if and only if it is second countable.

PROOF: A separable metric space is second countable (balls of rational radii around a dense set of points form a base for the topology). The converse (indeed, the much more general result that every second countable topological space is separable) is trivial if the Countable AC is assumed. We give a proof in the σ-complete case which requires no Choice.

Fix a countable base $\mathcal{U} = \{U_k : k \in \mathbb{N}\}$ for the topology of X consisting of nonempty sets. For each k we define an $x_k \in U_k$ using the algorithm of XIII.15.9.8.. The sequence (x_k) is dense in X, so (X, ρ) is separable. 👍

XIII.15.9.10. Some completeness is needed here (in the absence of the Countable AC). COHEN constructed an infinite Dedekind-finite subset X of $\mathbb{R}$. X with the relative topology is metrizable and second countable, but not separable since X has no countably infinite subsets at all.

It is unclear whether XIII.15.9.9. holds just under Cauchy completeness. But combining it with XIII.15.9.7. (cf. XIII.15.9.4.), we obtain:

XIII.15.9.11. COROLLARY. Let (X, ρ) be a separable (Cauchy) complete metric space. Then every closed subspace of X is also separable.

PROOF: Any subspace of a second countable space is second countable. 👍

Metric Versions of Local Compactness

There are (at least) two relevant versions of local compactness for metric spaces. The next definition is nonstandard:

XIII.15.9.12. DEFINITION. Let (X, ρ) be a metric space.

(i) (X, ρ) is a *Heine-Borel space* if every closed bounded subset of X is compact.

(ii) (X, ρ) is *compact im kleinen* if there is an $\eta > 0$ such that every closed subset of diameter $\leq \eta$ is compact.

XIII.15.9.13. Note that these are *metric* notions, not topological ones. A Heine-Borel space is compact im kleinen; a metric space which is compact im kleinen is σ-complete, and locally compact as a topological space; a Heine-Borel space is locally compact and σ-compact. The Heine-Borel Theorem of course says that $\mathbb{R}^n$ with its usual metric is a Heine-Borel space.

A space can be Heine-Borel or compact im kleinen under one metric but not under an equivalent metric (XIII.15.9.14.(i)). A space which is compact im kleinen under one metric is compact im kleinen under any uniformly equivalent metric; thus compactness im kleinen is a uniform property.

XIII.15.9.14. Examples. (i) $(0,1)$ with its usual metric is not compact im kleinen (and *a fortiori* not Heine-Borel). But it is homeomorphic to $\mathbb{R}$, which is Heine-Borel.

(ii) $\mathbb{R}$ with the metric $\tilde{\rho}$ defined by $\tilde{\rho}(x,y) = \min(|x-y|,1)$ is compact im kleinen but not Heine-Borel. The metric $\tilde{\rho}$ is uniformly equivalent to the usual metric. Thus the Heine-Borel property is not a uniform property.

(iii) An infinite set with the discrete metric is compact im kleinen but not Heine-Borel. An uncountable set with the discrete metric is compact im kleinen but not σ-compact.

XIII.15.9.15. It can be shown using the Tietze Extension Theorem (XIII.15.12.10.) that every locally compact, σ-compact metrizable space has a metric making it a Heine-Borel space. But it is unclear whether this can be shown without some Choice. It is possible that every locally compact metrizable space has a metric making it compact im kleinen.

XIII.15.10. Compact and Precompact Uniform Spaces

Compact Hausdorff Spaces

XIII.15.10.1. Theorem. Let $(X,\mathcal{T})$ be a compact Hausdorff space. Let $\mathfrak{U}$ be the set of all neighborhoods of $X \times X$ in the product topology. Then $\mathfrak{U}$ is an admissible uniformity for $\mathcal{T}$ on X, and is the unique admissible uniformity on X. In particular, every compact Hausdorff space is uniformizable.

Proof: It follows immediately from XIII.11.4.5. that $\mathfrak{U}$ is an admissible uniformity for $\mathcal{T}$ on X. To show uniqueness, let $\mathfrak{V}$ be an admissible uniformity for $\mathcal{T}$ on X. 👍

XIII.15.10.2. If some Choice (e.g. DC) is assumed, compact Hausdorff spaces are completely regular, so the last statement follows from (); but the above proof does not use any Choice.. If no Choice is assumed, a compact Hausdorff space X can be constructed which is not completely regular (XIII.11.14.20.). Then $\mathfrak{U}$ is a uniformity on X which does not come from a family of pseudometrics (i.e. has no gauge).

XIII.15.10.3. Corollary. Let $(X,\mathcal{T})$ be a (uniformizable) Hausdorff topological space

Totally Bounded Uniformities

XIII.15.10.4. DEFINITION. Let $(X, \mathfrak{U})$ be a uniform space, and $\mathscr{U}$ the corresponding covering uniformity. Then $(X, \mathscr{U})$ is *totally bounded* if every $\mathcal{U} \in \mathscr{U}$ has a finite subcover. $(X, \mathfrak{U})$ is *compact* if $(X, \mathcal{T}_{\mathfrak{U}})$ is compact.

XIII.15.10.5. Being compact is a topological concept, but being totally bounded is a uniform concept: if $\mathfrak{U}$ and $\mathfrak{V}$ are uniformities admissible for a (noncompact) topology $\mathcal{T}$ on X, it can happen that $(X, \mathfrak{U})$ is totally bounded but $(X, \mathfrak{V})$ is not (). A compact uniform space is totally bounded. A totally bounded uniform space is not necessarily compact, since it is far from true that every open cover is a uniform cover (for example, every subspace of a totally bounded uniform space is totally bounded in the subspace uniformity). In the case of a metric space, the space is totally bounded with respect to the metric uniformity if and only if it is totally bounded in the sense of ().

XIII.15.10.6. PROPOSITION. Let $(X, \mathfrak{U})$ be a uniform space, and $\mathscr{U}$ the corresponding covering uniformity. Then $(X, \mathfrak{U})$ is totally bounded if and only if $\mathscr{U}$ has a subbase consisting of finite covers of X.

XIII.15.10.7. PROPOSITION. Let $\mathfrak{U}$ and $\mathfrak{V}$ be uniformities on a set X. If $\mathfrak{U} \subseteq \mathfrak{V}$ and $(X, \mathfrak{V})$ is totally bounded (resp. compact), then $(X, \mathfrak{U})$ is totally bounded (resp. compact).

XIII.15.10.8. PROPOSITION. Let $\{\mathfrak{U}_i : i \in I\}$ be uniformities on a set X, and $\mathfrak{U} = \bigvee_i \mathfrak{U}_i$. Then $(X, \mathfrak{U})$ is totally bounded if and only if each $(X, \mathfrak{U}_i)$ is totally bounded.

XIII.15.10.9. PROPOSITION. Let $\{(X_i, \mathfrak{U}_i) : i \in I\}$ be a set of uniform spaces. Assume $X = \prod_i X_i \neq \emptyset$, and set $(X, \mathfrak{U}) = \prod_i (X_i, \mathfrak{U}_i)$. Then $(X, \mathfrak{U})$ is totally bounded if and only if each $(X, \mathfrak{U}_i)$ is totally bounded.

XIII.15.10.10. PROPOSITION. Let $(X, \mathfrak{U})$ be a separated uniform space. If $(X, \mathfrak{U})$ is totally bounded, so is $(\bar{X}, \bar{\mathfrak{U}})$.

The following theorem requires at least the BPI if not the full AC:

XIII.15.10.11. THEOREM. Let $(X, \mathfrak{U})$ be a uniform space. Then $(X, \mathfrak{U})$ is compact if and only if it is complete and totally bounded.

XIII.15.10.12. COROLLARY. A (separated) uniform space $(X, \mathfrak{U})$ is totally bounded if and only if $(\bar{X}, \bar{\mathfrak{U}})$ is compact.

XIII.15.10.13. As a result, a totally bounded uniform space is often called *precompact*.

XIII.15.10.14. Theorem XIII.15.10.11. cannot be proved without at least the BPI: the statement implies the BPI (Tikhonov's Theorem for Hausdorff Spaces, cf. XIII.15.12.9.). An arbitrary product of separated complete uniform spaces is complete in the product uniformity (XIII.15.7.8.), and an arbitrary product of totally bounded uniform spaces is totally bounded in the product uniformity by XIII.15.10.9.. Thus an arbitrary product of compact Hausdorff spaces is both complete and totally bounded in the product uniformity.

Completions and Compactifications

To make things simpler, in this subsection we will assume the AC (or at least DC+BPI), and only work with separated uniformities. Under these assumptions, all uniform spaces have gauges and all uniformizable spaces are completely regular, a uniform space is compact if and only if it is complete and totally bounded, an arbitrary product of compact uniform spaces is compact, and the completion of a totally bounded uniform space is compact.

XIII.15.10.15. Let $(X,\mathcal{T})$ be a completely regular space. Let $\mathfrak{U}$ and $\mathfrak{V}$ be (necessarily separated) admissible uniformities for $\mathcal{T}$, and $(Y,\bar{\mathfrak{U}})$ and $(Z,\bar{\mathfrak{V}})$ the completions of $(X,\mathfrak{U})$ and $(X,\mathfrak{V})$ respectively. Then $(Y,\bar{\mathfrak{U}})$ and $(Z,\bar{\mathfrak{V}})$ are not necessarily compact (i.e. $(X,\mathfrak{U})$ and $(X,\mathfrak{V})$ are not necessarily totally bounded). If they are not compact, we can have Y and Z homeomorphic over X even if $\mathfrak{U}\neq\mathfrak{V}$ (they could even both be just X, cf. XIII.15.6.6.). However, if one (hence both) are totally bounded, this cannot happen:

XIII.15.10.16. Proposition. In the above situation, if $(X,\mathfrak{U})$ is totally bounded and Y and Z are homeomorphic over X (hence compact, so $(X,\mathfrak{V})$ is also necessarily totally bounded), then $\mathfrak{U}=\mathfrak{V}$.

Proof: We can identify Y and Z and call the space $\bar{X}$. We have $\mathfrak{U}=(\bar{\mathfrak{U}})_X$ and $\mathfrak{V}=(\bar{\mathfrak{V}})_X$. But $\bar{X}$ is compact, so it has a unique admissible uniformity (XIII.15.10.1.), and thus $\bar{\mathfrak{U}}=\bar{\mathfrak{V}}$. 👍

XIII.15.10.17. Corollary. Let $(X,\mathcal{T})$ be a completely regular space. Then there is a one-one correspondence between compactifications of $(X,\mathcal{T})$ and admissible totally bounded uniformities for $\mathcal{T}$. The partially ordered set of admissible totally bounded uniformities for $\mathcal{T}$, ordered by inclusion, is order-isomorphic to the set of equivalence classes of compactifications of $(X,\mathcal{T})$ with the domination partial order ().

XIII.15.10.18. If $(X,\mathcal{T})$ is completely regular, there is a largest compactification of $(X,\mathcal{T})$, βX. Thus there is a largest admissible totally bounded uniformity for $\mathcal{T}$, namely $\mathfrak{C}^*$ (XIII.15.1.28.). There is a smallest totally bounded admissible uniformity for $\mathcal{T}$ if and only if $(X,\mathcal{T})$ is locally compact (XIII.11.14.21.), namely the uniformity from the one-point compactification $X^\dagger$. This is extendible to arbitrary admissible uniformities:

XIII.15.10.19. PROPOSITION. Let $(X, \mathcal{T})$ be a completely regular space, and $\mathfrak{U}$ an admissible uniformity for $\mathcal{T}$. Then the uniformity $\mathfrak{U} \wedge \mathfrak{C}^*$ (XIII.15.5.6.) is admissible for $\mathcal{T}$, and is the largest totally bounded uniformity contained in $\mathfrak{U}$. In particular, every admissible uniformity on X contains an admissible totally bounded uniformity.

XIII.15.10.20. Thus, if $(X, \mathcal{T})$ is a completely regular space, there is always a largest admissible uniformity, the fine uniformity; but there is a minimal admissible uniformity if and only if $(X, \mathcal{T})$ is locally compact, in which case the uniformity from $X^\dagger$ is the smallest admissible uniformity.

XIII.15.11. Topologically Complete Spaces

One of the deepest theorems in the subject of uniform spaces is SHIROTA'S near-characterization of topologically complete spaces as realcompact (XIII.11.13.4.) spaces [**?**]:

XIII.15.11.1. THEOREM. Let $(X, \mathcal{T})$ be a completely regular topological space. Then

(i) If $(X, \mathcal{T})$ is realcompact, then it is topologically complete.

(ii) If $(X, \mathcal{T})$ is topologically complete, and every closed discrete subspace of X has nonmeasurable cardinality (II.10.4.2.), then $(X, \mathcal{T})$ is realcompact.

XIII.15.11.2. Thus if $(X, \mathcal{T})$ is completely regular, and is not of absurdly large size, then it is topologically complete if and only if it is realcompact. The cardinality restriction is necessary: every discrete space is topologically complete (the discrete metric is complete), but a discrete space of measurable cardinality is not realcompact.

XIII.15.11.3. Every completely metrizable space (XIII.10.3.1.) is obviously topologically complete. But a topologically complete metrizable space need not be completely metrizable. In fact:



XIII.15.11.4. THEOREM. Every metrizable topological space is topologically complete.

See [GJ76] for proofs of XIII.15.11.1. and XIII.15.11.4.. The proofs use the AC.

XIII.15.12. Exercises

XIII.15.12.1. Prove all the propositions in this section given without proof.

XIII.15.12.2. Let $(X, \mathfrak{U})$ be a uniform space. Set $R = \cap_{U \in \mathfrak{U}} U$.

(i) Show that R is an equivalence relation on X, and is the equality relation if and only if $\mathfrak{U}$ is separated.

(ii) Let $\tilde{X} = X/R$ be the set of equivalence classes. Show that $\mathfrak{U}$ induces a unique uniformity $\tilde{\mathfrak{U}}$ on $\tilde{X}$.

(c) Show that the topology $\mathcal{T}_{\tilde{\mathfrak{U}}}$ on $\tilde{X}$ is the quotient topology from $\mathcal{T}_{\mathfrak{U}}$, that the quotient map π from X to $\tilde{X}$ is $(\mathfrak{U}, \tilde{\mathfrak{U}})$-uniformly continuous, and that $(\tilde{X}, \mathcal{T}_{\tilde{\mathfrak{U}}})$ is the maximal T_0 quotient () and the maximal Hausdorff quotient () of $(X, \mathcal{T}_{\mathfrak{U}})$.

(d) If $(Y, \mathfrak{V})$ is any separated uniform space, any $f : X \to Y$ which is $(\mathfrak{U}, \mathfrak{V})$-uniformly continuous factors uniquely through $(\tilde{X}, \tilde{\mathfrak{U}})$ as $\tilde{f} \circ \pi$, and $\tilde{f} : \tilde{X} \to Y$ is $(\tilde{\mathfrak{U}}, \mathfrak{V})$-uniformly continuous.

Thus the theory of general uniform spaces can be essentially completely reduced to separated uniform spaces. Some authors even require uniform spaces to be separated, calling a nonseparated uniformity a "preuniformity."

XIII.15.12.3. (a) If X is a set, there are at most $2^{2^{|X|}}$ uniformities on X. [A covering uniformity on X is an element of $\mathcal{P}(\mathcal{P}(X))$.] So there are at most $2^{2^{\aleph_0}}$ uniformities on $\mathbb{N}$.

(b) $\mathbb{N}$ has $2^{2^{\aleph_0}}$ distinct compactifications (XIII.11.14.9.). So there are at least (hence exactly) $2^{2^{\aleph_0}}$ uniformities on $\mathbb{N}$ admissible for the discrete topology (XIII.15.10.17.).

XIII.15.12.4. Uniformities on $\mathbb{Q}$. Give $\mathbb{Q}$ its topology $\mathcal{T}$ from the usual metric (i.e. the relative topology from $\mathbb{R}$). In this problem, say a uniform structure on $\mathbb{Q}$ is *admissible* if it is admissible for $\mathcal{T}$. Since $\mathcal{T}$ is metrizable, there is an admissible uniform structure, in fact many, including the metric uniformity and the uniformities $\mathfrak{C}$ and $\mathfrak{C}^*$ from $\mathcal{C}$ and $\mathcal{C}^*$ (XIII.15.1.28.). In this problem we describe a large family of other admissible uniformities on $\mathbb{Q}$.

(a) Let A be a set of irrational numbers which is dense in $\mathbb{R}$. Let $\mathcal{I}_A$ be the set of intervals in $\mathbb{Q}$ of the form $[a, b] \cap \mathbb{Q} = (a, b) \cap \mathbb{Q}$ for $a, b \in A$, and let

$$F_A = \{\chi_I : I \in \mathcal{I}_A\}$$

be the set of indicator functions of intervals in $\mathcal{I}_A$. Each $I \in \mathcal{I}_A$ is clopen in $\mathbb{Q}$, so the functions in F_A are continuous. Let $\mathcal{P}_A$ be the set of pseudometrics corresponding to the functions in F_A as in XIII.15.1.24.. Then $\mathcal{P}_A \subseteq \mathcal{C}^*$. Let $\mathfrak{U}_A$ be the uniformity on $\mathbb{Q}$ defined by $\mathcal{P}_A$. Show that $\mathfrak{U}_A$ is an admissible uniformity on $\mathbb{Q}$. [$\mathcal{I}_A$ is a base for the topology of $\mathbb{Q}$.] $\mathfrak{U}_A$ is metrizable if and only if A is countable.

(b) Let b, c be irrational numbers with $b < c$, and $I = [b, c] \cap \mathbb{Q}$. Then I is clopen in $\mathbb{Q}$, so $\chi_I \in BC(\mathbb{Q})$. Show that χ_I is $\mathfrak{U}_A$-uniformly continuous if and only if $b, c \in A$, and that the pseudometric ρ_{χ_I} is uniformly dominated by $\mathcal{P}_A$ if and only if $b, c \in A$. Thus if A and B are distinct dense sets of irrationals, then $\mathfrak{U}_A$ and $\mathfrak{U}_B$ are distinct admissible uniformities. Since there are $2^{2^{\aleph_0}}$ dense sets of irrationals in $\mathbb{R}$, there are $2^{2^{\aleph_0}}$ distinct admissible uniformities (by XIII.15.12.3. there are no more than $2^{2^{\aleph_0}}$ since $\mathbb{Q}$ is countable).

(c) Let A and B be sets of irrational numbers dense in $\mathbb{R}$. Show that $\mathfrak{U}_A \vee \mathfrak{U}_B = \mathfrak{U}_{A \cup B}$.

(d) Let A and B be sets of irrational numbers dense in $\mathbb{R}$. Show that there is an admissible uniformity $\mathfrak{V}$ contained in both $\mathfrak{U}_A$ and $\mathfrak{U}_B$ if and only if $A \cap B$ is dense in $\mathbb{R}$, in which case $\mathfrak{U}_{A \cap B} = \mathfrak{U}_A \wedge \mathfrak{U}_B$.

(e) If A and B are disjoint dense sets of irrationals (e.g. $A = \sqrt{2} + \mathbb{Q}$, $B = \pi + \mathbb{Q}$), then $\mathfrak{U}_A \wedge \mathfrak{U}_B$ is the indiscrete uniformity.

There are *many* other admissible uniformities on $\mathbb{Q}$ (but exactly $2^{2^{\aleph_0}}$ in all): for example, any compactification of $\mathbb{R}$ gives one. In fact, by XIII.13.9.16., every completely regular separable first countable space without isolated points gives at least one admissible uniformity (actually many in general; cf. XIII.11.14.24.).

XIII.15.12.5. (a) Let $(X, \mathfrak{U})$ be a uniform space. Prove without any form of Choice (e.g. without using ()) that $(X, \mathcal{T}_{\mathfrak{U}})$ is T_3 (regular if it is T_0, i.e. if $(X, \mathfrak{U})$ is separated).

***(b) In the absence of any form of Choice, give a topological characterization of what (T_0) topological spaces are uniformizable. [Any subspace of a product of compact Hausdorff spaces is uniformizable by **??**,

XIII.15.5.9., and XIII.15.7.2.. Are these the only ones? This class includes all completely regular spaces (XIII.15.1.28.), and all compactly embeddable spaces (homeomorphic to a subspace of a compact Hausdorff space). Note that one needs Tikhonov's theorem for Hausdorff spaces, i.e. the BPI (II.6.3.11.), to obtain that every completely regular space is compactly embeddable.]

XIII.15.12.6. Let $(X, \mathcal{T})$ be a T_0 topological space. Give $X \times X$ the product topology. If the set of all neighborhoods of Δ is a uniformity on X admissible for $\mathcal{T}$, show that $(X, \mathcal{T})$ is collectionwise normal, and in particular normal. [Let $\{E_i\}$ be a collection of closed subsets of X such that each E_i is disjoint from the closed set $F_i = \cup_{j \neq i} E_j$. Let U be the union of the sets $(X \setminus F_i) \times (X \setminus F_i)$. Then U is an open neighborhood of Δ. If there is an open neighborhood V of Δ with $V \circ V \subseteq U$, consider the sets $V(E_i)$.]

Thus, if $(X, \mathcal{T})$ is completely regular but not collectionwise normal (e.g. not normal), then the fine uniformity for $\mathcal{T}$ (XIII.15.5.4.) does not include all neighborhoods of the diagonal.

XIII.15.12.7. Completions of Nonseparated Uniform Spaces. In this problem we outline the modifications needed to construct a completion (the "minimal" one) of a uniform space which is not separated.

(a) Let $(X, \mathfrak{U})$ be a uniform space which is not necessarily separated, and let $\mathscr{C}$ and $\sim$ be as in XIII.15.8.2.. Set

$$\bar{X} = \{(x, [\mathcal{F}]) : \mathcal{F} \in \mathscr{C} \text{ converges to } x\} \cup \{[\mathcal{F}] : \mathcal{F} \in \mathscr{C} \text{ does not converge}\}$$

and identify $x \in X$ with $(x, [\mathcal{F}])$ for a filter $\mathcal{F}$ converging to x, e.g. the neighborhood filter of x. Modify the definition of the uniform structure $\bar{\mathfrak{U}}$ on $\bar{X}$ to this setting, and prove analogs of XIII.15.8.6. and XIII.15.8.7..

(b) Show that if $(Y, \mathfrak{V})$ is a complete uniform space containing $(X, \mathfrak{U})$, then there is a unique $f : Y \to \bar{X}$ which is the identity on X and with $\mathfrak{V} = f^{-1}(\bar{\mathfrak{U}})$. Thus $(\bar{X}, \bar{\mathfrak{U}})$ is the "minimal" completion.

(c) Let $\approx$ be the T_0-equivalence relation, i.e. $x \approx y$ if $\{x\}$ and $\{y\}$ have the same closure (). Show that $(\bar{X}, \bar{\mathfrak{U}})$ is the unique completion with the property that no $y \in \bar{X} \setminus X$ is T_0-equivalent to any other point of $\bar{X}$.

(d) A uniform space can have many "completions." Let $X = (\mathbb{R} \setminus \{0\}) \times \{0, 1\}$, where $\{0, 1\}$ has the indiscrete topology. Then X is uniformizable under the uniformity $\mathfrak{V} = \pi^{-1}(\mathfrak{U})$, where $\mathfrak{U}$ is the usual uniformity on $\mathbb{R} \setminus \{0\}$ and π is the projection of X onto the first coordinate. Then $(Y, \mathfrak{W})$ is a "completion" of $(X, \mathfrak{V})$, where $Y = \mathbb{R} \times \{0, 1\}$ and $\mathfrak{W}$ the preimage under π of the usual metric uniformity on $\mathbb{R}$. In this case, $\bar{X}$ can be identified with $X \cup \{(0, 0)\}$ with $\bar{\mathfrak{U}} = \mathfrak{W}_{\bar{X}}$.

(e) Even a (separated) complete metrizable uniform space can have a larger "completion" which is not separated. Let $(\mathbb{R}, \mathfrak{U})$ be the metric uniform space from the standard metric on $\mathbb{R}$, $Y = \mathbb{R} \times \{0, 1\}$ (where $\{0, 1\}$ has the indiscrete topology), and $\mathfrak{V} = \pi^1(\mathfrak{U})$. Then $(Y, \mathfrak{V})$ is a "completion" of $(\mathbb{R}, \mathfrak{U})$. For a perhaps more interesting example, let $\mathfrak{U}$ be the usual metric uniformity on $\mathbb{R} \setminus \{0\}$. The space Y of XIII.20.1.3.(v) with doubled origin has a compatible uniform structure $\mathfrak{V}$ which makes $(Y, \mathfrak{V})$ a "completion" of $(X, \mathfrak{U})$.

XIII.15.12.8. A Bizarre Metric Space Without Countable AC. [] Let (X_n) be a sequence of pairwise disjoint nonempty sets such that the product of any infinite subcollection is empty (such a sequence can be constructed in a model of ZF (), but is of course incompatible with the Countable AC). Set $X = \cup_n X_n$. Put a metric ρ on X by setting $\rho(x, y) = \frac{1}{n}$ if $x, y \in X_n$, $x \neq y$, and $\rho(x, y) = \frac{1}{n} + \frac{1}{m}$ if $x \in X_n$, $y \in X_m$, $n \neq m$.

(a) Show that ρ is a metric on X, and the metric topology is the discrete topology.

(b) Show that every ρ-Cauchy sequence is eventually constant. [A ρ-Cauchy sequence not eventually constant would have terms in infinitely many X_n.] Thus (X, ρ) is Cauchy complete.

(c) For each n, let $F_n = \cup_{k \geq n} X_k$. Then (F_n) is a decreasing sequence of closed sets in X. Show that $diam(F_n) = \frac{1}{n} + \frac{1}{n+1}$, so $diam(F_n) \to 0$. But $\cap_n F_n = \emptyset$. So (X, ρ) is not σ-complete, hence not uniformly complete.

(d) Suppose each X_k is finite. Let $X^\dagger = X \cup \{\infty\}$ be the one-point compactification of X. Extend ρ to $\bar{\rho}$ on $X^\dagger$ by setting $\bar{\rho}(x, \infty) = \bar{\rho}(\infty, x) = \frac{1}{n}$ if $x \in X_n$. Show that $\bar{\rho}$ is a metric on $X^\dagger$ defining its topology, so $(X^\dagger, \bar{\rho})$ is a compact metric space. X is dense in $X^\dagger$, i.e. ∞ is in the closure of X in $X^\dagger$, but no sequence in X converges to ∞. $(X^\dagger, \bar{\rho})$ is not separable or second countable.

XIII.15.12.9. Another Bizarre Metric Space Without Countable AC. Let (I_n) be a set of Russell socks, a sequence of pairwise disjoint two-element sets such that the product of any infinite subcollection is empty (), and let $I = \cup_n I_n$. Let $X = \{0,1\}^I$, the set of all functions from I to $\{0,1\}$. Put a metric on X as follows. If $f, g \in X$, $f \neq g$, let $n(f,g)$ be the smallest n such that f and g disagree on I_n, and set $\rho(f,g) = 2^{-n(f,g)}$.

(a) Show that ρ (extended by $\rho(f,f) = 0$ for all f) is a metric on X, and the metric topology is the product topology (taking the discrete topology on $\{0,1\}$).

(b) Show that (X, ρ) is uniformly complete and totally bounded. [For completeness, apply XIII.15.7.8. or give a direct argument.]

(c) Show that every $f \in X$ is constant on I_n for all sufficiently large n.

(d) For each n let $J_n = \cup_{k \leq n} I_k$. Let $S_n \subseteq \{0,1\}^{J_n}$ be the set of all functions s from J_n to $\{0,1\}$ for which s is constant on I_n. For $s \in S_n$, set $\ell(s) = n$ and U_s the set of $f \in X$ which extend s. Show that U_s is a clopen set in X.

(e) Let $\mathcal{U}$ be the collection of all U_s for all $s \in \cup_n S_n$. Show that $\mathcal{U}$ is an open cover of X.

(f) Show that $\mathcal{U}$ does not have a finite subcover. [If $\{U_{s_1}, \ldots, U_{s_m}\}$ is a finite subset of $\mathcal{U}$, choose $n > \ell(s_k)$ for all k and let $s \in S_n$ which is not constant on I_k for $k < n$, and extend s to $f \in X$. Then $f \notin U_{s_k}$ for any k.] Thus X is not compact.

This is perhaps the simplest explicit example of a metric space which is uniformly complete and totally bounded, but not compact, and also of a product of compact Hausdorff spaces which is not compact (a metrizable counterexample to Tikhonov's Theorem for Hausdorff spaces). It of course requires a drastic set-theoretic assumption incompatible with the Countable AC or the BPI.

This example appears (cryptically) in [**?**]; the above argument is due to I. Farah.

XIII.15.12.10. Let (X, ρ) be a locally compact, σ-compact metric space. Let (U_n) be an increasing sequence of open sets in X such that $\bar{U}_n$ is compact and contained in U_{n+1}, with $X = \cup_n U_n$.

(a) For each n let $F_n = \bar{U}_n \setminus U_n$. In $X^\dagger$, define a continuous function f from $F = \cup_n F_n \cup \{\infty\}$ by setting $f = \frac{1}{n}$ on F_n and $f(\infty) = 0$. Use the Tietze Extension Theorem to extend f to a continuous function $g : X^\dagger \to [0,1]$ with $g(x) > 0$ for all $x \in X$.

(b) Show that X is homeomorphic to the set

$$\left(x, \frac{1}{g(x)}\right) \subseteq X \times \mathbb{R}$$

and that the product metric is an equivalent metric to ρ making X into a Heine-Borel space.

XIII.16. Order Topology

XIII.16.1. Definition and Properties

XIII.16.1.1. THEOREM. Any subspace of a space with the order topology is completely normal. In particular, any space with the order topology is normal.

PROOF: [Cat06] Let X be an ordered space with the order topology, and let Y be a subspace. We will show that Y has the following stronger property: if $\{A_i : i \in I\}$ is a collection of subsets of Y which are *separated*, i.e. each A_i is disjoint from the closure (in Y) of $\cup_{j \neq i} A_j$, then the A_i have disjoint neighborhoods V_i in Y (Y is said to be *completely collectionwise normal*).

Note that by (), if the A_i are separated in Y, then the A_i are separated in X, and if $\{U_i\}$ are disjoint neighborhoods in X, and $V_i = U_i \cap Y$, then $\{V_i\}$ are disjoint neighborhoods in Y. Thus for the proof we may assume $Y = X$.

Suppose $\{A_i\}$ are separated in X (and in particular disjoint), and set $A = \cup_i A_i$. For each convex component C of $X \setminus A$, choose a point $f(C) \in C$. For $p \in A_i$, p not the largest point of X (if there is one), choose $t_p \in X$ as follows. There is a $u_p > p$ such that $[p, u_p)$ contains no point of $A \setminus A_i$.

> Choose t_p to be a point of $A_i \cap (p, u_p)$ if $A_i \cap (p, u_p)$ is nonempty.
>
> Choose $t_p = u_p$ if $(p, u_p) = \emptyset$.
>
> If $(p, u_p) \neq \emptyset$ but $A_i \cap (p, u_p) = \emptyset$, then (p, u_p) is contained in a unique convex component C of $X \setminus A$; in this case set $t_p = f(C)$.

Similarly, if p is not the smallest point of X, choose $s_p < p$ in the same manner with inequalities reversed. Set $I_p = (s_p, t_p)$. (If p is the largest point of X, set $I_p = (s_p, p]$, and if p is the smallest point of X, set $I_p = [p, t_p)$.) Then I_p is an open interval in X containing p. This defines I_p for each $p \in A$.

For each i, set $U_i = \cup_{p \in A_i} I_p$. Then U_i is an open neighborhood of A_i. We claim the U_i are disjoint. If $x \in U_i \cap U_j$ for $i \neq j$, there are $p \in A_i$ and $q \in A_j$ such that $x \in I_p \cap I_q$. We may suppose $p < q$. Then $s_q < x < t_p$. Since $q \notin I_p$ and $p \notin I_q$, $p < s_q < x < t_p < q$. Thus $t_p \in I_q$, and hence $t_p \notin A_i$, and $(p, t_p) \neq \emptyset$. Similarly, $s_q \notin A_j$ and $(s_q, q) \neq \emptyset$. By the way t_p and s_q were selected, we have $t_p = s_q = f(C)$ for the same convex component C of $X \setminus A$, contradicting that $s_q < t_p$. Thus the U_i are disjoint. 👍

Note that the AC is needed in several places in this proof. In fact, the AC (or something only slightly weaker) is necessary for this result [vD85].

XIII.16.2. Completion of an Order Topology

XIII.17. Function Spaces

XIII.17.1. Function Space Notation

There are several notations for function spaces used in the literature, and the usage is not always consistent. There is one predominant notation:

XIII.17.1.1. Definition. Let X and Y be topological spaces. Denote by Y^X the set of all maps (continuous functions) from X to Y.

XIII.17.1.2. The notation $C(X,Y)$ is also frequently used for Y^X. Some authors use Y^X to denote the set of all functions from X to Y, which we will write $F(X,Y)$. Although it is occasionally useful to consider this set, the set $C(X,Y)$ arises far more often. Our use of Y^X to denote $C(X,Y)$, if done with care, includes all standard examples of this exponential notation:

XIII.17.1.3. Examples. (i) If X has the discrete topology, then for any Y we have $Y^X = C(X,Y) = F(X,Y)$. If X and Y are just sets, with no topology, we can consider them to be topological spaces with the discrete topology so the notation $Y^X = F(X,Y)$ used in set theory is consistent with the topological notation. Note that the topology on Y has no effect on the definition of Y^X as a set in this case.

(ii) If Y is a topological space (or just a set, regarded as a topological space with the discrete topology), and $n \in \mathbb{N}$, we can regard n as the subset $\{1,\dots,n\}$ of $\mathbb{N}$ (or, more commonly and properly in set theory, $\{0,1,\dots,n-1\} \subseteq \mathbb{N}_0$), with the discrete topology. Then Y^n (in the sense of XIII.17.1.1.) is naturally the set of n-tuples of elements of Y, the usual meaning of Y^n. Similarly, $Y^{\mathbb{N}}$ denotes the set of sequences in Y (where $\mathbb{N}$ has its usual discrete topology). This latter set is often, but somewhat less precisely, denoted Y^∞.

(iii) More generally, if κ is a cardinal, we denote by Y^κ the set of all functions from a set of cardinality κ to Y. To be precise, we regard κ as a set of ordinals as in (), and give it the discrete topology. This set Y^κ can be naturally identified with the Cartesian product of κ copies of Y.

(iv) If X is a set, the power set $\mathcal{P}(X)$ is often denoted 2^X. If X is a topological space, 2^X is often used to denote the set of closed subsets of X. Both notations are consistent with XIII.17.1.1. if we give $2 = \{0,1\}$ the Sierpiński space topology with $\{0\}$ open and $\{1\}$ closed. We then identify a continuous function $f : X \to 2$ with the subset $f^{-1}(\{1\})$ of X.

XIII.17.1.4. Both the notations Y^X and $C(X,Y)$ are slight abuses, since the notation does not reflect the topologies of X and Y which are needed to determine Y^X as a set. (The notation $F(X,Y)$ is not ambiguous.) The abuse is the same as in the common statement "Let X be a topological space."

XIII.17.1.5. If X is a topological space, the set $\mathbb{R}^X = C(X,\mathbb{R})$ is usually denoted $C_{\mathbb{R}}(X)$. Similarly, $\mathbb{C}^X = C(X,\mathbb{C})$ is usually denoted $C_{\mathbb{C}}(X)$ or just $C(X)$. Here $\mathbb{R}$ and $\mathbb{C}$ are always assumed to have their usual topologies.

XIII.17.2. U.C. Convergence and the Compact-Open Topology

There are many important notions of convergence for functions, i.e. many topologies on spaces of functions. The most generally useful topology on Y^X is the "topology of uniform convergence on compact sets" ("u.c. convergence") (IV.3.2.). If Y is metrizable, or more generally if the topology on Y comes from a uniform structure, this topology can be defined in a straightforward way; the idea is that a net (f_i) converges to f if and only if, for every compact subset K of X, for sufficiently large i $f_i(x)$ should be close to $f(x)$ for all $x \in K$. Even for a general topological space Y, there is a clean way to define such a topology on Y^X, called the compact-open topology.

XIII.17.2.1. DEFINITION. Let X and Y be topological spaces. The *compact-open topology* on Y^X is the topology generated by the subbase

$$\{U_{K,O} : K \subseteq X \text{ compact}, O \subseteq Y \text{ open}\}$$
$$U_{K,O} = \{f \in Y^X : f(K) \subseteq O\} .$$

XIII.17.2.2. The set of $U_{K,O}$ does not give a base for a topology in general; one needs to take finite intersections. So a basic open set for the compact-open topology is defined by taking a finite number of compact sets $K_1, \ldots, K_n$ in X, a corresponding finite number of open sets $O_1, \ldots, O_n$ in Y, and taking all maps from X to Y such that $f(K_j) \subseteq O_j$ for $1 \leq j \leq n$.

To obtain a base for the compact-open topology, it suffices to let the O's range over a base for the topology of Y.

The compact-open topology can be described in terms of convergence:

XIII.17.2.3. PROPOSITION. Let X and Y be topological spaces. Then a net (f_i) in Y^X converges to $f \in Y^X$ in the compact-open topology if and only if, for every compact subset K of X, and every open set O in Y such that $f(K) \subseteq O$, there is an i_0 such that $f_i(K) \subseteq O$ for all $i \geq i_0$.

The proof is immediate from the definition, and is left to the reader.

Since singleton sets are compact, we obtain:

XIII.17.2.4. PROPOSITION. Let X and Y be topological spaces. If a net (f_i) in Y^X converges to $f \in Y^X$ in the compact-open topology, then $f_i \to f$ pointwise on X, i.e. $f_i(x) \to f(x)$ for all $x \in X$.

PROOF: Let $x \in X$. If O is any neighborhood of $f(x)$ in Y, then $f(\{x\}) \subseteq O$, so there is an i_0 such that $f_i(\{x\}) \subseteq O$ for all $i \geq i_0$, i.e. $f_i(x) \in O$ for $i \geq i_0$. 👍

XIII.17.2.5. The compact-open topology is thus stronger than the topology of pointwise convergence; it is usually strictly stronger. But if every compact subset of X is finite, e.g. if X is discrete, the two topologies coincide. (There are nondiscrete spaces where every compact subset is finite, e.g. XIII.11.14.10.; but such a space cannot have any nontrivial convergent sequences. See ().)

XIII.17.2.6. Corollary. If Y is a topological space and X a set with the discrete topology, then Y^X can be identified with the Cartesian product of κ copies of Y, where $\kappa = card(X)$, and the compact-open topology is the product topology.

XIII.17.2.7. If X and Y are topological spaces, any constant function from X to Y is continuous. The subset C of Y^X of constant functions can be identified with Y. It is immediate from the definitions that this identification is a homeomorphism if C is given the relative compact-open topology. If Y is Hausdorff, then C is closed in Y^X in the compact-open topology. (The converse does not hold in general, e.g. if X is a one-point space.)

We now examine separation properties.

XIII.17.2.8. Proposition. Let X and Y be topological spaces. Give Y^X the compact-open topology.

(i) Y^X is Hausdorff if and only if Y is Hausdorff.

(ii) Y^X is regular if and only if Y is regular.

Proof: Y is homeomorphic to a subset of Y^X (XIII.17.2.7.), so if Y^X is Hausdorff or regular, so is Y. Conversely, suppose Y is Hausdorff, and f and g are distinct elements of Y^X. Then there is an $x \in X$ with $f(x) \neq g(x)$. Let O_1 and O_2 be disjoint open neighborhoods of $f(x)$ and $g(x)$ in Y. Then $U_{\{x\},O_1}$ and $U_{\{x\},O_2}$ are disjoint neighborhoods of f and g in Y^X.

Now suppose Y is regular. Let $f \in Y^X$ and V a compact-open neighborhood of f. Then there are compact sets $K_1, \ldots, K_n$ in X and open sets $O_1, \ldots, O_n \in Y$ with $f(K_j) \subseteq O_j$ for $1 \leq j \leq n$ and

$$U_{K_1,O_1} \cap \cdots \cap U_{K_n,O_n} \subseteq V \ .$$

Fix j. Since $f(K_j)$ is compact, by () there is an open set O'_j in Y with

$$f(K_j) \subseteq O'_j \subseteq \bar{O}'_j \subseteq O_j \ .$$

Then $W = U_{K_1,O'_1} \cap \cdots \cap U_{K_n,O'_n}$ is a neighborhood of f in Y^X. We claim $\bar{W} \subseteq V$. Suppose (g_i) is a net in W and $g_i \to g \in Y^X$. If $x \in K_j$, then $g_i(x) \in O'_j$ for all i, so $g(x) \in \bar{O}'_j \subseteq O_j$. Thus $g \in V$. 👍

XIII.17.2.9. If Y is normal, then Y^X is not necessarily normal: if S is the Sorgenfrey line (XIII.7.10.14.), and $2 = \{0, 1\}$ has the discrete topology, then S is normal, but S^2 is not normal.

XIII.17.2.10. Example. Let X be a topological space. Consider 2^X, where 2 has the Sierpiński topology. Then 2^X can be identified with the set of closed subsets of X as in XIII.17.1.3.(iv). What is the corresponding topology on the collection of closed subspaces? Since 2 is not Hausdorff, limits in 2^X will not be unique. If (E_i) is a net of closed subspaces of X and E a closed subspace of X, then $E_i \to E$ in this topology if and only if, whenever K is a compact subspace of X disjoint from E, then there is an i_0 such

that K is disjoint from E_i for all $i \geq i_0$. In particular, if $E_i \to E$, and $x \notin E$, then there is an i_0 such that $x \notin E_i$ for any $i \geq i_0$. Thus we have

$$\bigcap_i \left[\bigcup_{j \geq i} E_j \right] \subseteq E \ .$$

This condition is in general not sufficient to guarantee that $E_i \to E$, however. It is sufficient that

$$E_0 := \bigcap_i \left[\overline{\bigcup_{j \geq i} E_j} \right] \subseteq E$$

for if E is a closed set containing E_0, and K is a compact set disjoint from E, then K is disjoint from E_0, so if W_i is the complement of $\overline{\cup_{j \geq i} E_j}$ the W_i form an increasing open cover of K, which has a finite subcover, i.e. $K \subseteq W_{i_0}$ for some i_0. Then K is disjoint from E_i for all $i \geq i_0$. Thus $E_i \to E$.

If every point of X has local base of compact neighborhoods, e.g. if X is a locally compact Hausdorff space, $E_i \to E$, and $x \notin E$, then since E is closed x has a compact neighborhood K disjoint from E. Then K is disjoint from E_i for all sufficiently large i, so $x \notin E_0$, i.e. $E_0 \subseteq E$. So in this case $E_i \to E$ if and only if E is a closed set containing E_0.

We now examine the case where Y is metrizable. In this case, the compact-open topology is precisely the topology of u.c. convergence:

XIII.17.2.11. Proposition. Let X be a topological space and (Y, ρ) a metric space. If (f_i) is a net in Y^X, then $f_i \to f \in Y^X$ if and only if, for every compact subset K of X and every $\epsilon > 0$, there is an i_0 such that $\rho(f_i(x), f(x)) < \epsilon$ for all $i \geq i_0$ and all $x \in K$.

Proof: Suppose the ϵ-condition is satisfied. Let K be a compact subset of X and O an open set in Y. Then, since $f(K)$ is compact, there is an $\epsilon > 0$ such that $\rho(f(x), Y \setminus O) \geq \epsilon$ for all $x \in K$. Fix i_0 such that $\rho(f_i(x), f(x)) < \epsilon$ for all $i \geq i_0$ and $x \in K$. Then $f_i(K) \subseteq O$ for all $i \geq i_0$. So $f_i \to f$ in the compact-open topology.

Conversely, if $f_i \to f$ in the compact-open topology, let K be a compact subset of X and $\epsilon > 0$. For each $x \in X$ let V_x be the open ball of radius $\frac{\epsilon}{4}$ centered at $f(x)$. Then finitely many of these balls, say $\{V_{x_1}, \ldots, V_{x_n}\}$, cover $f(K)$ since $f(K)$ is compact. For each j, $1 \leq j \leq n$, let $K_j = K \cap f^{-1}(\bar{V}_{x_j})$ and O_j the open ball of radius $\frac{\epsilon}{3}$ around $f(x_j)$. For each j, K_j is compact and $f(K_j) \subseteq O_j$. There is an i_0 such that $f_i(K_j) \subseteq O_j$ for all j and for all $i \geq i_0$. Thus, if $x \in K$, then $x \in K_j$ for some j, so for any $i \geq i_0$ we have

$$\rho(f_i(x), f(x)) \leq \rho(f_i(x), f_i(x_j)) + \rho(f_i(x_j), f(x_j)) + \rho(f(x_j), f(x)) < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon \ .$$

XIII.17.2.12. More generally, if Y has a (separated) uniform structure $\mathcal{U}$, the compact-open topology is the topology of uniform convergence on compact sets (the proof is essentially identical). The topology of u.c. convergence appears to depend on the uniformity, but the compact-open topology (and thus the u.c. topology) only depends on the topology of the uniformity and not on the specific uniformity. This should not be surprising since a continuous image of a compact set is compact, and a compact Hausdorff space has a unique uniformity compatible with its topology.

XIII.17.2.13. If X and Y are topological spaces, the compact-open topology can be defined in the same way on $F(X,Y)$ as in XIII.17.2.1.. The relative topology on the subset $Y^X = C(X,Y)$ is exactly the compact-open topology defined in XIII.17.2.1.. However, the compact-open topology on $F(X,Y)$ does not reasonably coincide with the topology of uniform convergence on compact sets, and Y^X is not closed in $F(X,Y)$ in general. See Exercise XIII.17.4.1..

Joint Continuity

XIII.17.2.14. If X and Y are topological spaces, there is a natural evaluation map from $Y^X \times X$ to Y sending (f,x) to $f(x)$. If $\mathcal{T}$ is a topology on Y^X, giving a product topology on $Y^X \times X$, we discuss continuity of the evaluation map. We will be particularly concerned with the topology $\mathcal{P}$ of pointwise convergence and the stronger compact-open topology $\mathcal{C}$.

XIII.17.2.15. Definition. Let X and Y be topological spaces, $\mathcal{F}$ a subset of Y^X, and A a subset of X. Let $\mathcal{T}$ be a topology on Y^X. Then $\mathcal{T}$ is *jointly continuous on A for $\mathcal{F}$* if the evaluation map $\mathcal{F} \times A \to Y$ is continuous for the product topology. $\mathcal{T}$ is *jointly continuous for $\mathcal{F}$* if it is jointly continuous on X, and *jointly continuous* if it is jointly continuous for Y^X. (One could more generally take $\mathcal{T}$ to be a topology just on $\mathcal{F}$, but this creates technical complications.)

XIII.17.2.16. Under the hypotheses, $\mathcal{T}$ is jointly continuous for $\mathcal{F}$ on A if and only if, whenever (f_i) is a net in $\mathcal{F}$ $\mathcal{T}$-converging to $f \in \mathcal{F}$, and (x_i) (same index set) is a net in A converging to $x \in A$, we have $f_i(x_i) \to f(x)$. This is called *continuous convergence.*

XIII.17.2.17. Examples such as () show that $\mathcal{P}$ is not jointly continuous for Y^X in general, even if X and Y are $[0,1]$.

XIII.17.2.18. Theorem. Let X and Y be topological spaces and $\mathcal{F} \subseteq Y^X$. Then

(i) The topology $\mathcal{C}$ is jointly continuous for $\mathcal{F}$ on compact subsets of X.

(ii) If $\mathcal{T}$ is a topology on $\mathcal{F}$ which is jointly continuous for $\mathcal{F}$ on compact subsets of X, then $\mathcal{T}$ is stronger than $\mathcal{C}$ on $\mathcal{F}$ on compact subsets of X (i.e. the restriction of $\mathcal{T}$ to $\mathcal{F}|_K$ is stronger than the restriction of $\mathcal{C}$ for every compact subset K of X, where $\mathcal{F}|_K$ denotes the restriction to K of the functions in $\mathcal{F}$).

(iii) If X is locally compact (Hausdorff), then $\mathcal{C}$ is jointly continuous for any $\mathcal{F} \subseteq Y^X$, and every jointly continuous topology for $\mathcal{F}$ is stronger than $\mathcal{C}$ on $\mathcal{F}$. In particular, $\mathcal{C}$ is jointly continuous, and every jointly continuous topology is stronger than $\mathcal{C}$.

XIII.17.2.19. This theorem implies that if $\mathcal{P}$ is jointly continuous for $\mathcal{F} \subseteq Y^X$ on compact subsets of X, then $\mathcal{P}$ and $\mathcal{C}$ agree on $\mathcal{F}$ on compact subsets of X, and if X is locally compact and $\mathcal{P}$ is jointly continuous for $\mathcal{F}$, then $\mathcal{P}$ and $\mathcal{C}$ agree on $\mathcal{F}$.

XIII.17.3. The Arzelà-Ascoli Theorem

The Arzelà-Ascoli Theorem gives a useful condition for a set $\mathcal{F}$ of functions between topological spaces X and Y to be subcompact in the compact-open topology. Various versions of the theorem appear in the literature; we will prove a general version and deduce the common versions as special cases, including the versions stated by Arzelà and Ascoli themselves.

For ease of exposition, we will assume the Axiom of Choice. Dependence of the various forms of the theorem on the AC will be discussed at the end of the section. We will also almost exclusively work with Hausdorff spaces; a few things work also in a non-Hausdorff setting. In fact, most spaces will be completely regular (e.g. locally compact Hausdorff).

Our exposition is adapted from [Kel75].

Subcompact Sets

XIII.17.3.1. Definition. Let X be a topological space. A (nonempty) subset A of X is *subcompact* if A is is a subset of a compact subset of X.

XIII.17.3.2. If X is Hausdorff, a subset A of X is subcompact if and only if $\bar{A}$ is compact (since compact subsets of X are closed). This can fail in non-Hausdorff spaces: a compact subspace need not be closed, and its closure need not be compact ().

XIII.17.3.3. Examples. (i) Any compact subset of a topological space X is subcompact.

(ii) A subset of $\mathbb{R}^n$ is subcompact if and only if it is bounded.

(iii) A subset of a complete metric space (or, more generally, a complete uniform space) is subcompact if and only if it is totally bounded. (A totally bounded subset of a uniform space is sometimes called a *precompact* subset. Thus a subcompact subset is precompact, but not necessarily conversely if the uniform space is not complete. A precompact uniform space is one whose completion is compact.)

XIII.17.3.4. Proposition. Let X be a completely regular space, and A a nonempty subset. Then A is subcompact if and only if every net in A has a subnet which converges in X. If X is first countable, then A is subcompact if and only if every sequence in A has a subsequence which converges in X.

Proof: Suppose $\bar{A}$ is not compact. Then $\bar{A}$ is a dense proper subspace of its Stone-Čech compactification $\beta\bar{A}$. A is dense in $\bar{A}$ and hence in $\beta\bar{A}$. If $p \in \beta\bar{A} \setminus \bar{A}$, then there is a net in A converging to p. This net cannot have a cluster point in $\bar{A}$ since $\beta\bar{A}$ is Hausdorff, hence cannot have a cluster point in X. The converse is obvious. 👍

It is unclear how much the completely regular hypothesis can be relaxed.

Even Continuity and Equicontinuity

XIII.17.3.5. DEFINITION. Let X be a topological space, Y a set, $x \in X$, and ρ a pseudometric on Y. A set $\mathcal{F}$ of functions from X to Y is *ρ-equicontinuous at x* if for every $\epsilon > 0$ there is a neighborhood U of x such that $\rho(f(z), f(x)) < \epsilon$ for every $z \in U$. $\mathcal{F}$ is *ρ-equicontinuous* if it is ρ-equicontinuous at x for every $x \in X$. If $A \subseteq X$, then $\mathcal{F}$ is *ρ-equicontinuous on A* if the set of restrictions of the functions in $\mathcal{F}$ to A is ρ-equicontinuous (as functions from A to Y).

If $(Y, \mathcal{V})$ is a uniform space, and $\mathcal{F} \subseteq Y^X$, then $\mathcal{F}$ is *$\mathcal{V}$-equicontinuous* if it is ρ-equicontinuous for every ρ in the gauge of $\mathcal{V}$. When $\mathcal{V}$ is understood, we just say $\mathcal{F}$ is *equicontinuous.* If $A \subseteq X$, then $\mathcal{F}$ is *$\mathcal{V}$-equicontinuous on A* if the set of restrictions of the functions in $\mathcal{F}$ to A is $\mathcal{V}$-equicontinuous (as functions from A to Y).

XIII.17.3.6. If $(Y, \mathcal{V})$ is a uniform space, and $\mathcal{F}$ an equicontinuous set of functions from X to Y, the functions in $\mathcal{F}$ are automatically continuous with respect to the uniform topology on Y. The equicontinuity condition roughly says that for any $x \in X$, if $z \in X$ is close enough to x, then $f(z)$ is uniformly close to $f(x)$ for all $f \in \mathcal{F}$. It is enough to check ρ-equicontinuity for enough ρ's to determine $\mathcal{V}$; in particular, if $\mathcal{V}$ is the uniformity from a metric ρ, then $\mathcal{F}$ is $\mathcal{V}$-equicontinuous if and only if it is ρ-equicontinuous. This is the most common situation in applications.

The equicontinuity condition can be stated without reference to pseudometrics: $\mathcal{F}$ is $\mathcal{V}$-equicontinuous at x if and only if, for every $V \in \mathcal{V}$, there is a neighborhood U of x such that $f(U) \subseteq V[x]$ for all $f \in \mathcal{F}$ (i.e. $(f(z), f(x)) \in V$).

XIII.17.3.7. **Caution:** If $\mathcal{F}$ is equicontinuous on a set A, it does not mean $\mathcal{F}$ is equicontinuous at every $x \in A$, since it is the *restrictions* of the functions to A which are equicontinuous. See V.2.2.2. for the analogous distinction for continuity.

There is a topological version of equicontinuity. The definition is a bit curious but has alternate characterizations, and is useful to work with.

XIII.17.3.8. DEFINITION. Let X and Y be topological spaces, and $\mathcal{F} \subseteq Y^X$. Then $\mathcal{F}$ is *evenly continuous* if, for every $x \in X$, $y \in Y$, and open neighborhood U of Y, there are open neighborhoods V of x and W of y, $W \subseteq U$, such that for all $f \in \mathcal{F}$ with $f(x) \in W$, $f(V) \subseteq U$. If $A \subseteq X$, then $\mathcal{F}$ is *evenly continuous on A* if the set $\mathcal{F}|_A$ of restrictions of the functions in $\mathcal{F}$ to A is evenly continuous as a set of functions from A to Y.

XIII.17.3.9. PROPOSITION. Let X and Y be topological spaces, and $\mathcal{F} \subseteq Y^X$. Then the following are equivalent:

(i) $\mathcal{F}$ is evenly continuous.

(ii) For every $x \in X$, $y \in Y$, and open neighborhood U of Y, there are open neighborhoods V of x and W of y, $W \subseteq U$, such that the evaluation map sends $\{f \in \mathcal{F} : f(x) \in W\} \times V$ into U.

(iii) Whenever (f_i, x_i) is a net in $\mathcal{F} \times X$ such that $x_i \to x \in X$, $f_i(x) \to y \in Y$, then $f_i(x_i) \to y$.

Version (ii) shows the close connection between even continuity and joint continuity of evaluation.

XIII.17.3.10. PROPOSITION. Let X and Y be topological spaces, with Y regular. Let $\mathcal{F} \subseteq Y^X$, and let $\tilde{\mathcal{F}}$ be the closure of $\mathcal{F}$ in the set $F(X,Y)$ of all functions in the topology $\mathcal{P}$ of pointwise convergence. If $\mathcal{F}$ is evenly continuous, then $\tilde{\mathcal{F}}$ is also evenly continuous, and $\mathcal{P}$ is jointly continuous on $\tilde{\mathcal{F}}$. In particular, every function in $\tilde{\mathcal{F}}$ is continuous.

PROOF: Let $x \in X$, $y \in Y$, and U an open neighborhood of y. There is an open neighborhood U' of y such that $\bar{U}' \subseteq U$ since Y is regular. Since $\mathcal{F}$ is evenly continuous, there are neighborhoods V of x and W of y, $W \subseteq U'$, such that for all $f \in \mathcal{F}$ with $f(x) \in W$, $f(V) \subseteq U'$. Let $g \in \tilde{\mathcal{F}}$ with $g(x) \in W$. Let (f_i) be a net in $\mathcal{F}$ converging pointwise to g; then $f_i(x) \in W$ for sufficiently large i since W is open. Thus if $z \in V$, $f_i(z) \in U'$ for sufficiently large i, and since $f_i(z) \to g(z)$, $g(z) \in \bar{U}' \subseteq U$.
For the last statement, note that $\{f \in \tilde{\mathcal{F}} : f(x) \in W\}$ is $\mathcal{P}$-open in $\tilde{\mathcal{F}}$, so $\{f \in \tilde{\mathcal{F}} : f(x) \in W\} \times V$ is open in $\tilde{\mathcal{F}} \times X$. 👍

If the target space is a uniform space, equicontinuity and even continuity are essentially the same thing:

XIII.17.3.11. THEOREM. Let X be a topological space and $(Y, \mathcal{V})$ a separated uniform space. Give Y the uniform topology. Let $\mathcal{F} \subseteq Y^X$. Then
(i) If $\mathcal{F}$ is equicontinuous, then $\mathcal{F}$ is evenly continuous.
(ii) If $\mathcal{F}[x]$ is subcompact for every $x \in X$, and $\mathcal{F}$ is evenly continuous, then $\mathcal{F}$ is equicontinuous. [$\mathcal{F}[x] = \{f(x) : f \in \mathcal{F}\}$.]

This has an interesting consequence:

XIII.17.3.12. COROLLARY. Let X be a topological space, Y a completely regular space, and $\mathcal{F} \subseteq Y^X$ with the property that $\mathcal{F}[x]$ is subcompact for every $x \in X$. If $\mathcal{V}$ and $\mathcal{W}$ are uniform structures for Y compatible with its topology, then $\mathcal{F}$ is $\mathcal{V}$-equicontinuous if and only if $\mathcal{F}$ is $\mathcal{W}$-equicontinuous.

This is reminiscent of the fact that a compact Hausdorff space has a unique compatible uniform structure.

XIII.17.3.13. Equicontinuity (and even continuity) are *pointwise* conditions. As a consequence, if $\mathcal{F}$ is equicontinuous [tesp. evenly continuous] on a neighborhood of each point, it is equicontinuous [resp. evenly continuous]. In particular, if X is locally compact, then $\mathcal{F}$ is equicontinuous if and only if it is equicontinuous on every compact subset of X.

There is a uniform version of equicontinuity:

XIII.17.3.14. DEFINITION. Let $(X,\mathcal{U})$ and $(Y,\mathcal{V})$ be uniform spaces, and $\mathcal{F} \subseteq Y^X$. $\mathcal{F}$ is *uniformly equicontinuous* [*$(\mathcal{U},\mathcal{V})$-uniformly equicontinuous*] if for every $V \in \mathcal{V}$ there is a $U \in \mathcal{U}$ such that $(f(z), f(x)) \in V$ whenever $(z,x) \in U$ and $f \in \mathcal{F}$.

XIII.17.3.15. If $\mathcal{F}$ is uniformly equicontinuous, then it is equicontinuous, and every $f \in \mathcal{F}$ is uniformly continuous. For example, a set $\mathcal{F}$ consisting of only one continuous function f is always equicontinuous, but is uniformly equicontinuous if and only if f is uniformly continuous. Informally, a set $\mathcal{F}$ is uniformly equicontinuous if the functions in $\mathcal{F}$ are "uniformly uniformly continuous."

If (X, ρ) and (Y, σ) are metric spaces and $\mathcal{F} \subseteq Y^X$, then $\mathcal{F}$ is uniformly equicontinuous (with respect to the metric uniformities) if and only if, for every $\epsilon > 0$, there is a $\delta > 0$ such that $\sigma(f(z), f(x)) < \epsilon$ whenever $\rho(x, z) < \delta$ and $f \in \mathcal{F}$.

The next result generalizes the result that a continuous function from a compact space to a uniform space is uniformly continuous ():

XIII.17.3.16. Proposition. Let $(X, \mathcal{U})$ and $(Y, \mathcal{V})$ be uniform spaces with X compact, and $\mathcal{F} \subseteq Y^X$. Then $\mathcal{F}$ is uniformly equicontinuous if and only if it is equicontinuous.

Statement and Proof

Here is our most general version of the Arzelà-Ascoli Theorem. Recall () that a *k-space* is a compactly generated Hausdorff space; any Hausdorff space which is locally compact or first countable is a k-space.

XIII.17.3.17. Theorem. [Arzelà-Ascoli Theorem, General Version] Let X be a k-space and Y a regular space, and let $\mathcal{F} \subseteq Y^X$. Then $\mathcal{F}$ is subcompact in the compact-open topology if and only if

(i) $\mathcal{F}$ is evenly continuous on compact subsets of X.

(ii) $\mathcal{F}[x]$ is subcompact in Y for all $x \in X$.

If these conditions are satisfied, the compact-open closure $\bar{\mathcal{F}}$ coincides with the closure in the topology of pointwise convergence, and the compact-open topology coincides with the topology of pointwise convergence on $\bar{\mathcal{F}}$.

Note that if X is locally compact, then condition (i) is simply that $\mathcal{F}$ is evenly continuous ().

Proof: Let $\tilde{\mathcal{F}}$ be the $\mathcal{P}$-closure of $\mathcal{F}$ in the set $F(X, Y)$ of all functions from X to Y. By (), if K is a compact subset of X, $\tilde{\mathcal{F}}|_K$ consists of continuous functions on K. Thus $\tilde{\mathcal{F}}$ consists of continuous functions from X to Y since X is a k-space, i.e. $\tilde{\mathcal{F}}$ is the $\mathcal{P}$-closure of $\mathcal{F}$ in Y^X.
The set $\tilde{\mathcal{F}}$ is a closed subset of $\prod_{x \in X} \overline{\mathcal{F}[x]}$, which is $\mathcal{P}$-compact by Tikhonov's Theorem (for Hausdorff spaces). Thus $\tilde{\mathcal{F}}$ is $\mathcal{P}$-compact.
We now show that the topologies $\mathcal{P}$ and $\mathcal{C}$ coincide on $\tilde{\mathcal{F}}$. It suffices to show that

$$U_{K,O} = \{f \in \mathcal{F} : f(K) \subseteq O\}$$

is $\mathcal{P}$-open for $K \subseteq X$ compact and $O \subseteq Y$ open. The set

$$U^K_{K,O} = \{f \in \mathcal{F}|_K : f(K) \subseteq O\}$$

is $\mathcal{P}$-open in Y^K by (). The restriction map $\rho : \mathcal{F} \to \mathcal{F}|_K$ is $\mathcal{P}$-continuous, so $\rho^{-1}(U^K_{K,O}) = U_{K,O}$ is $\mathcal{P}$-open.

So $\bar{\mathcal{F}} \subseteq \tilde{\mathcal{F}}$ is $\mathcal{P}$-closed, hence $\mathcal{P}$-compact. Since $\mathcal{F}$ is dense in both, it follows that $\bar{\mathcal{F}} = \tilde{\mathcal{F}}$.

For the converse, suppose $\mathcal{F}$ is $\mathcal{C}$-compact (we may assume $\mathcal{F}$ is closed since by XIII.17.3.10. $\bar{\mathcal{F}}$ also satisfies the hypotheses). Then it is also compact in the weaker $\mathcal{P}$-topology since $\mathcal{P}$ is Hausdorff, and the two topologies agree on $\mathcal{F}$. Thus for any $x \in X$ the set $\mathcal{F}[x]$ is compact since it is the continuous image of $\mathcal{F}$ under evaluation at x, which is a continuous function from $\mathcal{F}$ to Y.

Let K be a compact subset of X. We show that $\mathcal{F}|_K$ is evenly continuous. We have that $\mathcal{F}|_K$ is compact since it is a continuous image of $\mathcal{F}$. Let $x \in K$, $y \in Y$, and let U be an open neighborhood of y. Let W be an open neighborhood of y with $\bar{W} \subseteq U$. Let

$$\mathcal{E} = \{f \in \mathcal{F}|_K : f(x) \in \bar{W}\} .$$

Then $\mathcal{E}$ is a closed set in $\mathcal{F}|_K$, hence compact. The evaluation map $\rho : \mathcal{F}|_K \times K \to Y$ is continuous, so $\rho^{-1}(U)$ is open, and contains the compact set $\mathcal{E} \times \{x\}$. Thus by XIII.11.14.19. there is a (relatively) open set V in K such that $\mathcal{E} \times V \subseteq \rho^{-1}(U)$, i.e. if $f \in \mathcal{F}|_K$ satisfies $f(x) \in W$, then $f(V) \subseteq U$. 👍

XIII.17.3.18. In [Are46b] it is shown that if X is a completely regular space and the compact-open topology on $[0,1]^X$ is jointly continuous, then X is locally compact, suggesting that even continuity on compact subsets does not imply even continuity if X is not locally compact.. But I do not know an example of a k-space X, a regular space Y, and a $\mathcal{C}$-compact $\mathcal{F} \subseteq Y^X$ such that $\mathcal{F}$ is not evenly continuous.

Common Special Cases

A commonly stated version is just the special case of XIII.17.3.17. where the target space Y is completely regular, by XIII.15.1.27. and XIII.17.3.11.:

XIII.17.3.19. Theorem. [Arzelà-Ascoli Theorem, Uniform Version] Let X be a k-space and $(Y, \mathcal{V})$ a separated uniform space, and let $\mathcal{F} \subseteq Y^X$. Then $\mathcal{F}$ is subcompact in the compact-open topology (topology of u.c. convergence) if and only if

(i) $\mathcal{F}$ is equicontinuous on compact subsets of X.

(ii) $\mathcal{F}[x]$ is subcompact in Y for all $x \in X$.

If these conditions are satisfied, the compact-open closure $\bar{\mathcal{F}}$ coincides with the closure in the topology of pointwise convergence, and the topology of u.c. convergence coincides with the topology of pointwise convergence on $\bar{\mathcal{F}}$.

Note that if X is locally compact, then condition (i) is simply that $\mathcal{F}$ is equicontinuous (XIII.17.3.13.).

A commonly stated and used version is the case $Y = \mathbb{R}$:

XIII.17.3.20. THEOREM. [ARZELÀ-ASCOLI THEOREM, REAL VERSION] Let X be a k-space, and let $\mathcal{F} \subseteq \mathbb{R}^X = C_{\mathbb{R}}(X)$. Then $\mathcal{F}$ is subcompact in the topology of u.c. convergence if and only if $\mathcal{F}$ is pointwise bounded and equicontinuous on compact subsets of X.

If these conditions are satisfied, the closure $\bar{\mathcal{F}}$ in the topology of u.c. convergence coincides with the closure in the topology of pointwise convergence, and the topology of u.c. convergence coincides with the topology of pointwise convergence on $\bar{\mathcal{F}}$.

The same statement holds if $\mathbb{R}$ is replaced by $\mathbb{C}$ or $\mathbb{R}^n$ (or any metric space where closed bounded sets are compact).

If X is compact, u.c. convergence is uniform convergence. We state the real version, but XIII.17.3.19. can be restated in the same way.

XIII.17.3.21. THEOREM. [ARZELÀ-ASCOLI THEOREM, COMPACT VERSION] Let X be a compact Hausdorff space, and let $\mathcal{F} \subseteq \mathbb{R}^X = C_{\mathbb{R}}(X)$. Then $\mathcal{F}$ is subcompact in the topology of uniform convergence if and only if $\mathcal{F}$ is uniformly bounded and equicontinuous.

If these conditions are satisfied, the closure $\bar{\mathcal{F}}$ in the topology of uniform convergence coincides with the closure in the topology of pointwise convergence, and the topology of uniform convergence coincides with the topology of pointwise convergence on $\bar{\mathcal{F}}$.

In fact, under the hypotheses, if $\mathcal{F}$ is $\mathcal{C}$-closed, the following are equivalent:

(i) $\mathcal{F}$ is compact.

(ii) $\mathcal{F}$ is pointwise bounded and equicontinuous.

(iii) $\mathcal{F}$ is uniformly bounded and uniformly equicontinuous.

(To show (ii) $\Rightarrow$ (iii), use XIII.17.3.16. and joint continuity.)

We then get the "classic" version of the theorem:

XIII.17.3.22. THEOREM. [ARZELÀ-ASCOLI THEOREM, CLASSIC VERSION] Let X be a compact Hausdorff space, and let $\mathcal{F} \subseteq \mathbb{R}^X = C_{\mathbb{R}}(X)$. If $\mathcal{F}$ is uniformly bounded and equicontinuous, then any sequence in $\mathcal{F}$ has a uniformly convergent subsequence. Any sequence in $\mathcal{F}$ which converges pointwise converges uniformly.

To obtain this from XIII.17.3.21., note that the uniform topology on $C_{\mathbb{R}}(X)$ is metrizable if X is compact ().

XIII.17.3.23. Actually Arzelà's version was the case where X is a closed bounded interval in $\mathbb{R}$, and his proof left much to be desired. Subsequently, Ascoli stated a slightly more general version and gave a more satisfactory proof. We give a direct proof of XIII.17.3.22. if X is metrizable which is essentially the proof of Arzelà and Ascoli (they assumed uniform equicontinuity, which follows from XIII.17.3.16.):

Proof: Let ρ be a metric on X giving its topology. X is separable by (). Let $D = \{x_1, x_2, \dots\}$ be a countable dense set in X. If (f_n) is a sequence in $\mathcal{F}$, then the sequence $(f_n(x_1))$ is bounded since $\mathcal{F}$ is pointwise bounded, so has a convergent subsequence $(f_{k_{n,1}}(x_1))$ by the Bolzano-Weierstrass Theorem. The sequence $(f_{k_{n,1}}(x_2))$ has a convergent subsequence $(f_{k_{n,2}}(x_2))$. Continuing inductively, we successively construct subsequences $(f_{k_{n,m}})$ for each m such that $f_{k_{n,m}}(x_m))$ converges; then $(f_{k_{n,m}}(x_r))$ converges for $1 \leq r \leq m$.

Construct the "diagonal" sequence $(f_{k_{n,n}})$. This is a subsequence of (f_n). We claim that $(f_{k_{n,n}})$ converges uniformly on X. We have that $(f_{k_{n,n}}(x_r))$ converges for all r. Let $\epsilon > 0$. Since X is compact, $\mathcal{F}$ is uniformly equicontinuous by XIII.17.3.16.; so there is a $\delta > 0$ such that whenever $x, y \in X$ and $\rho(x, y) < \delta$, then $|f(x) - f(y)| < \frac{\epsilon}{3}$ for all $f \in \mathcal{F}$. Since X is compact, it can be covered by finitely many open balls of radius $\frac{\delta}{2}$. Thus there is a p such that every $x \in X$ satisfies $\rho(x, x_j) < \delta$ for some j, $1 \leq j \leq p$. There is an n_0 such that

$$|f_{k_{n,n}}(x_j) - f_{k_{m,m}}(x_j)| < \frac{\epsilon}{3}$$

for all $n, m > n_0$ and $1 \leq j \leq p$ (only finitely many such j's).

Let $x \in X$ and $n, m > n_0$. Choose j, $1 \leq j \leq p$, with $\rho(x, x_j) < \delta$. Then

$$|f_{k_{n,n}}(x) - f_{k_{m,m}}(x)| \leq |f_{k_{n,n}}(x) - f_{k_{n,n}}(x_j)| + |f_{k_{n,n}}(x_j) - f_{k_{m,m}}(x_j)| + |f_{k_{m,m}}(x_j) - f_{k_{m,m}}(x)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon\,.$$

Thus $(f_{k_{n,n}})$ is a uniform Cauchy sequence, hence converges uniformly (). 👍

XIII.17.3.24. Example. A subset of even $[0,1]^{[0,1]}$ can be $\mathcal{P}$-compact and not $\mathcal{C}$-compact (and thus not equicontinuous). Let (f_n) be the sequence of IV.3.2.4.(iii) with peaks of height 1, and let ✋

$$\mathcal{F} = \{f_n : n \in \mathbb{N}\} \cup \{0\}$$

where 0 is the function which is identically 0. More generally, let g be a continuous function from $[0,1]$ to $[0,1]$ which is not identically 0, but with $g(0) = g(1) = 0$, and set $f_n(t) = g(t^n)$. Then $\mathcal{F} = \{f_n : n \in \mathbb{N}\} \cup \{0\}$ is $\mathcal{P}$-compact, but not $\mathcal{C}$-compact or equicontinuous. Of course, the topology of pointwise convergence does not agree with the compact-open topology (topology of uniform convergence) on such an $\mathcal{F}$.

Arzelà-Ascoli and the AC

XIII.17.3.25. It is clear that the general Arzelà-Ascoli Theorem () must require at least the BPI, since Tikhonov's Theorem for Hausdorff spaces is a special case (the case where X is discrete and Y is compact Hausdorff, and $\mathcal{F}$ is all of Y^X, cf. XIII.11.14.23.).

XIII.17.3.26. The direct proof of the classic version given in XIII.17.3.23. seems to require DC to choose the subsequences. But the proof of the Bolzano-Weierstrass Theorem in IV.1.7.1. gives a recipe for specifying a convergent subsequence of a bounded sequence of real numbers which does not require any choice. Since the rest of the proof only requires finitely many choices (the metric ρ and the dense sequence D; all other choices can be made systematically), no Choice is needed for this proof.

XIII.17.4. Exercises

XIII.17.4.1. (a) Let $X = \mathbb{N}^\dagger$, the one-point compactification of $\mathbb{N}$, and let $Y = [0,1]$. Let $f_n : X \to Y$ be the function which is 1 on $\{1,\dots,n\}$ and 0 otherwise, and let f be the function which is 1 on $\mathbb{N}$ and 0 at ∞. Then $f_n \to f$ pointwise, each f_n is continuous, but f is discontinuous. Obviously (f_n) does not converge to f uniformly, hence does not converge uniformly on compact sets since X is compact. Show that $f_n \to f$ in the compact-open topology on $F(X,Y)$.

(b) Let $X = \mathbb{N} \cup \{\alpha\}$ be the space of XIII.11.14.10., and let $Y = [0,1]$. The only compact subsets of X are the finite subsets; thus uniform convergence on compact sets is the same as pointwise convergence. If f_n is defined as above, and f is 1 on $\mathbb{N}$ and $f(\alpha) = 0$, then $f_n \to f$ uniformly on compact sets (pointwise), but each f_n is continuous and f is discontinuous. In this case, $f_n \to f$ also in the compact-open topology on $F(X,Y)$.

(c) Let $X = \mathbb{N}^\dagger$ and $Y = \mathbb{R}$. Define $f : X \to Y$ by $f(k) = k$ for $k \in \mathbb{N}$, and $f(\infty) = 0$. Let $f_n(k) = k - \frac{1}{n}$, and $f_n(\infty) = 0$. Then $f_n \to f$ uniformly. Let

$$O = (-1,1) \cup \bigcup_{k\in\mathbb{N}} \left(k - \frac{1}{k}, k + \frac{1}{k}\right) .$$

Then O is open in Y. We have $f \in U_{X,O}$, but $f_n \notin U_{X,O}$ for any n. Thus (f_n) does not converge to f in the compact-open topology.

With some care, this example can be modified to make $Y = [0,1]$.

(d) Show that the proof of Proposition XIII.17.2.11. shows the following:

PROPOSITION. Let (f_i) be a net of functions from a topological space X to a metric space (Y,ρ), and let f be a function from X to Y. If f is continuous on each compact subset of X, then $f_i \to f$ in the compact-open topology if and only if $f_i \to f$ uniformly on each compact subset K of X.

Note that there is no continuity assumption (or any other assumption) on the f_i. Even the limit function need not be continuous on all of X. (If X is first countable, then continuity on compact subsets implies continuity on all of X (); however, this can fail in general, as in (b).)

(e) If each f_i is continuous (or just continuous on compact sets), and $f_i \to f$ uniformly on compact sets, then f is continuous on compact sets (), and thus $f_i \to f$ in the compact-open topology.

Note that problem 8b in section 9.1 of [**?**] is wrong as stated.

XIII.18. Cantor Sets

The Cantor set, and its variations, is a complicated but interesting closed subset of the unit interval in $\mathbb{R}$. It is well worth while to develop a good understanding of this set, since it is an important gateway to a number of varied aspects of analysis and topology.

XIII.18.1. The Cantor Set

XIII.18.2. Other Cantor Sets

The construction of the Cantor set can be obviously generalized: at each step we only need to remove an open interval from the interior of each of the remaining closed intervals at that step; it does not have to be exactly the middle third (the lengths of the remaining intervals must just approach zero). By varying the lengths of the removed intervals, we can obtain versions of the Cantor set with various measure-theoretic properties.

Symmetric Cantor Sets

We first consider the case where all removed intervals at each stage are centered and of the same length, so that all the remaining intervals at each step will have the same length. This restriction is not necessary, but without it notation becomes unwieldy, and the Cantor sets obtained with the restriction, while quite general, retain a pleasant amount of symmetry and homogeneity.

XIII.18.2.1. Let $\alpha = (\alpha_0, \alpha_1, \alpha_2, \dots)$ be a sequence of real numbers such that $\alpha_0 = 1$ and $0 < \alpha_n < \frac{\alpha_{n-1}}{2}$ for $n \geq 1$. Begin with $K_0 = [0,1]$. Remove an open interval from the middle of K_0 to obtain the set K_1 consisting of two closed intervals of length α_1; the removed interval has length $1 - 2\alpha_1 = \alpha_0 - 2\alpha_1$. Continue inductively: at the n'th stage remove open intervals from the middle of each of the closed intervals in K_{n-1} to obtain 2^n closed intervals of length α_n; the 2^{n-1} removed intervals each have length $\alpha_{n-1} - 2\alpha_n$. Let K_n be the union of the resulting intervals. We have that K_n is a closed subset of $[0,1]$, hence compact (), and $K_{n+1} \subseteq K_n$ for all n.

XIII.18.2.2. Definition. The set $K_\alpha = \cap_n K_n$ is the *Cantor set of type* α.

XIII.18.2.3. We thus obtain a large variety of Cantor sets. The Cantor set of () is the Cantor set of type α, where $\alpha_n = 3^{-n}$ for all n. This set will simply be denoted K and called "the Cantor set," i.e. the phrase "the Cantor set" without additional qualification always means the standard one.

All Cantor sets constructed in this way are topologically the same as the Cantor set, and embedded in $[0,1]$ in the same way:

XIII.18.2.4. Proposition. Let K_α be any Cantor set. Then there is a strictly increasing continuous function $\phi : [0,1] \to [0,1]$ with $\phi(K) = K_\alpha$. In particular, K_α is homeomorphic to K.

We always have $\alpha_n < 2^{-n}$, so $\alpha_n \to 0$ rather rapidly. The faster the α_n converge to 0, the "smaller" the set K_α is. By choosing this rate carefully, one can obtain arbitrary Hausdorff and packing dimensions for K_α:

XIII.18.2.5. THEOREM. Let K_α be a Cantor set of type $\alpha = (\alpha_0, \alpha_1, \alpha_2, \dots)$. Then

(i) The Hausdorff dimension () of K_α is

$$Dim_H(K_\alpha) = \liminf_{n\to\infty} \frac{n \log 2}{-\log \alpha_n} \ .$$

(ii) The packing dimension () of K_α is

$$Dim_P(K_\alpha) = \limsup_{n\to\infty} \frac{n \log 2}{-\log \alpha_n} \ .$$

In particular, if $0 \le s_H \le s_P \le 1$, there is a K_α with $Dim_H(K_\alpha) = s_H$ and $Dim_P(K_\alpha) = s_P$.

XIII.18.2.6. EXAMPLES. (i) If $0 < s < 1$, set $\alpha_n = 2^{-n/s}$ for all n. Then $Dim_H(K_\alpha) = Dim_P(K_\alpha) = s$. (ii) If $\alpha_n = 2^{-n^2}$ for all n, then $Dim_H(K_\alpha) = Dim_P(K_\alpha) = 0$.

XIII.18.2.7. If $Dim_H(K_\alpha) = s$, this result does not give information about $\mathcal{H}^s(K_\alpha)$. By a more delicate analysis, we can make this measure arbitrary as long as $0 < s < 1$ (Since $\mathcal{H}^0$ is counting measure, we have $\mathcal{H}^0(K_\alpha) = \infty$ for all α, and $\mathcal{H}^1$ is Lebesgue measure on $\mathbb{R}$, so $\mathcal{H}^1(K_\alpha) \le \mathcal{H}^1([0,1]) = 1$ for all α). We can do the same for packing measure. We have:

If $Dim_H(K_\alpha) = 1$, we can make K_α have positive Lebesgue measure:

XIII.18.2.8. EXAMPLE. Let $0 < d < 1$. We will construct a K_α such that $\lambda(K_\alpha) = d$. Set

$$c = \frac{3 - 2d}{1 - d} \in (3, \infty)$$

and note that $d = \frac{c-3}{c-2}$. Choose the α_n so that the removed open intervals at the n'th stage have length c^{-n}. This can be done inductively by setting $\alpha_0 = 1$ and solving $c^{-n} = \alpha_n - 2\alpha_{n-1}$ for α_n. If $b = c^{-1}$, we have successively

$$\alpha_1 = \frac{1-b}{2}, \quad \alpha_2 = 1 - b + b^2, \quad \alpha_3 = 2 - 2b + 2b^2 + b^3, \ \cdots$$

At the n'th step, 2^{n-1} intervals of length c^{-n} are removed; hence the complement of K_α in $[0,1]$ has Lebesgue measure

$$\sum_{n=1}^{\infty} \frac{2^{n-1}}{c^n} = \frac{1}{2} \sum_{n=1}^{\infty} \left[\frac{2}{c}\right]^n = \frac{1}{2} \cdot \frac{\frac{2}{c}}{1 - \frac{2}{c}} = \frac{1}{c-2}$$

since the removed intervals are disjoint, and thus we have

$$\lambda(K_\alpha) = 1 - \frac{1}{c-2} = \frac{c-3}{c-2} = d \ .$$

Denote this Cantor set by K_d. We could also take $d = 0$ in this construction ($c = 3$) to get the ordinary Cantor set, although this set does not have Hausdorff dimension 1.

Note that there is no K_α with $\lambda(K_\alpha) = 1$, since for any α we have that $[0,1] \setminus K_\alpha$ is a nonempty open set and thus has strictly positive Lebesgue measure.

More general Hausdorff and packing dimensions

Asymmetric Cantor Sets

For Cantor sets with Lebesgue measure zero, there is a feasible alternate notation which works also in the asymmetric case.

XIII.18.2.9. Let $\lambda = (\lambda_k)$ be a sequence of strictly positive numbers with $\sum_{k=1}^{\infty} \lambda_k = 1$. At the first stage, remove a suitable open interval of length λ_1 from the interior of $I_1 = [0,1]$. It will turn out that there is a uniquely located such subinterval for which the subsequent process works, not necessarily centered in $[0,1]$. This leaves two closed subintervals I_2 and I_3 (numbered from left to right). Remove a suitable open subinterval of length λ_2 from I_2, and one of length λ_3 from I_3, leaving four closed intervals $I_4 - I_7$. Continue the process, and let C_λ be the remaining set at the end. Then C_λ is a Cantor set of Lebesgue measure zero (Exercise ()).

XIII.18.2.10. To make the process work, the lengths of the remaining intervals at each stage must be sums of suitable subseries of the original series. The length of I_2 must be $\sum \lambda_k$, where the k are all numbers of the form $2^m + j$, $0 \le j \le 2^{m-1} - 1$ for $m \ge 1$, and the length of I_3 must be $\sum \lambda_k$, where the k are of the form $2^m + j$, $2^{m-1} \le j \le 2^m - 1$, $m \ge 1$, i.e.

$$\ell(I_2) = \lambda_2 + \lambda_4 + \lambda_5 + \lambda_8 + \lambda_9 + \lambda_{10} + \lambda_{11} + \lambda_{16} + \cdots$$

$$\ell(I_3) = \lambda_3 + \lambda_6 + \lambda_7 + \lambda_{12} + \lambda_{13} + \lambda_{14} + \lambda_{15} + \lambda_{24} + \cdots$$

and these uniquely determine I_2 and I_3 (note that $1 = \ell(I_1) = \lambda_1 + \ell(I_2) + \ell(I_3)$) and hence uniquely determine the interval which must be removed from I_1. Lengths of subsequent intervals are uniquely determined in the same manner.

XIII.18.2.11. The Cantor set C_λ depends on the ordered sequence λ. In rare cases, a reordering of the sequence can yield the same Cantor set; two sequences yielding the same Cantor set must be reorderings of each other.

XIII.18.2.12. The sets K_α which have Lebesgue measure 0 are among the C_λ. In fact, if (α_n) is given, take $\lambda_k = \alpha_{n-1} - 2\alpha_n$ for $2^{n-1} \le k < 2^n$.

More generally, any Cantor set of Lebesgue measure 0 in $[0,1]$ containing 0 and 1 is a C_λ. If C is such a Cantor set, then $U_1 = [0,1] \setminus C$ is a countable union of disjoint open intervals. Let λ_1 be the maximum length of the disjoint subintervals. Remove an interval of length λ_1 from U, leaving open sets U_2 and U_3 on either side. Let λ_2 be the maximum length of an open subinterval in U_2, etc. This generates a sequence (λ_k), and it is easily seen that $C = C_\lambda$ (Exercise ()).

The same construction can be done in any interval $[a,b]$ of $\mathbb{R}$, starting with a series $\sum_{k=1}^{\infty} \lambda_k$ whose sum is $b - a$. Every Cantor subset of $\mathbb{R}$ of Lebesgue measure 0 is such a set, i.e. a translated and scaled version of a C_λ.

The Hausdorff measure and Hausdorff dimension of the C_λ have been studied by a number of authors: [Bor49], [BT54], [CMMS04], [CMPS05], [CHM10].

XIII.18.3. Exercises

XIII.18.3.1. (i) Show that for every $a \in [0,1]$, there is a $b \in [0,1]$ such that $a+b$ and $a-b$ are in K. [Write a in its ternary expansion. Let the n'th digit of the ternary expansion of b be 1 if the n'th digit of the expansion of a is 1, and 0 otherwise.]

(ii) Show that every number in $[0,2]$ can be written as the sum of two elements of K. [If $x \in [0,2]$, set $a = x/2$ and choose b as in (i).] How many ways are there of doing it?

(iii) Conclude from (ii), (), and () that the Hausdorff dimension of K is at least $1/2$. (In fact, by XIII.18.2.5., the Hausdorff dimension of K is $\frac{\log 2}{\log 3} \approx 0.63$.)

XIII.18.4. Space-Filling Curves

In this section, we will show that there is a continuous function from $[0,1]$ *onto* the square $[0,1]^2$. This remarkable fact was first discovered by Peano in -. There are actually many ways to construct such a "space-filling curve"; our approach will be somewhat unconventional, exploiting properties of the Cantor set. For other constructions of space-filling curves, see, for example, [] or [] (or many other references.) Essentially the same construction will yield a continuous function from $[0,1]$ onto $[0,1]^n$ for any n, and even a continuous function from $[0,1]$ onto the Hilbert cube $[0,1]^{\mathbb{N}}$ (an "ultimate space-filling curve").

Recall that the Cantor set K is the subset of $[0,1]$ consisting of all numbers which have a triadic expansion consisting only of 0's and 2's. Every element of K has a unique such expansion. K is an uncountable compact subset of $[0,1]$ which is totally disconnected but without isolated points, which is nowhere dense and has Lebesgue measure 0.

The proof of the next proposition is straightforward and is left to the reader.

XIII.18.4.1. PROPOSITION. The map

$$(a_1, a_2, \dots) \mapsto .a_1 a_2 \cdots$$

is a homeomorphism from $\{0,2\}^{\mathbb{N}}$ onto K.

XIII.18.4.2. Thus K is homeomorphic to a countably infinite product of 2-point spaces. The key fact about K we will need, which follows immediately from this (since $\aleph_0 = n\aleph_0 = \aleph_0^2$), is that K is homeomorphic to $K^2 = K \times K$, and also to K^n for any n and even to $K^{\mathbb{N}}$. An explicit homeomorphism from K to K^2 is given by:

$$\rho(.a_1a_2a_3\cdots) = (.a_1a_3a_5\cdots, .a_2a_4a_6\cdots)$$

where the a_i are the triadic digits, either 0 or 2. A similar homeomorphism can be given from K to K^n for any n. To get a homeomorphism from K to $K^{\mathbb{N}}$, let γ be a bijection from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$, for example $\gamma(m,n) = 2^{m-1}(2n-1)$, and set

$$\sigma(.a_1a_2a_3\cdots) = (\sigma_1(.a_1a_2a_3\cdots), \sigma_2(.a_1a_2a_3\cdots), \sigma_3(.a_1a_2a_3\cdots), \cdots)$$

where $\sigma_m(.a_1a_2a_3\cdots) = .a_{\gamma(m,1)}a_{\gamma(m,2)}a_{\gamma(m,3)}\cdots$. So, for the above γ,

$$\sigma(.a_1a_2a_3\cdots) = (.a_1a_3a_5\cdots, .a_2a_6a_{10}\cdots, .a_4a_{12}a_{20}\cdots, \cdots)$$

XIII.18.4.3. There is a standard continuous function from K onto $[0,1]$, defined by

$$\theta(.a_1a_2a_3\cdots) = .b_1b_2b_3\cdots$$

where the a_n are triadic digits, either 0 or 2, $b_n = a_n/2$ (i.e. $b_n = 0$ if $a_n = 0$ and $b_n = 1$ if $a_n = 2$), and $.b_1b_2b_3\cdots$ is the *binary* expansion of the image number. This function θ is just the restriction of the Cantor function () to K.

Applying θ coordinatewise, we obtain continuous functions θ^n from K^n onto $[0,1]^n$, and $\theta^{\mathbb{N}}$ from $K^{\mathbb{N}}$ onto $[0,1]^{\mathbb{N}}$. Composing these with the homeomorphisms ρ and σ above, we obtain continuous surjective maps

$$\phi = \theta^2 \circ \rho : K \to [0,1]^2, \quad \psi = \theta^{\mathbb{N}} \circ \sigma : K \to [0,1]^{\mathbb{N}}$$

Let ϕ_n $(n = 1, 2)$ be the coordinate functions of ϕ, and ψ_n $(n \in \mathbb{N})$ the coordinate functions of ψ. By (), ϕ_n and ψ_n extend to continuous functions on all of $[0, 1]$ by making them linear on the intervals in $[0, 1] \setminus K$; also call these extensions ϕ_n and ψ_n respectively. Using these as coordinate functions, we obtain surjective continuous functions $\phi : [0, 1] \to [0, 1]^2$ and $\psi : [0, 1] \to [0, 1]^{\mathbb{N}}$, which is what we wanted. Thus:

XIII.18.4.4. THEOREM. There are continuous functions ϕ from $[0, 1]$ onto $[0, 1]^2$, and ψ from $[0, 1]$ onto $[0, 1]^{\mathbb{N}}$.

XIII.18.4.5. The functions we have defined have the following properties:

(1) The coordinate functions are differentiable on $[0, 1] \setminus K$, hence differentiable a.e.

(2) For any $(s_1, s_2) \in [0, 1]^2$, there is a $t \in [0, 1]$ with $\phi_1(t) = s_1$ and $\phi_2(t) = s_2$. Similarly, for any sequence $(s_1, s_2, \cdots)$ in $[0, 1]$, there is a $t \in [0, 1]$ with $\psi_n(t) = s_n$ for all n.

XIII.18.4.6. The functions ϕ and ψ cannot be injective, since if they were they would be homeomorphisms by (); but cubes of different dimensions are not homeomorphic (). However, ϕ is not terribly far from being injective: no point of $[0, 1]^2$ is traversed more than 5 times (the point $(\frac{1}{2}, \frac{1}{2})$ is traversed exactly 5 times, as are countably many other points). Individual coordinate functions are very far from being injective, however. (In fact, see XIII.18.5.3..)

XIII.18.4.7. The coordinate functions ϕ_n and ψ_n, although differentiable a.e., are not of bounded variation. The graphs of ϕ_1 and ϕ_2 look as follows:

The graphs have a somewhat fractal character in that the same patterns are repeated in miniature. However, they are not truly fractal since the scaling factors are different on the horizontal and vertical axes ($\frac{1}{3}$ on the horizontal, $\frac{1}{2}$ on the vertical).

XIII.18.5. Exercises

XIII.18.5.1. Describe the "motion" of $\phi(t)$ through the unit square as t goes from 0 to 1. [First describe $\phi(t)$ for $t \in [\frac{1}{3}, \frac{2}{3}]$, $t \in [\frac{1}{9}, \frac{2}{9}]$, $t \in [\frac{7}{9}, \frac{8}{9}]$, etc., and then find $\phi([0, \frac{1}{3}])$, $\phi([\frac{2}{3}, 1])$, $\phi([0, \frac{1}{9}])$, $\phi([\frac{2}{9}, \frac{1}{3}])$, etc.]

XIII.18.5.2. An occasional attempt is made to define a space-filling curve ω from $[0, 1]$ onto $[0, 1]^2$ by the formula

$$\omega(.b_1b_2b_3\cdots) = (.b_1b_3b_5\cdots, .b_2b_4b_6\cdots)$$

where the b_n (0 or 1) are *binary* digits. Show why this approach does not work. Try to modify the formula to give a well-defined continuous surjection.

XIII.18.5.3. Each of the coordinate functions of ϕ or ψ is a continuous function from $[0, 1]$ to $[0, 1]$ taking each value in $[0, 1]$ uncountably many times. [Fix $a \in [0, 1]$. For each $b \in [0, 1]$ there is a $t_b \in [0, 1]$ such that $\phi(t_b) = (a, b)$, so $\phi_1(t_b) = a$ for all b.]

XIII.18.6. A Plane Curve With Positive Area

It is easy to give examples of topologically one-dimensional, and even zero-dimensional, compact subsets of $\mathbb{R}^2$ which have positive planar (Lebesgue) measure (e.g. XIII.18.6.4.). In this section, we construct a simple closed curve in $\mathbb{R}^2$ (Jordan curve), i.e. a subset of $\mathbb{R}^2$ homeomorphic to S^1, which has positive planar measure.

Such a set, regarded as a curve, must be highly nondifferentiable: the image of a piecewise continuously differentiable curve (or, more generally, a Lipschitz function) has Hausdorff dimension 1, and hence plane measure 0 ().

The first example of a simple closed curve with positive planar measure was given by Osgood in 1902 ([Osg03], complete with color pictures!); another example was constructed by Besicovitch and Schoenberg in [BS61].

Cantor Sets

XIII.18.6.1. Recall one standard construction of Cantor sets in $\mathbb{R}$.

Let α be a real number ≥ 3. The Cantor set K_α is constructed by beginning with the closed unit interval $[0,1]$ and at the n'th stage removing the open subintervals of length α^{-n} centered in each of the 2^{n-1} remaining intervals; K_α is the intersection. K_α is a compact totally disconnected set without isolated points, homeomorphic to a countable product of two-point spaces.

Since the total length of the removed subintervals is

$$\sum_{n=1}^{\infty} \frac{2^{n-1}}{\alpha^n} = \frac{1}{\alpha} \sum_{m=0}^{\infty} \left(\frac{2}{\alpha}\right)^m = \frac{1}{\alpha(1-\frac{2}{\alpha})} = \frac{1}{\alpha-2}$$

we obtain that the (one-dimensional) Lebesgue measure of K_α is

$$1 - \frac{1}{\alpha-2} = \frac{\alpha-3}{\alpha-2}.$$

XIII.18.6.2. The usual Cantor set is K_3, just denoted K. The measure of K is zero. K consists precisely of all elements of $[0,1]$ with a ternary (base 3) decimal expansion consisting only of 0's and 2's, and every element of K has a unique such expansion.

XIII.18.6.3. Given any α and β, there is an evident homeomorphism of $\mathbb{R}$ (strictly increasing function) carrying K_α onto K_β. (Note that this shows the well-known fact () that a homeomorphism of $\mathbb{R}$ need not preserve sets of measure zero.)

XIII.18.6.4. If $K_\alpha^2 = K_\alpha \times K_\alpha$, then K_α^2 is a compact subset of $[0,1]^2$ homeomorphic to K (a specific homeomorphism will be described below.) The planar measure of K_α^2 is

$$\left(\frac{\alpha-3}{\alpha-2}\right)^2.$$

A Wandering Curve

We will construct a homeomorphic image of $[0,1]$ passing through every point of K^2_α; thus if $\alpha > 3$ the image will have positive planar measure.

For notational simplicity, we will describe the construction for $\alpha = 3$. The construction for $\alpha > 3$ can be done directly in an analogous way, or by simply composing the curve for $\alpha = 3$ with a homeomorphism of $\mathbb{R}^2$ sending K^2 onto K^2_α.

XIII.18.6.5. We first define an auxiliary function

$$\psi : \{0\} \cup [1/9, 2/9] \cup [1/3, 2/3] \cup [7/9, 8/9] \cup \{1\} \to [0,1]^2$$

by setting $\psi(0) = (0,0)$, $\psi(1/9) = (0, 1/3)$, $\psi(2/9) = (0, 2/3)$, $\psi(1/3) = (1/3, 2/3)$, $\psi(2/3) = (2/3, 2/3)$, $\psi(7/9) = (1, 2/3)$, $\psi(8/9) = (1, 1/3)$, $\psi(1) = (1,1)$, and letting ψ be linear (affine) on each of the subintervals. See Figure XIII.17.

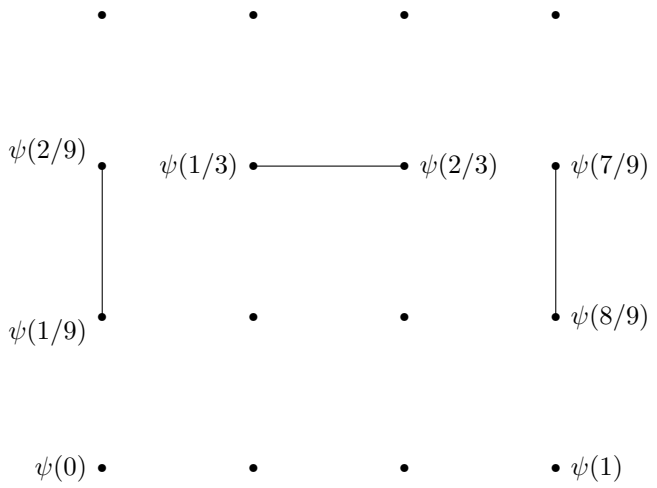


Figure XIII.17: The auxiliary function ψ

XIII.18.6.6. Before defining the curve $\phi : [0,1] \to [0,1]^2$, let us make a definition. A square S with sides parallel to the axes, of length 3^{-n} for some n, will be called *full* if the four corners of S are in the closure of the intersection of K^2 with the interior of S. The whole unit square is full; and if a full square is divided into nine subsquares, it is the four corner subsquares which are also full.

XIII.18.6.7. We will now define ϕ in stages. It is convenient to do the first step of the definition somewhat differently from the others, in order to have $\phi(0) = (1/3, 0)$ and $\phi(1) = (2/3, 0)$; then the segment from $(2/3, 0)$ back to $(1/3, 0)$ can be added to give a simple closed curve. (This segment will be included in the pictures.) So in addition to these definitions of $\phi(0)$ and $\phi(1)$, in the first step we define ϕ on $[1/9, 2/9]$ to map linearly to the segment from $\phi(1/9) = (1/3, 1/3)$ to $\phi(2/9) = (1/3, 2/3)$; define ϕ on $[1/3, 2/3]$ to

map linearly to the segment from $\phi(1/3) = (1/3, 1)$ to $\phi(2/3) = (2/3, 1)$, and similarly ϕ maps $[7/9, 8/9]$ linearly to the segment from $\phi(7/9) = (2/3, 2/3)$ to $\phi(8/9) = (2/3, 1/3)$.

XIII.18.6.8. We now proceed inductively. After the n'th step ϕ is defined except on open intervals of the form $\left(\frac{k}{9^n}, \frac{k+1}{9^n}\right)$, where k is 0,2,6, or 8 mod 9. For each such k, there is a unique full square S with adjacent corners $\phi(k/9^n)$ and $\phi((k+1)/9^n)$, and a unique affine map θ from the unit square onto S with $\theta(0,0) = \phi(k/9^n)$ and $\theta(1,0) = \phi((k+1)/9^n)$. Then for $t \in [1/9, 2/9] \cup [1/3, 2/3] \cup [7/9, 8/9]$ define

$$\phi\left(\frac{k+t}{9^n}\right) = \theta \circ \psi(t)$$

After the third step, the curve looks like Figure XIII.18.

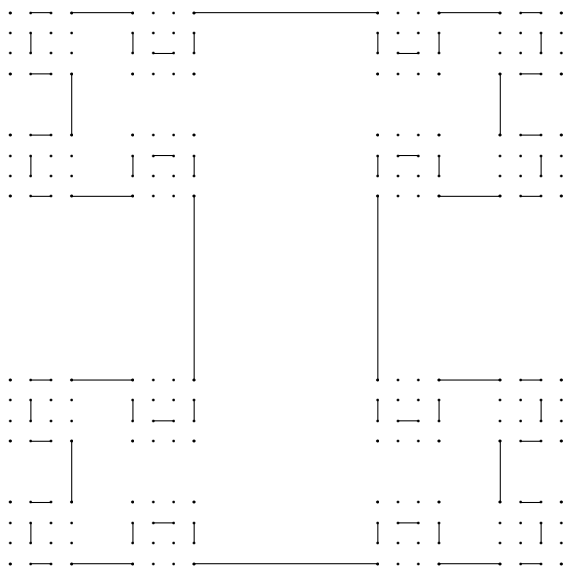


Figure XIII.18: Three steps of the definition of ϕ

XIII.18.6.9. After the inductive construction, ϕ is defined on a dense subset of $[0, 1]$. By construction, if $s, t \in [0, 1]$ and $|s - t| < 9^{-n}$, and $\phi(s)$ and $\phi(t)$ are defined, then $\|\phi(s) - \phi(t)\| < 3^{-n}\sqrt{2}$; thus ϕ is uniformly continuous and extends to a continuous map from all of $[0, 1]$ to $[0, 1]^2$.

XIII.18.6.10. PROPOSITION. The image of ϕ contains K^2.

PROOF: This is obvious, since after the n'th step in the construction every point of K^2 lies within $3^{-n}\sqrt{2}$ of a point on the curve. 👍

XIII.18.6.11. PROPOSITION. The map ϕ is a homeomorphism onto its image.

PROOF: By compactness, it is only necessary to show that ϕ is one-to-one. It clearly maps $[0,1]\backslash K$ one-one into $[0,1]^2\backslash K^2$, so we need only show that ϕ maps K one-one into (actually onto) K^2. This is geometrically clear, but is rather complicated to write down explicitly. In fact, there is a precise (but complicated) formula giving the first n places of the ternary decimal expansion of the two coordinates of $\phi(t)$ $(t \in K)$ in terms of the first $2n$ places of the ternary decimal expansion of t. For example, if $t = .a_1a_2\cdots$, $\phi(t) = (.b_1\cdots, .c_1\cdots)$, where b_1 is always equal to a_1 and $c_1 = a_2$ if $a_1 = 0$, $c_1 = 2 - a_2$ if $a_1 = 2$. It can be checked that if $s, t \in K$ differ in the first $2n$ ternary decimal places, then $\phi(s)$ and $\phi(t)$ differ in the first n ternary decimal places in at least one coordinate. 👍

XIII.18.6.12. The final curve is depicted in Figure XIII.19. The curve has a fractal character in that all of the full squares with sides of length $\leq 1/3$ are similar up to scaling and rotation.

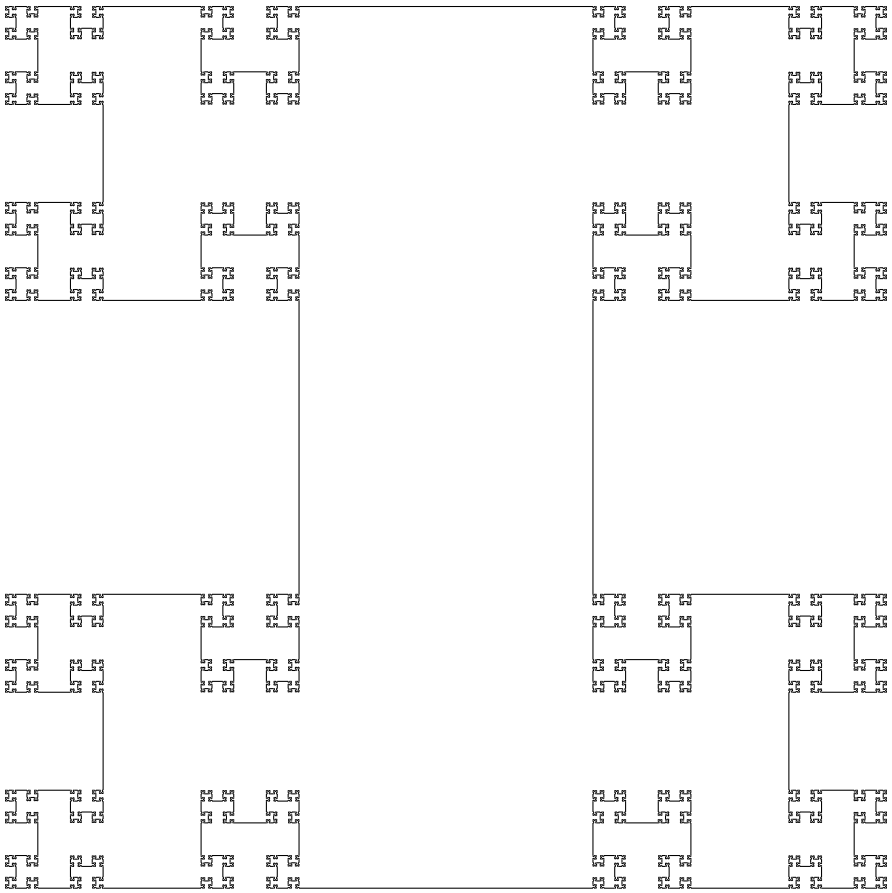


Figure XIII.19: The closed curve ϕ

Connection with Space-Filling Curves

XIII.18.6.13. This construction bears a close resemblance to one version of the Peano space-filling curve (see, for example, [Edg08, p. 64].) We obtain a space-filling curve by composing ϕ with $p \times p$, where p is the usual map from K onto $[0,1]$ given by

$$p(.a_1a_2a_3\cdots) = .b_1b_2b_3\cdots$$

where $.a_1a_2\cdots$ is the ternary decimal expansion of $t \in K$, and $.b_1b_2\cdots$ is the binary decimal expansion of $p(t)$, with $b_n = a_n/2$. The resulting map $(p \times p) \circ \phi$ maps $[0,1]$ continuously onto $[0,1]^2$. (It should be thought of as being constant on the intervals of $[0,1]\backslash K$.)

XIII.18.6.14. The map p is not one-to-one since the binary decimal expansion of dyadic rational numbers is not unique, and any cross section $q : [0,1] \to K$ for p is necessarily discontinuous; however, the map $\sigma = (p \times p) \circ \phi \circ q$ is continuous and independent of q, and maps $[0,1]$ onto $[0,1]^2$. This map σ is a more usual type of space-filling curve.

Summary and Variations

XIII.18.6.15. If C_α is the image of the curve ϕ done for the Cantor set K_α, with the last segment filled in between the endpoints, then C_α is a simple closed curve, contained in $[0,1]^2$, with planar measure at least the measure of K_α^2. Actually, C_α is the union of K_α^2 and a countable number of straight line segments, so the measure of C_α is exactly

$$\left(\frac{\alpha-3}{\alpha-2}\right)^2.$$

This can be any (nonnegative) number less than 1. If α is large enough ($\alpha > 4+\sqrt{2}$ will do), the measure of this curve is greater than the measure of the region inside the curve!

There cannot be a simple closed curve in $[0,1]^2$ of planar measure exactly one, since any nonempty open set in the plane has positive measure.

XIII.18.6.16. The curve C_α is a Jordan curve, but the region inside the curve is an open set in $\mathbb{R}^2$ (homeomorphic to an open disk) which is not a Jordan region () if $\alpha > 3$ since its boundary does not have measure zero.

XIII.18.6.17. The example in [BS61] has the additional property that for any s, t, $0 \le s < t \le 1$, the arc from $\phi(s)$ to $\phi(t)$ has positive planar measure. Our example can be modified to have this property, by replacing each straight line segment by countably many small versions of the curve.

XIII.18.6.18. It would appear that a construction similar to the one of this section can be done in $[0,1]^n$ for any n, giving a simple closed curve passing through every point of K_α^n and thus having positive n-dimensional Lebesgue measure. However, the details might be complicated to write carefully. Perhaps a similar construction can even be done in the Hilbert cube. See [PS92] for a version of Osgood's example in $\mathbb{R}^n$.

XIII.19. The Topology of $\mathbb{R}^n$

XIII.19.1. Invariance of Domain and the Jordan Curve Theorem

In this subsection, we describe two remarkable results about Euclidean space. Both of these fall into the category of facts which at first glance are "intuitively obvious," but which become highly nonobvious when thought about more carefully.

We begin with the original Jordan Curve Theorem in $\mathbb{R}^2$.

XIII.19.1.19. DEFINITION. A *Jordan curve* in $\mathbb{R}^2$ is the image of a one-to-one continuous function from the circle S^1 to $\mathbb{R}^2$. A Jordan curve is also called a *simple closed curve*.

A Jordan curve can also be defined as the image of a continuous function γ from a closed bounded interval $[a, b]$ into $\mathbb{R}^2$ which is one-to-one except that $\gamma(a) = \gamma(b)$. There is no assumption of smoothness or differentiability. Such a "parametrization" of the curve is highly nonunique and is not part of the structure of the curve.

A Jordan curve is intuitively a closed loop in the plane which does not cross itself. In fact, since S^1 is compact and Hausdorff, a Jordan curve is homeomorphic to S^1 by (). (Conversely, any subset of $\mathbb{R}^2$ which is homeomorphic to S^1 is trivially a Jordan curve.) Geometric circles, ellipses, and polygons are elementary examples of Jordan curves. A "figure 8" is not a Jordan curve. For a more complicated example, see Figure XIII.20. For more extreme examples, see Figures XIII.21, XIII.19, and XIII.18.6.17..

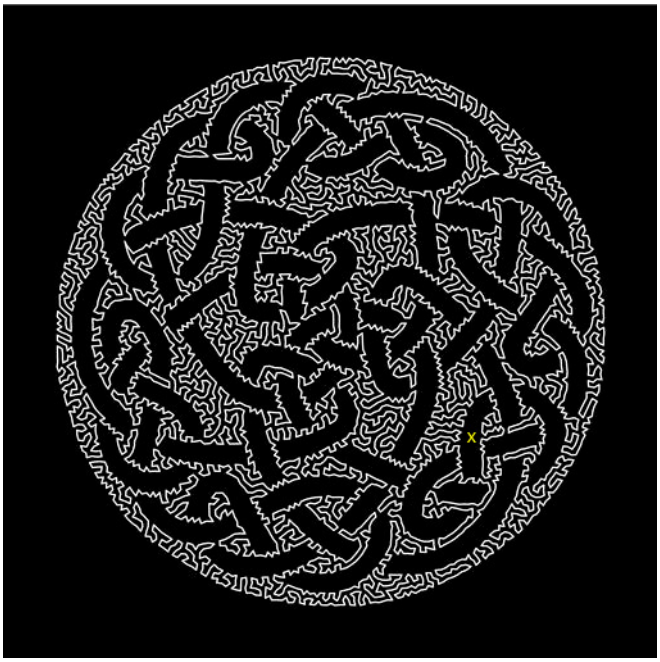


Figure XIII.20: A complicated Jordan curve (in white). Artwork by Robert Bosch

One would intuitively expect that a Jordan curve would divide the plane into two disjoint regions, the "inside" and the "outside" of the curve, just as a geometric circle does. But when more complicated examples are considered, this conclusion is less obvious; for example, is it obvious what is the "inside" and "outside" of the curve in Figure XIII.20? (In particular, is the point marked with the yellow x inside or outside? Try to find a path out from this point.) Nonetheless, it does turn out to be true:

XIII.19.1.20. THEOREM. [JORDAN CURVE THEOREM] Let C be a Jordan curve in $\mathbb{R}^2$. Then $\mathbb{R}^2 \setminus C$ has exactly two connected components, one bounded and one unbounded, whose common boundary is C.

This result is due to C. JORDAN in 1887. It is questionable whether JORDAN'S proof can be regarded as complete; certainly his exposition leaves much to be desired. Conventional wisdom says that JORDAN'S proof was flawed and that the first complete, correct proof was given by O. VEBLEN in 1905; however, this assertion may not stand up under examination [Hal07]. There have been many other proofs given over the years; see [Mae84] for a particularly nice one. We will not give a proof here. If the curve is polygonal or piecewise-smooth, the proof is relatively simple since such a curve locally divides the plane in two parts; however, the general case is difficult since a Jordan curve can be locally pathological such as the example in Figure XIII.21 (although it turns out to still locally divide the plane).

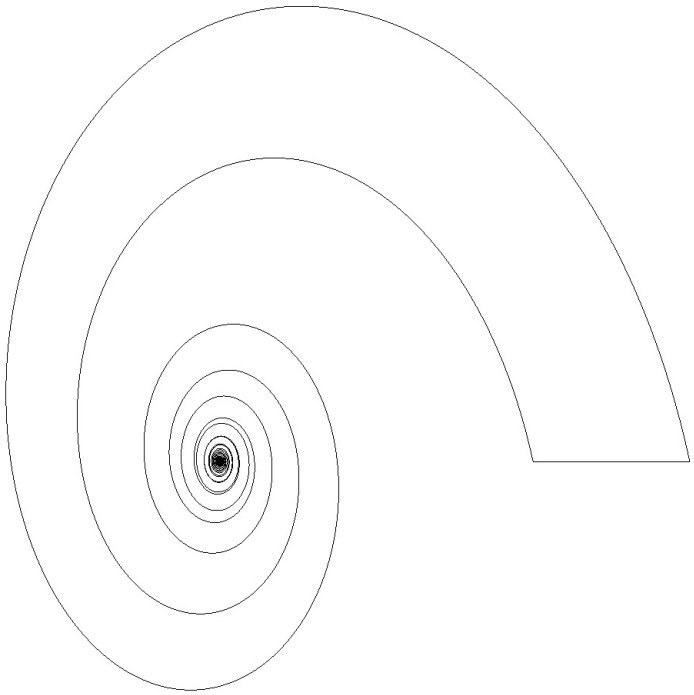


Figure XIII.21: A locally pathological Jordan curve

This curve, made of two intertwined spirals, is only locally pathological at one point, but small copies of the spirals can be successively added densely, replacing small arcs of the curve, to make a curve locally pathological at every point. (Much more complicated local pathology is possible too!)

There is a more refined version of the Jordan Curve Theorem, first proved by A. Schönflies in 1908(?) []. This result is commonly called the Jordan-Schönflies theorem.

XIII.19.1.21. Theorem. [Jordan-Schönflies Theorem] Let S^1 be the unit circle in $\mathbb{R}^2$, and $\gamma : S^1 \to \mathbb{R}^2$ a one-to-one continuous function. Then γ extends to a homeomorphism from $\mathbb{R}^2$ onto $\mathbb{R}^2$.

It is easy to see that the Jordan-Schönflies Theorem implies XIII.19.1.20. (Exercise ()). The conclusion is stronger: the "inside" (bounded component of the complement) is not only connected, but even homeomorphic to the open unit disk (and hence to $\mathbb{R}^2$ itself), and the "outside" (the unbounded component of the complement) is homeomorphic to $\mathbb{R}^2$ with a point removed. It really says that every subset of $\mathbb{R}^2$ homeomorphic to S^1 is topologically embedded in $\mathbb{R}^2$ in the same way.

See XII.8.6.8. for a proof using Complex Analysis, or see [CV98], for example, for a direct proof.

The Jordan Curve Theorem in $\mathbb{R}^n$

One natural approach to generalizing the Jordan curve theorem is to consider subsets of $\mathbb{R}^n$ homeomorphic to S^{n-1}. Such a subset could be called a *Jordan sphere* in $\mathbb{R}^n$ (this is nonstandard terminology, and it is questionable whether the name is really appropriate, so we will avoid using it).

XIII.19.1.22. One might hope in analogy with the Jordan-Schönflies Theorem that a homeomorphism from the unit sphere S^{n-1} in $\mathbb{R}^n$ to a subset C of $\mathbb{R}^n$ would extend to a homeomorphism from $\mathbb{R}^n$ onto $\mathbb{R}^n$, i.e. that all Jordan spheres in $\mathbb{R}^n$ are topologically embedded in the same way, but the example of the "Alexander horned sphere" in $\mathbb{R}^3$ (XIII.19.5.40.) shows that this statement is false in general, and thus the Jordan-Schönflies Theorem is special to $\mathbb{R}^2$. But remarkably, the Jordan Curve Theorem remains true:

XIII.19.1.23. Theorem. [Jordan Curve Theorem in $\mathbb{R}^n$] Let C be a subset of $\mathbb{R}^n$ ($n \geq 2$) homeomorphic to S^{n-1}. Then $\mathbb{R}^n \setminus C$ has exactly two connected components, one bounded and one unbounded, whose common boundary is C.

This theorem is due to Lebesgue and Brouwer. We again omit the proof; see, for example, [].

XIII.19.1.24. The statement is not quite true for $n = 1$. The "unit sphere" S^0 in $\mathbb{R}^1$ consists of two points $\{\pm 1\}$ whose complement has three connected components. But topologists usually state the theorem by replacing $\mathbb{R}^n$ by its one-point compactification S^n (the statement is easily seen to be equivalent; cf. Exercise ()), and this version also applies when $n = 1$.

There is a related result (which is used in the proof of XIII.19.1.23.). Let B^n be the closed unit ball in $\mathbb{R}^n$. Note that B^n is homeomorphic to $[0, 1]^n$.

XIII.19.1.25. THEOREM. Let B be a subset of $\mathbb{R}^n$ ($n \geq 2$) homeomorphic to B^n (i.e. to $[0,1]^n$). Then $\mathbb{R}^n \setminus B$ is connected.

The statement also holds with $\mathbb{R}^n$ replaced by S^n, even for $n = 1$.

XIII.19.1.26. It has been an active project in topology, not yet completed, to generalize the Jordan-Schönflies Theorem to higher dimensions. Restrictions on the embedding maps must, of course, be made. Perhaps the best result so far is that the statement is true for "locally flat" embeddings (embeddings which extend to a thin spherical shell) ([**?**], [**?**]). The result for smooth embeddings is known for $n \neq 4$, but is still open for $n = 4$.

XIII.19.1.27. One might also wonder about subsets of $\mathbb{R}^n$ (or S^n) homeomorphic to S^m for $m < n-1$. Such subsets always have connected complements, but the ways they can be embedded (even leaving out "wild" embeddings such as the Alexander Horned sphere, e.g. restricting to smooth or piecewise-linear embeddings) are very complicated. The study of smooth embeddings of S^1 into $\mathbb{R}^3$ is a fascinating and active branch of modern topology called *knot theory* (XIII.19.5.35.).

Brouwer's Invariance of Domain

A very important consequence of the Jordan Curve Theorem in $\mathbb{R}^n$ is the theorem that open sets in Euclidean space are invariant under homeomorphisms. This result, called *invariance of domain*, was proved by L. BROUWER in 1912 [].

XIII.19.1.28. THEOREM. [INVARIANCE OF DOMAIN] Let U be a nonempty open subset of $\mathbb{R}^n$, and f a one-to-one continuous function from U to $\mathbb{R}^m$. Then

(i) $m \geq n$.

(ii) If $m = n$, then $f(V)$ is open in $\mathbb{R}^n$ for every open subset V of U (in particular, $f(U)$ is open in $\mathbb{R}^n$), i.e. f is an open mapping and a homeomorphism onto its range.

Before outlining the proof, we list some of the most important consequences.

XIII.19.1.29. COROLLARY. Let U be an open subset of $\mathbb{R}^n$. Then any subset of $\mathbb{R}^n$ which is homeomorphic to U is also open.

The next corollary is especially noteworthy:

XIII.19.1.30. COROLLARY. Let U be a nonempty open subset of $\mathbb{R}^n$. Then U is not homeomorphic to any subset of $\mathbb{R}^m$ for any $m < n$. In particular, $\mathbb{R}^n$ is not homeomorphic to $\mathbb{R}^m$ for $n \neq m$.

This result also follows from the fact that any nonempty open set in $\mathbb{R}^n$ is topologically n-dimensional ().

XIII.19.1.31. Corollary. Let U be an open subset of $\mathbb{R}^n$, and $f : U \to \mathbb{R}^n$ a continuous function which is locally one-to-one. Then f is an open mapping.

XIII.19.1.32. We now outline the proof of XIII.19.1.28.; actually we give a complete proof assuming XIII.19.1.23. and XIII.19.1.25.. We first prove (ii), so assume $m = n$. Let $x \in U$, and fix $\epsilon > 0$ with $A = \bar{B}_\epsilon(x) \subseteq U$. Since A is compact and f is one-to-one on A, $B = f(A)$ is homeomorphic to B^n, and hence the open set $\mathbb{R}^n \setminus B$ is connected by XIII.19.1.25.. If $C = f(\partial A)$, then C is homeomorphic to $\partial A \cong S^{n-1}$, and hence the open set $\mathbb{R}^n \setminus C$ has two components. But $D = f(B_\epsilon(x)) = f(A \setminus \partial A)$ is connected since $B_\epsilon(x)$ is connected (), and D and $\mathbb{R}^n \setminus B$ are disjoint connected sets whose union is $\mathbb{R}^n \setminus C$, so they must be the two components. But the components of an open set are open, so D is open in $\mathbb{R}^n$ and contains $f(x)$. This argument can be repeated for any $x \in U$, so $f(U)$ is open. The same argument works for any open subset V of U. Thus the inverse map from $f(U)$ to U is continuous.

To prove (i), Suppose $m < n$. Define $\tilde{f} : U \to \mathbb{R}^n \cong \mathbb{R}^m \times \mathbb{R}^{n-m}$ by $\tilde{f}(x) = (f(x), 0)$. Then $\tilde{f}$ is one-to-one and continuous, but $\tilde{f}(U)$ is not open, a contradiction.

XIII.19.1.33. If U is an open set in $\mathbb{R}^n$ and f is a one-to-one continuous function from U to $\mathbb{R}^m$ ($m \geq n$), then the restriction of f to any compact subset K of U is a homeomorphism from K to $f(K)$ by (). But if $m > n$, f does not give a homeomorphism from U to $f(U)$ in general. For example, there is an evident one-to-one continuous map from $\mathbb{R}$ onto a figure 8 in $\mathbb{R}^2$. There are more dramatic examples: there is a one-to-one continuous function from $\mathbb{R}$ onto a dense subset of the 2-torus (or n-torus for any $n > 1$) obtained by "wrapping around" at an irrational angle (). See also XIII.19.5.19..

XIII.19.2. Orientation

$\mathbb{R}^n$, or any connected open set in $\mathbb{R}^n$, has two orientations. It is a significant and nontrivial fact that a homeomorphism between connected open sets in $\mathbb{R}^n$ must be either orientation-preserving or orientation-reversing everywhere.

XIII.19.2.1. This is easily seen if $n = 1$. If I is an open interval in $\mathbb{R}$, and f is a one-to-one continuous function from I to $\mathbb{R}$, then (V.3.10.15.) f must be either strictly increasing (orientation-preserving everywhere) or strictly decreasing (orientation-reversing everywhere) on I.

XIII.19.2.2. If $n > 1$, we first need a notion of a one-to-one continuous function being locally orientation-preserving or locally orientation-reversing at a point. If U is an open set in $\mathbb{R}^n$ and $\boldsymbol{f}$ a one-to-one continuous function from U to $\mathbb{R}^n$ (hence a homeomorphism onto an open set V in $\mathbb{R}^n$), and $\boldsymbol{x} \in U$, let B be a closed ball in U centered at $\boldsymbol{x}$ and S the boundary sphere of B; then $\boldsymbol{f}(B)$ is a topological n-ball in V with boundary $\boldsymbol{f}(S)$, a topological $(n-1)$-sphere, with $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ in the interior of $\boldsymbol{f}(B)$. We can associate an index with this map: by composing $\boldsymbol{f}$ on the left and right with translations, we may assume $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{0}$, and then composing with radial projections we obtain from $\boldsymbol{f}$ a continuous map from the unit sphere S^{n-1} to itself, which has an index $i(\boldsymbol{f}_S) \in \mathbb{Z}$.

XIII.19.2.3. Proposition. With the above notation, $i(\boldsymbol{f}_S)$ is either $+1$ or -1.

XIII.19.2.4. It is easily seen that, for fixed $\boldsymbol{x}$, the number $i(\boldsymbol{f}_S)$ does not depend on the choice of B (and hence S), since as the radius of B varies the induced maps from S^{n-1} to itself are homotopic. It could *a priori* depend on $\boldsymbol{x}$, so we will denote it $i_{\boldsymbol{f}}(\boldsymbol{x})$.

XIII.19.2.5. DEFINITION. The function $\boldsymbol{f}$ is *locally orientation-preserving* at $\boldsymbol{x}$ if $i_{\boldsymbol{f}}(\boldsymbol{x}) = +1$ and *locally orientation-reversing* at $\boldsymbol{x}$ if $i_{\boldsymbol{f}}(\boldsymbol{x}) = -1$.

XIII.19.2.6. THEOREM. Let U be a connected open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ a one-to-one continuous function from U to $\mathbb{R}^n$. Then $\boldsymbol{f}$ is either locally orientation-preserving at every point of U or locally orientation-reversing at every point of U.

PROOF: The index $i_{\boldsymbol{f}}(\boldsymbol{x})$ is easily seen to be locally constant in $\boldsymbol{x}$, i.e. $i_{\boldsymbol{f}}$ is a locally constant (continuous) function from U to $\{+1, -1\}$. Since U is connected, $i_{\boldsymbol{f}}$ must be constant. 👍

Smooth Maps

There is an important consequence for smooth maps. Suppose $\boldsymbol{f}$ is a one-to-one $\mathcal{C}^1$ function from an open set U in $\mathbb{R}^n$ to $\mathbb{R}^n$. Then $det(D\boldsymbol{f})$ is a continuous function from U to $\mathbb{R}$.

XIII.19.2.7. PROPOSITION. If $\boldsymbol{f}$ is locally orientation-preserving at $\boldsymbol{x} \in U$, then $det(D\boldsymbol{f}(\boldsymbol{x})) \geq 0$; if $\boldsymbol{f}$ is locally orientation-reversing at $\boldsymbol{x} \in U$, then $det(D\boldsymbol{f}(\boldsymbol{x})) \leq 0$.

PROOF: This is equivalent to the statement that if $det(D\boldsymbol{f}(\boldsymbol{x})) > 0$, then $\boldsymbol{f}$ is locally orientation-preserving at $\boldsymbol{x}$, and if $det(D\boldsymbol{f}(\boldsymbol{x})) < 0$, then $\boldsymbol{f}$ is locally orientation-reversing at $\boldsymbol{x}$, which follows from (). 👍

We then obtain, as a corollary of XIII.19.2.6.:

XIII.19.2.8. COROLLARY. Let U be a connected open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ a one-to-one $\mathcal{C}^1$ function from U to $\mathbb{R}^n$. Then either $det(D\boldsymbol{f}(\boldsymbol{x})) \geq 0$ for all $\boldsymbol{x} \in U$ or $det(D\boldsymbol{f}(\boldsymbol{x})) \leq 0$ for all $\boldsymbol{x} \in U$.

There is an alternate proof of XIII.19.2.8. using the Poincaré Lemma (). If $\boldsymbol{f}$ is a diffeomorphism onto its range, the result is trivial by the Intermediate Value Theorem, since $det(D\boldsymbol{f})$ is never zero. But the general result XIII.19.2.8. is definitely nontrivial. (Note that a one-to-one smooth map is not necessarily a diffeomorphism onto its range even if $n = 1$, e.g. $f(t) = t^3$.)

XIII.19.3. $\mathbb{R}^n$ is n-Dimensional

Everyone "knows" that $\mathbb{R}^n$ is n-dimensional. But what does this mean? Specifically, what does "dimension" mean? And how can $\mathbb{R}^n$ be distinguished from $\mathbb{R}^m$ for $m \neq n$ by its properties?

$\mathbb{R}^n$ is a set with several types of structure: algebraic, geometric, topological, and analytic. Efforts have been made to define and/or describe the "dimension" of $\mathbb{R}^n$ in terms of each of these types of structure, and various theories of dimension have been studied in different contexts; the only thing these various theories have in common is that they all assign the *same* number n to $\mathbb{R}^n$, justifying the use of the name "dimension" for all of them.

XIII.19.3.1. Specifically, we have:

(i) The vector space dimension () of $\mathbb{R}^n$ as a real vector space is n.

(ii) The topological dimension () of $\mathbb{R}^n$ as a topological space is n.

(iii) The Hausdorff dimension () of $\mathbb{R}^n$ is n.

XIII.19.3.2. There is one respect in which all Euclidean spaces are the same, however: cardinality. Contrary to Cantor's expectations (he first thought the cardinality of $\mathbb{R}^n$ should be $\aleph_n$), the cardinality of $\mathbb{R}^n$ is $2^{\aleph_0}$ for all n (). Thus some additional structure beyond set theory is needed to distinguish $\mathbb{R}^n$ from $\mathbb{R}^m$ for $n \neq m$.

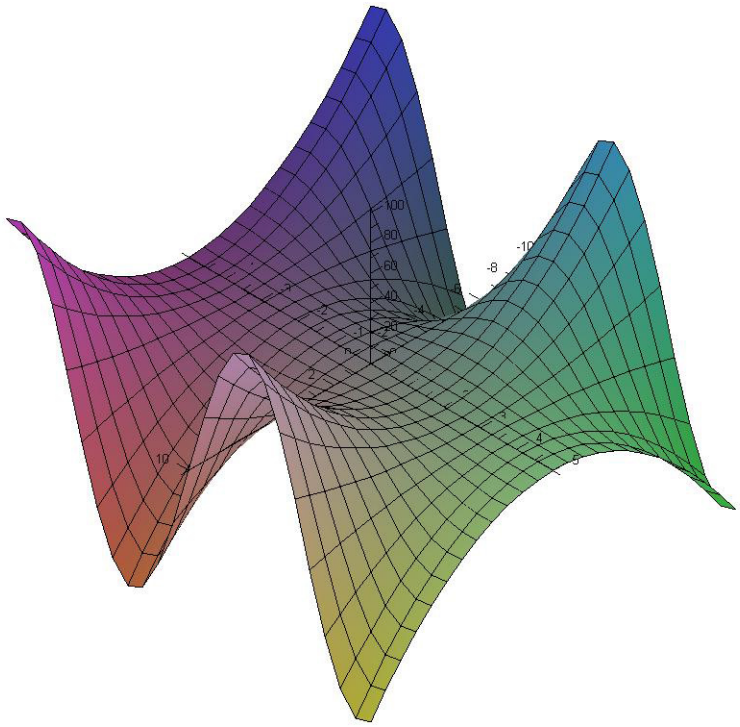


Figure XIII.22: Maple plot

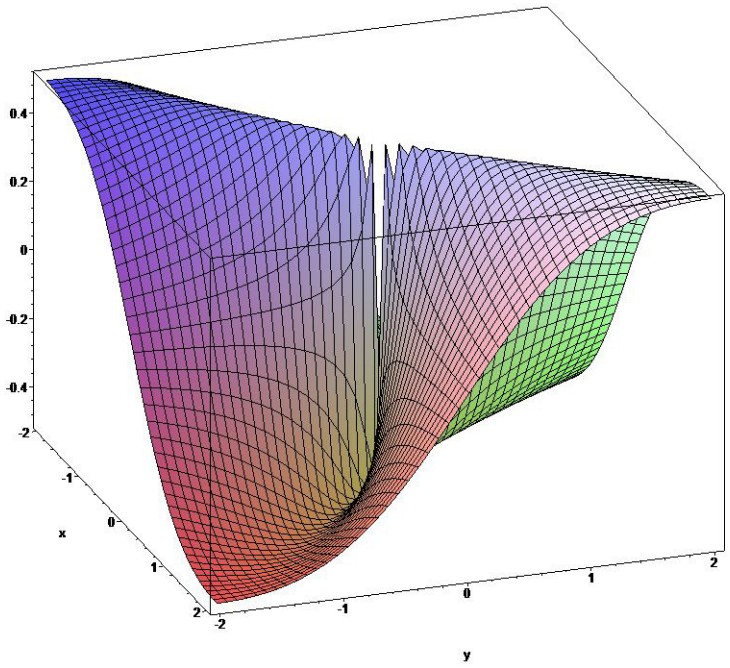


Figure XIII.23: Maple plot

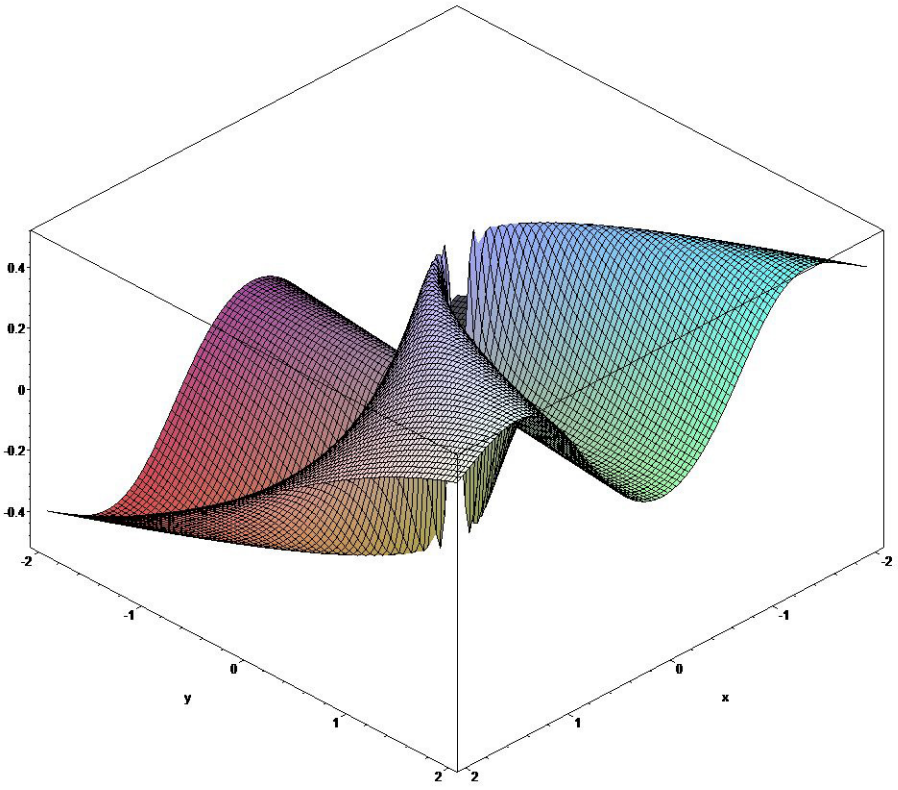


Figure XIII.24: Maple plot

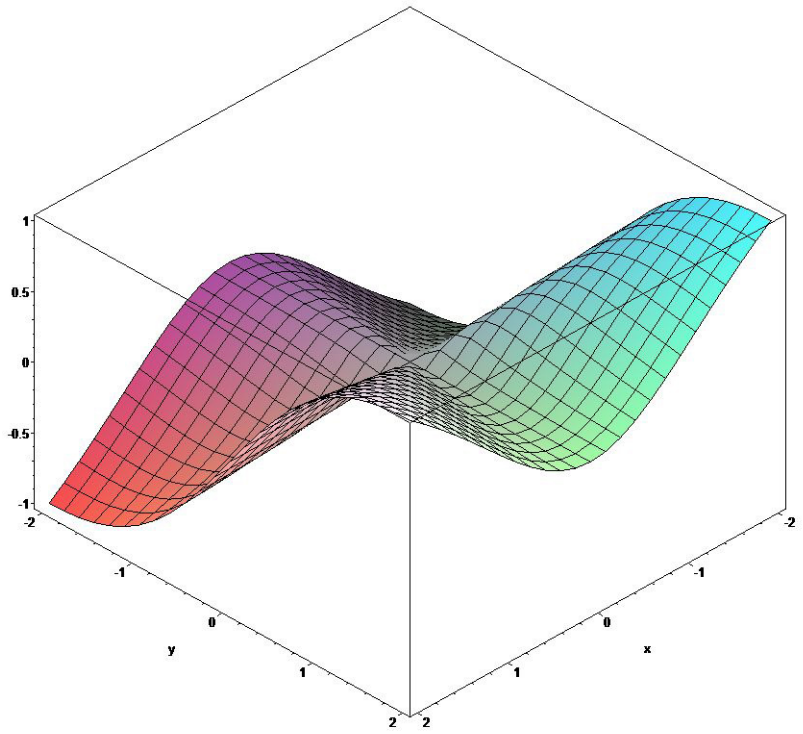


Figure XIII.25: Maple plot

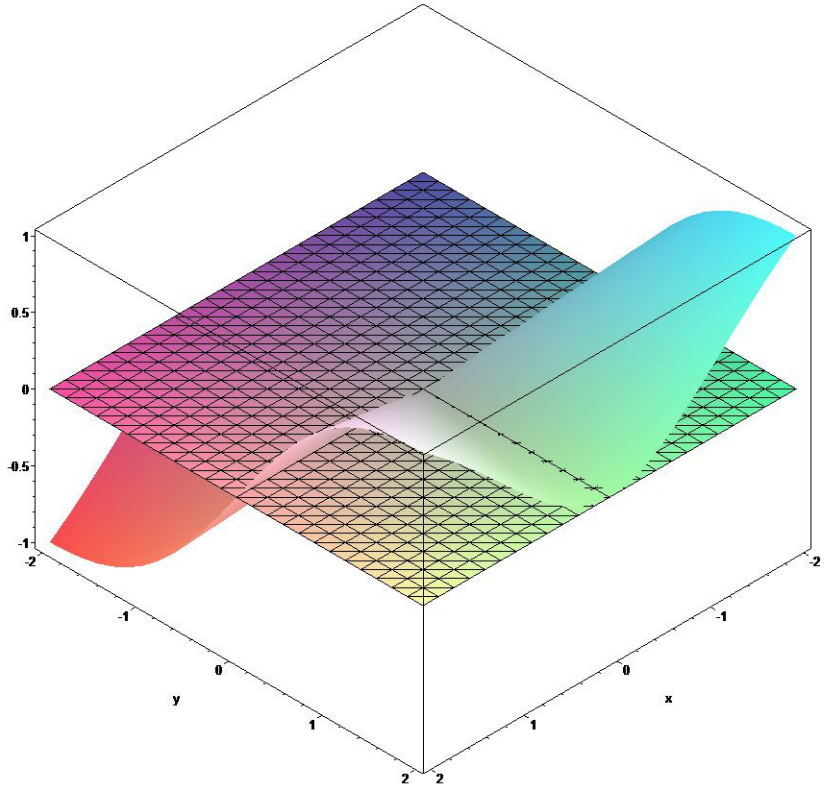


Figure XIII.26: Maple plot

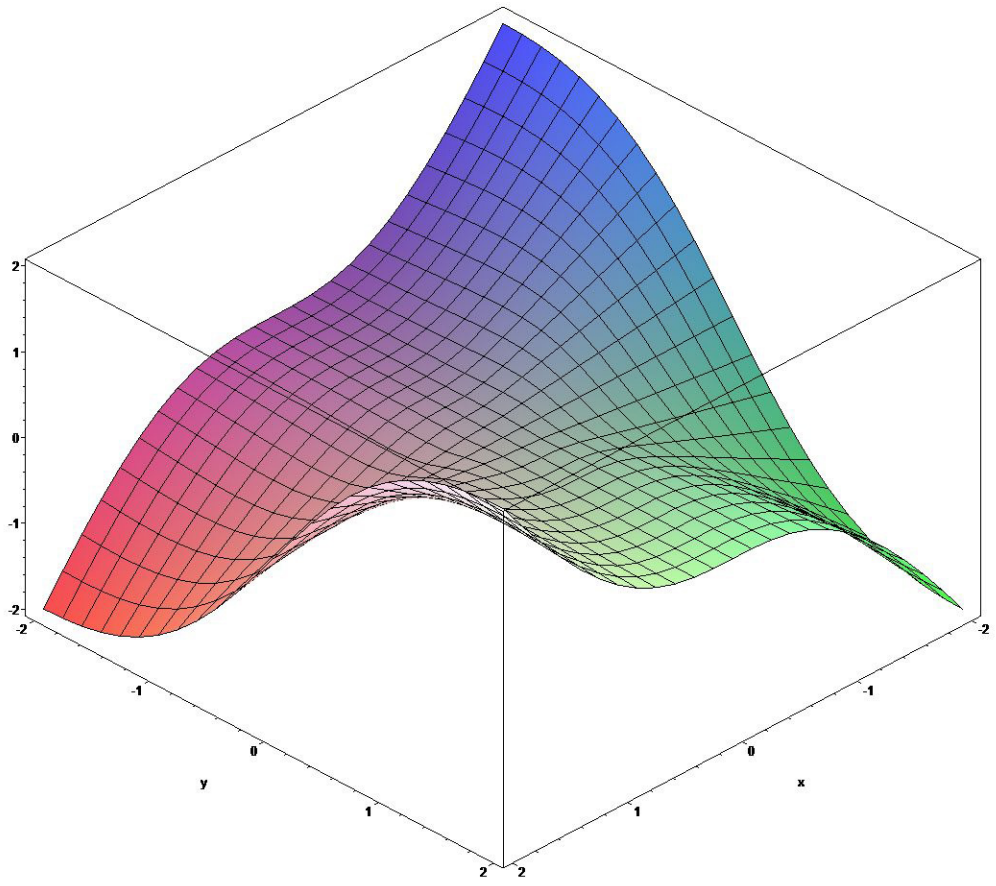


Figure XIII.27: Maple plot

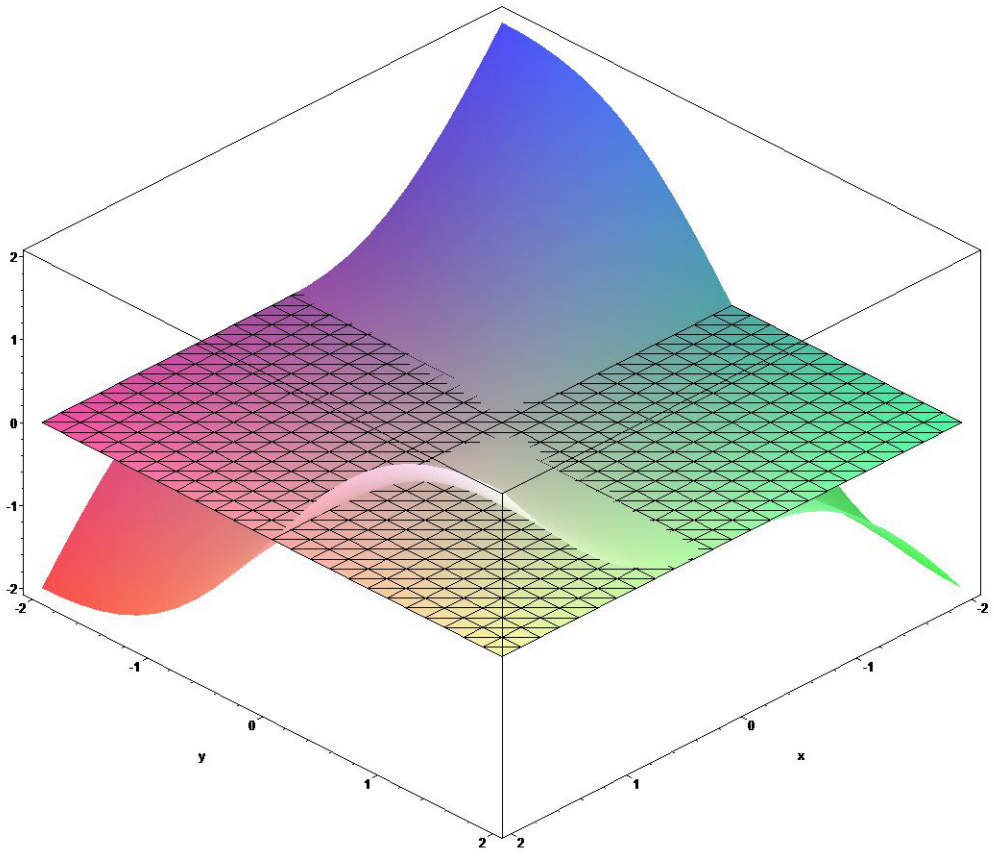


Figure XIII.28: Maple plot

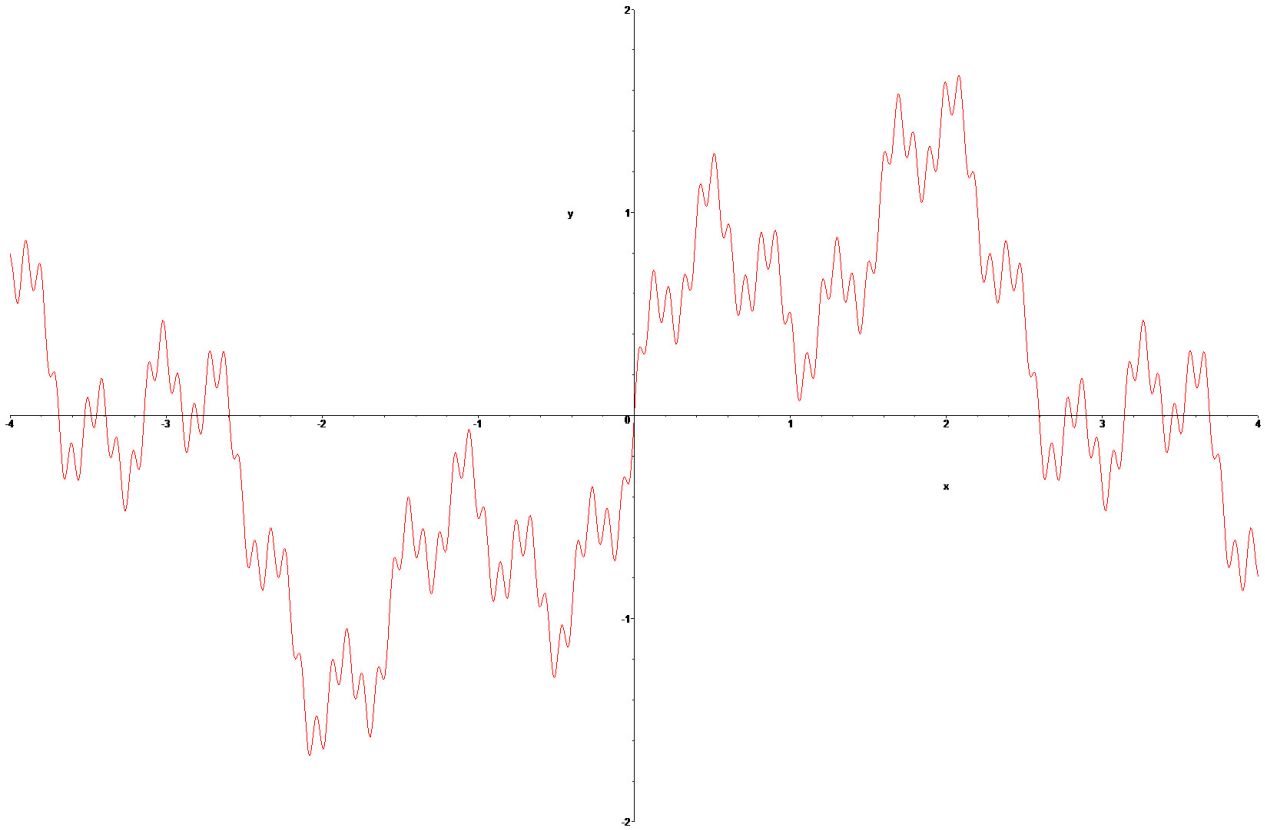


Figure XIII.29: Maple plot

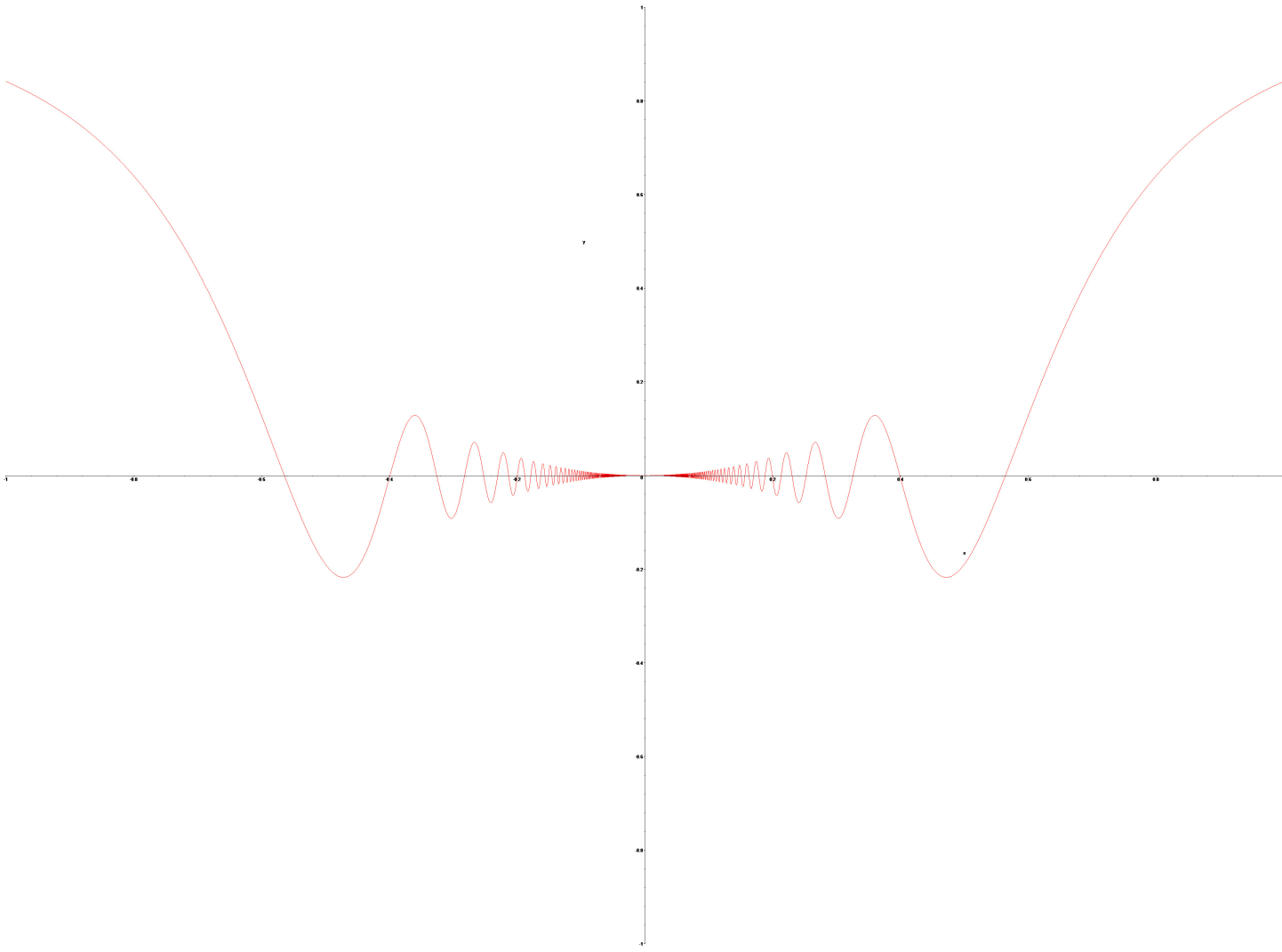


Figure XIII.30: Maple plot of graph of $f(x) = x^2 \sin\left(\frac{1}{x^2}\right)$

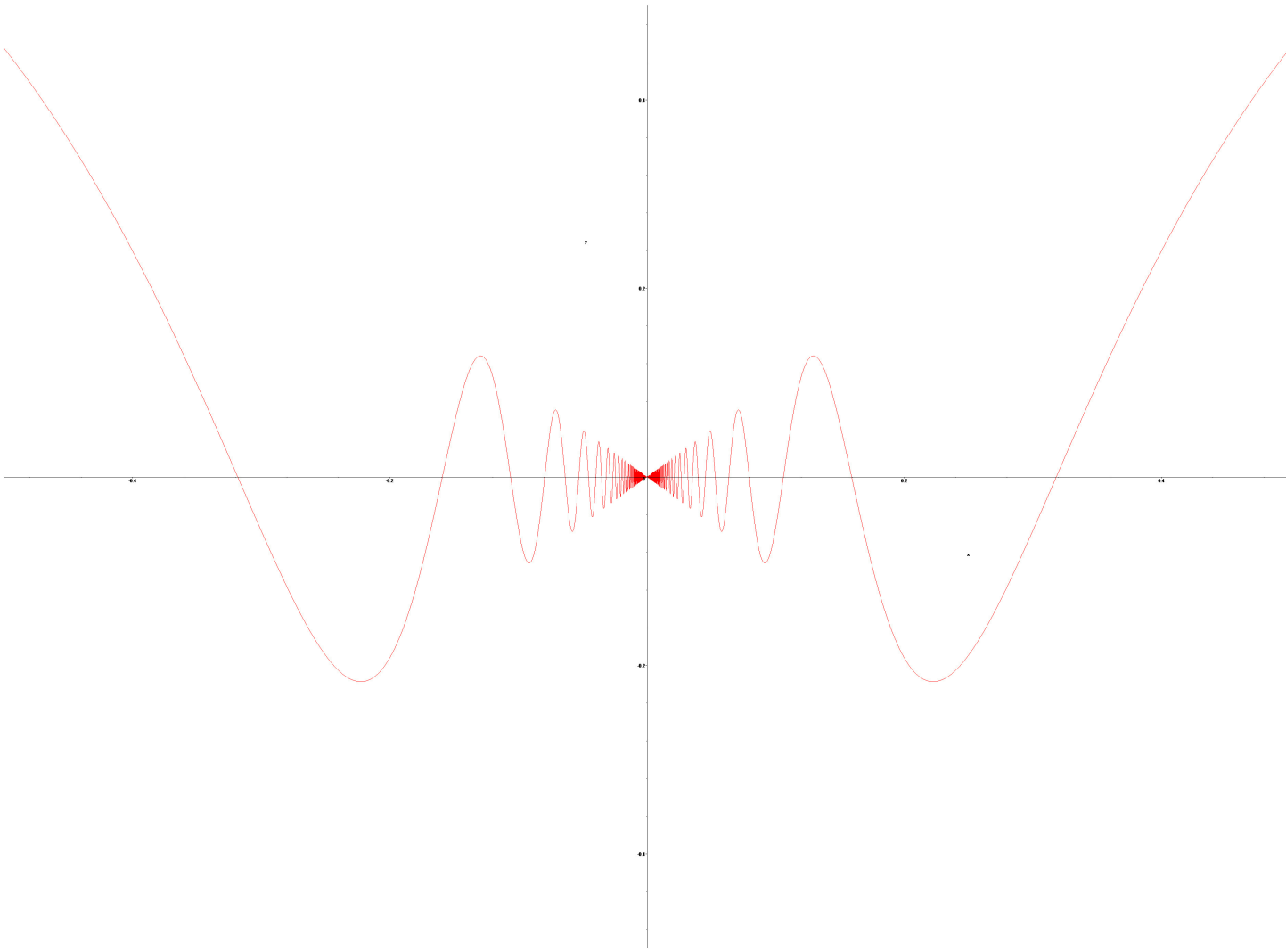


Figure XIII.31: Maple plot of graph of $f(x) = x \sin\left(\frac{1}{x}\right)$

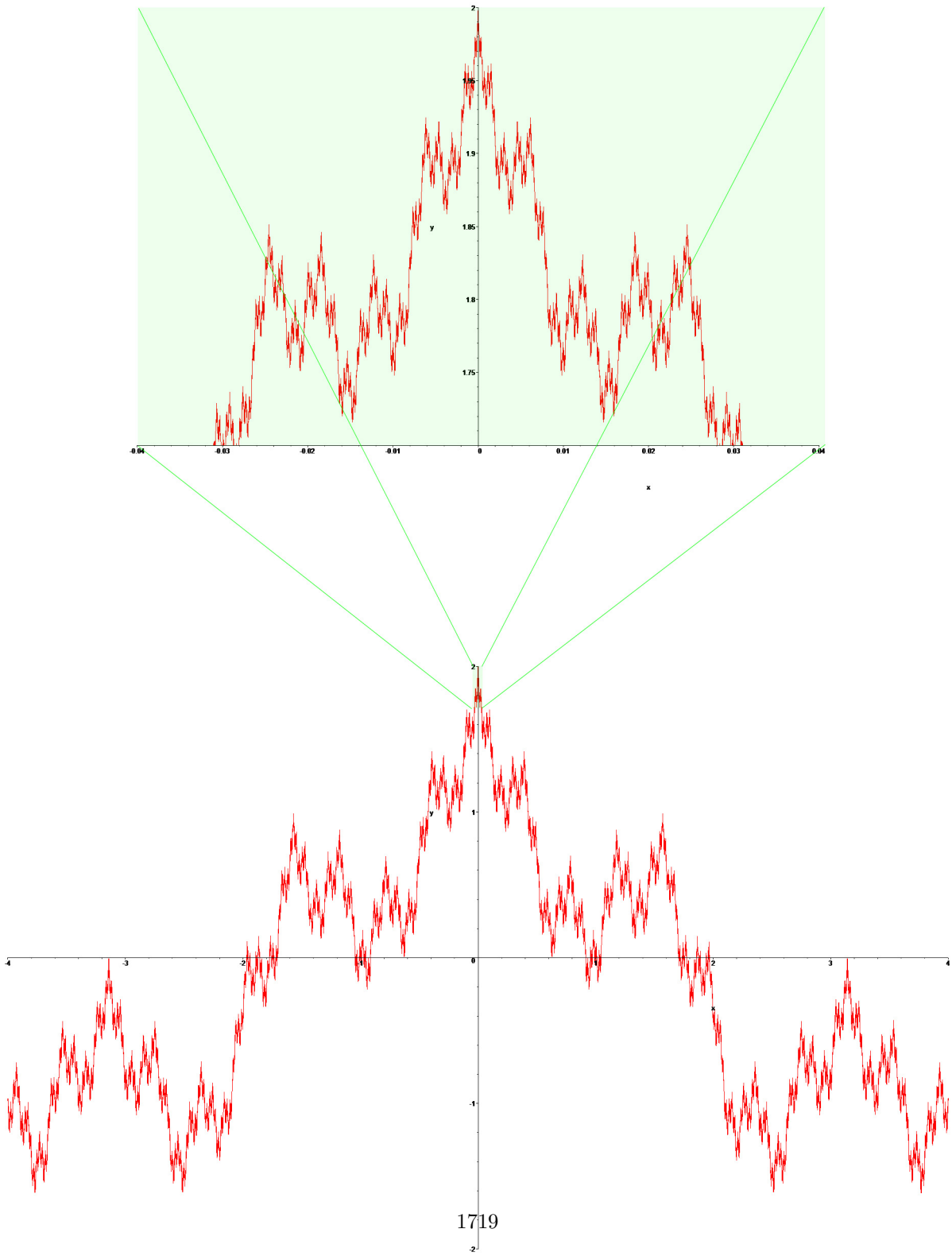


Figure XIII.32: Maple plot of $f(x) = \sum_{k=1}^{\infty} 2^{-k} \cos(4^k x)$

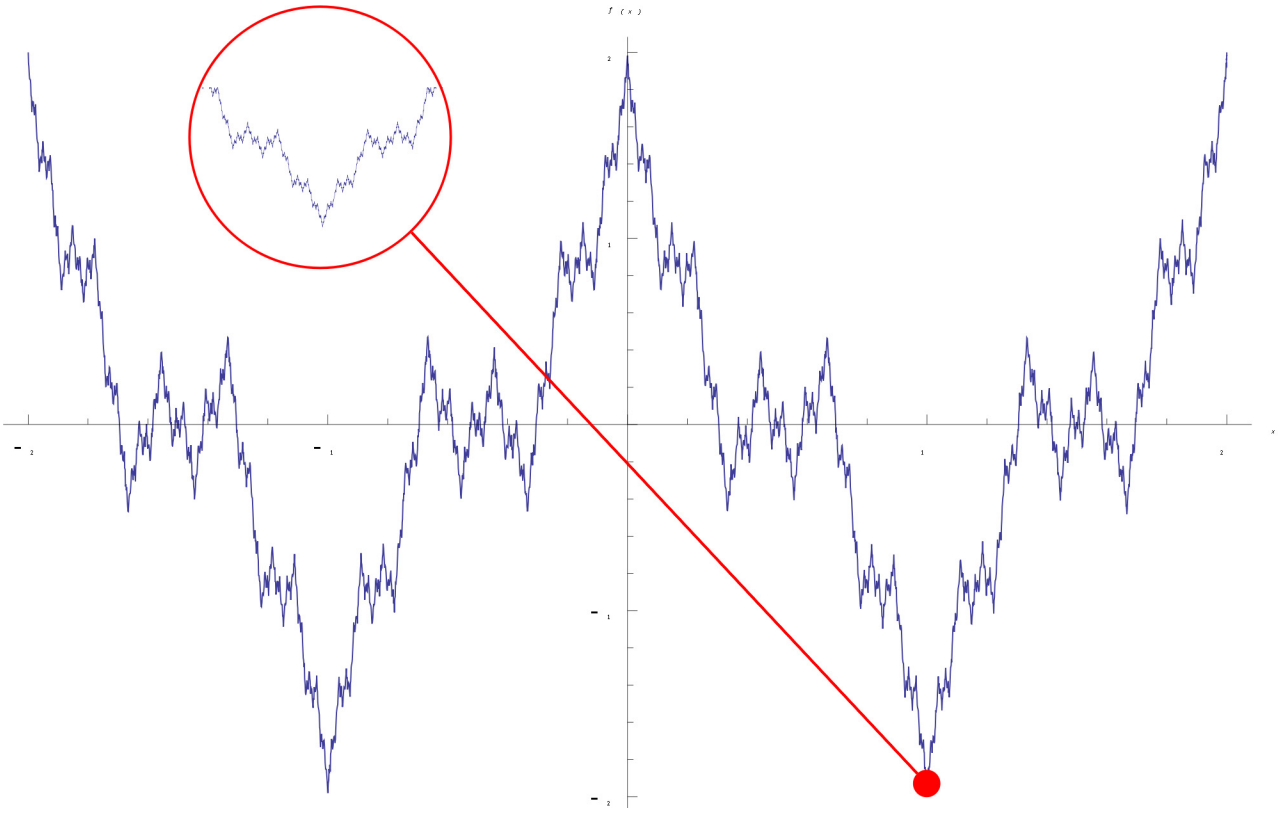


Figure XIII.33: Mathematica plot

XIII.19.4. When is $\mathbb{R}^n$ "Similar" to $\mathbb{R}^m$?

In this section, we ask a rather vague and subjective question: for which n and m are $\mathbb{R}^n$ and $\mathbb{R}^m$ "essentially similar"? The answer depends partly on what type of mathematician is answering it.

XIII.19.4.1. Of course, in a certain technical sense, $\mathbb{R}^n$ and $\mathbb{R}^m$ are different for any $n \neq m$ since they are not homeomorphic (or linearly isomorphic); thus they by definition differ in some describable topological (or algebraic) way. But qualitatively all Euclidean spaces, especially ones of large dimension, seem to be "similar." The question is really: how large does n have to be for Euclidean spaces of dimension $\geq n$ all to behave essentially the same way? Is there really such a dividing line at all?

XIII.19.4.2. In calculus or real analysis, the appropriate answer seems to be that all Euclidean spaces of dimension ≥ 2 are essentially similar, but essentially different from $\mathbb{R}^1$. Calculus on $\mathbb{R}^2$ is quite different in character from calculus on $\mathbb{R}^1$, but calculus on $\mathbb{R}^5$, or $\mathbb{R}^{100}$, is not essentially different from calculus on $\mathbb{R}^2$ except for the trivial difference that notation becomes more complicated in larger dimensions. This statement is not unrestrictedly true: there are some features and results of vector analysis which are special to $\mathbb{R}^2$ and $\mathbb{R}^3$, e.g. ones related to cross product, divergence, and curl (); but even these results are mostly just special cases or interpretations of the general Stokes' Theorem (), which holds in any dimension.

There are analytic properties unique to three- and four-dimensional space which arise in mathematical physics, e.g. stability of configurations under a gravitational force satisfying an inverse-square law. However, it is unclear whether we should regard these results as deep mathematical reasons why physical space must be three-dimensional, or space-time four-dimensional, or whether the three-dimensionality of the universe (as we currently perceive and understand it) and the compatibility of physical laws with three-dimensionality are just physical facts not entailed by mathematical considerations.

XIII.19.4.3. A functional analyst might well give an even less discriminating answer: that all finite-dimensional Hilbert spaces (in particular, all Euclidean spaces) are essentially similar and essentially trivial, and the first important distinction is between finite-dimensional and infinite-dimensional Hilbert spaces.

XIII.19.4.4. On the other hand, a topologist would likely say that the Euclidean spaces of dimension ≤ 5 are all essentially different, while those of dimension ≥ 5 are essentially similar. They would probably go on to add that $\mathbb{R}^4$ is "more different" than any of the others, for various reasons (). One can identify simple and important topological differences between $\mathbb{R}^2$ and $\mathbb{R}^3$, e.g. XIII.19.6.3., Classification of closed manifolds, Banach-Tarski, Poincaré Conjecture, and between $\mathbb{R}^3$ and $\mathbb{R}^n$ for $n \geq 5$ (). Topologists even consider "low-dimensional topology" (of dimension ≤ 4) to be a separate subdiscipline.

XIII.19.4.5. There is a qualitative difference between even- and odd-dimensional spaces, however; for example, even and odd spheres are qualitatively different. Only odd spheres have globally nonvanishing vector fields (), only even spheres have orientation-reversing homeomorphisms without fixed points, etc. The differences can be interpreted as reflecting the difference between the complex K-theories of even and odd spheres ().

There are many other differences in the structure of n-manifolds for different n: for example, every n-dimensional PL manifold has a smooth structure (in fact a unique one) if $n \leq 6$, but if $n \geq 7$ there are closed n-dimensional PL-manifolds which do not have a smooth structure. And smooth structures, if they exist, need not be unique. The n-sphere has a unique smooth structure for $n = 1, 2, 3, 5, 6, 12$ and 61, and

more than one for all other $n \leq 126$ except possibly $n = 4$ (open), and all other odd n []. The cases of even $n > 126$ are partially but not completely known.

It may be a stretch to regard the question as including the structure of manifolds of different dimensions, since the definition of a general manifold is rather subtle; but it is less of a stretch to include the structure of spheres, since the n-sphere is just the one-point compactification of $\mathbb{R}^n$, or the unit sphere in $\mathbb{R}^{n+1}$. One might also legitimately consider the topology of open subsets of $\mathbb{R}^n$ (cf. XIII.19.6.6.–XIII.19.6.8.).

XIII.19.4.6. The classification of (convex) regular polytopes is another difference. There are infinitely many in dimension 2, five in dimension 3, six in dimension 4, and three in dimension n for every $n \geq 5$. So besides the three generic ones which occur in every dimension ≥ 3 (the analogs of the tetrahedron, cube, and octahedron), "exotic" ones exist only in dimensions ≤ 4, of different character in each dimension.

XIII.19.4.7. There are more esoteric dimensional restrictions arising from topology. Two 1-spheres (circles) can be linked nontrivially in $\mathbb{R}^n$ if and only if $n = 3$. But if $m > 1$, the question of whether two disjoint (smoothly embedded) m-spheres can be nontrivially linked in $\mathbb{R}^n$ $(n > m)$ is much more subtle: for example, two 50-spheres can be nontrivially linked in $\mathbb{R}^n$ if and only if $52 \leq n \leq 95$ or $98 \leq n \leq 101$ [Rol76]. Smooth embeddings of single spheres (higher-dimensional knots) are also quite subtle.

XIII.19.4.8. An algebraist might say $\mathbb{R}^1$, $\mathbb{R}^2$, $\mathbb{R}^4$, and $\mathbb{R}^8$ are similar to each other and different from all other $\mathbb{R}^n$ because of III.6.3.2.. In fact, for arbitrarily large n, the algebraic structure of $\mathbb{R}^n$ is significantly impacted by the number-theoretic properties of n (`https://mathoverflow.net/questions/34186/does-the-truth-of-any-statement-of-real-matrix-algebra-stabilize-in-sufficiently`), and much of the algebraic structure can be interpreted geometrically.

XIII.19.4.9. In geometric measure theory, there is a significant difference between $\mathbb{R}^n$ for $n \leq 7$ and for $n \geq 8$: if $n \leq 7$, an area-minimizing rectifiable $(n-1)$-current in $\mathbb{R}^n$ is a smooth submanifold (), but this is not necessarily true if $n \geq 8$. See [Mor09, 10.7].

XIII.19.4.10. There is a potentially much higher dividing line from the Borsuk conjecture (or problem): Borsuk asked whether any bounded subset of $\mathbb{R}^n$ is the union of $n+1$ subsets of strictly smaller diameter. The statement is known to be true for $n \leq 3$ and false for $n \geq 64$ (). The smallest n for which the statement fails is in at least one combinatorial sense an important transition point.

XIII.19.4.11. There are also interesting and important questions concerning sphere packings. One tries to fill $\mathbb{R}^n$ with solid equal-size n-balls as densely as possible. The problem is even tricky to state precisely: we first seek the supremum of the density of all possible packings, and then to find a packing (or "all" packings) realizing the supremum (optimal packings; one should impose additional conditions on optimal packings such as some type of incompressibility, part of the difficulty in phrasing the problem carefully). In $\mathbb{R}^2$ a hexagonal lattice is easily shown to give the unique optimal packing. In $\mathbb{R}^3$ the cubic face centered lattice () is optimal (*Kepler's Conjecture*), although the proof is very difficult and was only recently completed [HF11]; it is not the unique optimal packing, although the other optimal packings are easily described variations. Optimal packings are known in dimensions 2, 3, 8, and 24 (and only these other than dimension 1), and are lattice packings in these dimensions. In dimensions 10, 11, and 13 there are irregular (but periodic) packings which are denser than any known lattice packing, and it may be that in these dimensions, and possibly in all sufficiently large dimensions, optimal packings are necessarily irregular and perhaps even aperiodic. It is

not obvious that there even are any optimal packings (i.e. ones realizing the supremum) in every dimension, but there are (). See [CS99] and [**?**] for details. Sphere packing problems in large numbers of dimensions have applications in number theory, coding, algebra, and mathematical physics as well as in geometry.

XIII.19.4.12. As to the general question of whether there exists a "dividing line," many (but certainly not all!) questions or phenomena stabilize for high dimensions and are difficult only due to "accidents of small dimension." From the right point of view, this philosophy applies even to such problems as the notorious one of computing higher homotopy groups $\pi_n(S^q)$ of spheres, which is not completely solved and is perhaps not even completely solvable in principle. Viewed in the usual way, by fixing q and letting n vary, the question can seem intractable. But if we instead fix n and let q vary, the problem stabilizes nicely for $q \geq n$ and the complications can be said to be only due to "accidents of small dimension."

Of course, as seen above, if dividing lines exist at all, they may vary from problem to problem.

XIII.19.4.13. Perhaps the best answer to the original question may be that it simply reflects the limitations of our current knowledge. Even topologists as recently as the 1960s might not have said that $\mathbb{R}^4$ is essentially different from $\mathbb{R}^3$; the most profound differences have all been discovered since then. Mathematicians 100 years from now may regard $\mathbb{R}^{18}$ and $\mathbb{R}^{19}$ (or maybe $\mathbb{R}^{17}$ and $\mathbb{R}^{19}$, to pick prime numbers) as being as different as $\mathbb{R}^3$ and $\mathbb{R}^4$ due to some phenomenon not yet imagined.

XIII.19.4.14. If you want to start some lively discussions, ask your colleagues or instructors this question! I decided to pose the question on MathOverflow (https://mathoverflow.net/questions/267909/when-are-mathbbrn-and-mathbbrm-essentially-similar) and the ensuing discussion has informed and improved my own discussion above.

XIII.19.5. Compact Subsets of $\mathbb{R}^n$

In this section, we describe a number of examples of compact subsets of Euclidean space with unusual, even bizarre, properties. We will state some of the properties of the examples without proof; readers with some topological background are encouraged to verify the asserted properties.

Mathematicians tend to have two types of reactions to examples of this sort. Some find them fascinating; others have more a sense of revulsion and want to sweep them into an obscure corner. I myself have the first reaction. But even if you fall into the second camp, it is still very important to be aware of the existence of such examples, since they graphically illustrate two vital mathematical principles:

(i) The number and variety of mathematical objects, even ones of relatively restricted type such as closed bounded subsets of the plane, is often vast, far beyond the limits of one's initial imagination.

(ii) When proving results about general mathematical objects of a specified type, it is necessary to stick strictly to arguments drawing logical conclusions from explicit axiomatic assumptions, without appeal to intuition or heuristic arguments, especially ones based on pictures or diagrams which may not accurately reflect the full range of possibilities for the objects studied.

Of course, various other examples, such as the examples of continuous and/or differentiable functions with pathological properties which are scattered through this book, also illustrate these principles.

It turns out that subsets similar to some of these occur naturally in mathematics, both pure and applied, such as strange attractors for nonlinear dynamical systems ().

Some of the examples of this section are fractals, and their fractal properties will be examined in ().

Uncountably Many Mutually Nonhomeomorphic Compact Subsets of $\mathbb{R}^2$

XIII.19.5.1. If $\boldsymbol{k} = (k_1, k_2, \dots)$ is a sequence of natural numbers with $k_1 \geq 2$, construct a compact subset $X_{\boldsymbol{k}}$ of $\mathbb{R}^2$ as follows. Begin with the closed segment from $(0,0)$ to $(1,0)$. For each n, add the closed line segments between $\left(\frac{1}{n}, 0\right)$ and $\left(\frac{1}{n+1}, \frac{1}{n(n+m)}\right)$ for $1 \leq m \leq k_n$.

For example, Figure XIII.34 shows $X_{\boldsymbol{k}}$ for $\boldsymbol{k} = (2, 3, 1, 1, 1, \dots)$.
It is clear that $X_{\boldsymbol{k}}$ is a closed bounded subset of $\mathbb{R}^2$. It can be shown with some straightforward elementary topological arguments ("meatball topology," cf. Exercise XIII.19.6.1.) that the $X_{\boldsymbol{k}}$ for distinct sequences $\boldsymbol{k}$ are mutually nonhomeomorphic. Since there are uncountably many $\left(2^{\aleph_0}\right)$ such $\boldsymbol{k}$, there are uncountably many $\left(2^{\aleph_0}\right)$ mutually nonhomeomorphic compact subsets of the plane. Since there are only $2^{\aleph_0}$ closed subsets of $\mathbb{R}^2$, there are exactly $2^{\aleph_0}$ homeomorphism classes of closed bounded subsets of the plane.

Each of the $X_{\boldsymbol{k}}$ is path-connected and topologically one-dimensional. They are even contractible; in fact, they are absolute retracts (dendrites) (). Note that these spaces are quite different from the infinite broom of XIII.13.9.9..

There is an enormous variety of variations possible on this construction.

XIII.19.5.2. There is a more spectacular version of this. Say that two topological spaces are *locally incomparable* if no nonempty open subset of either one can be topologically embedded in the other. Then:

XIII.19.5.3. THEOREM. ([**?**,] cf. [Bor67]) Let $n \in \mathbb{N}$. Then there exist $2^{\aleph_0}$ mutually locally incomparable n-dimensional (compact) absolute retracts in $\mathbb{R}^{n+1}$.

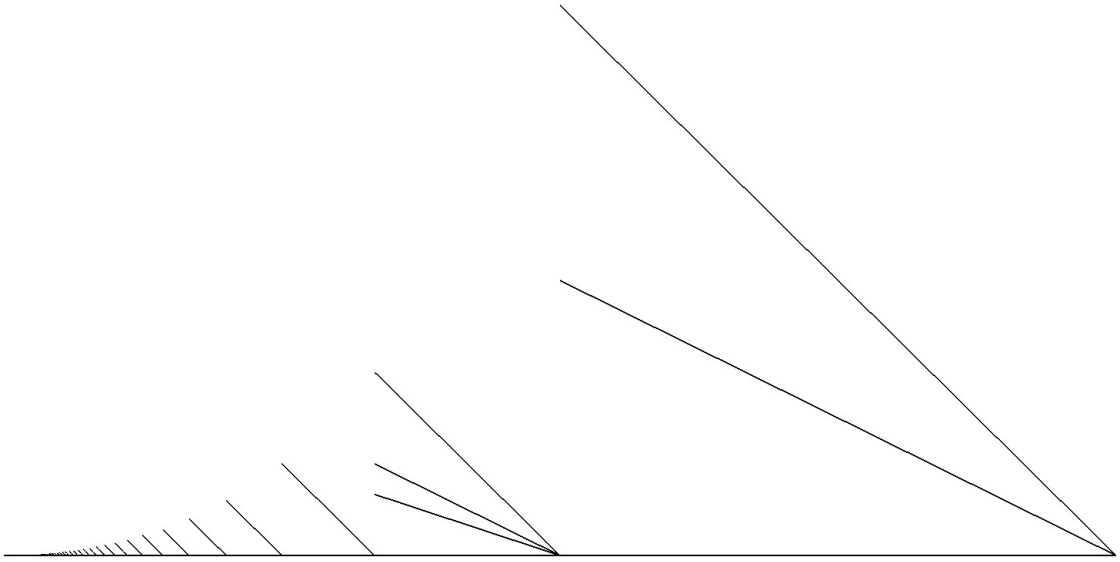


Figure XIII.34: $X_{\boldsymbol{k}}$ for $\boldsymbol{k} = (2, 3, 1, 1, 1, \dots)$.

Uncountably Many Mutually Nonhomeomorphic Compact Subsets of $\mathbb{R}$

XIII.19.5.4. Any countable ordinal () can be order-embedded in $\mathbb{R}$ (). Any countable ordinal which is not a limit ordinal can be embedded as a *compact* subset of $\mathbb{R}$ (with the order topology). There are $\aleph_1$ such ordinals.

However, these are not all distinct topologically. But the ordinals $\omega^\sigma + 1$ are all distinct topologically: the derived series () as a topological space has length σ. This space is countable if σ is countable. Thus there are at least $\aleph_1$ topologically distinct compact countable subsets of $\mathbb{R}$.

It is trickier to explicitly construct $2^{\aleph_0}$ topologically distinct compact subsets of $\mathbb{R}$. Here is a way (cf. [Rei62]). Let K be the Cantor set, and $c_n = 1 - \frac{1}{3^n}$. Then $c_n \in K$, and is the right endpoint of the rightmost open interval I_n removed at the n'th step of the construction of K. Let $\boldsymbol{m} = (m_1, m_2, \dots)$ be a strictly increasing sequence in $\mathbb{N}$, and for each n attach to K a copy of ω^{m_n} in I_n which is relatively closed and converges to c_n (see Figure ()). Let $X_{\boldsymbol{m}}$ be the resulting subset of $\mathbb{R}$. Then $X_{\boldsymbol{m}}$ is a compact totally disconnected subset of $\mathbb{R}$ for each $\boldsymbol{m}$, and the $X_{\boldsymbol{m}}$ are topologically distinct for distinct $\boldsymbol{m}$. There are $2^{\aleph_0}$ such $\boldsymbol{m}$, since they are in one-one correspondence with the infinite subsets of $\mathbb{N}$. See Exercise XIII.19.6.2.. Since there are only $2^{\aleph_0}$ compact subsets of $\mathbb{R}$, there are exactly $2^{\aleph_0}$ homeomorphism classes.

The Topologist's Sine Curve

XIII.19.5.5. Let X be the graph of $y = \sin\left(\frac{\pi}{x}\right)$ for $0 < x \leq 1$ (the right endpoint is not important), along with the closed segment on the y-axis between $(0, -1)$ and $(0, 1)$. Then X is the closure of the graph part, and is a compact subset of $\mathbb{R}^2$. See Figure XIII.35.

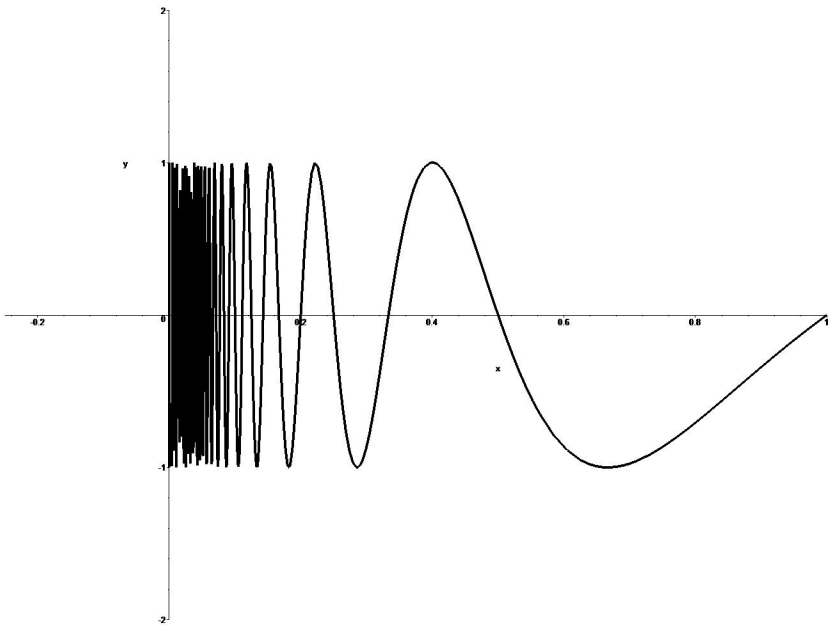


Figure XIII.35: The Topologist's Sine Curve

Since the graph is connected (even path-connected), X is the closure of a connected set and hence is connected (XIII.13.1.9.). However, X is not path-connected; it has two path components, the graph and the segment on the y-axis.

XIII.19.5.6. A variation is to "close up" X by adding an arc connecting $(1, 0)$ with $(0, -1)$. This is often called the "extended topologist's sine curve" or "Warsaw circle." See Figure XIII.36. This space is path-connected but not locally path-connected. In some references, the added arc goes from $(1, 0)$ around to $(0, 0)$ from the left; this space is topologically different but essentially similar.

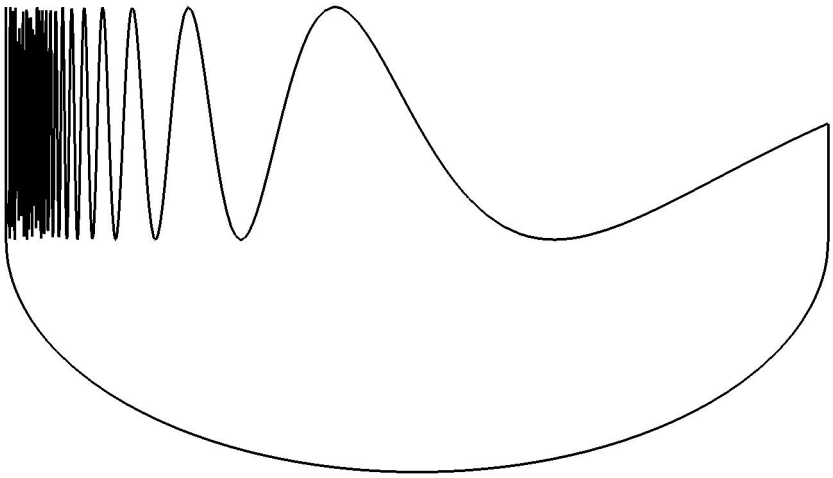


Figure XIII.36: The Extended Topologist's Sine Curve or Warsaw Circle

XIII.19.5.7. The topologist's sine curve can be used to construct a "counterexample" to the Four-Color Theorem (Figure XIII.37). Each of the five regions pictured has a boundary with each of the others which at least contains an entire arc, and thus they must all be colored differently. This shows that the Four-Color Theorem must be stated carefully to rule out examples of this sort (cf. [**?**]).

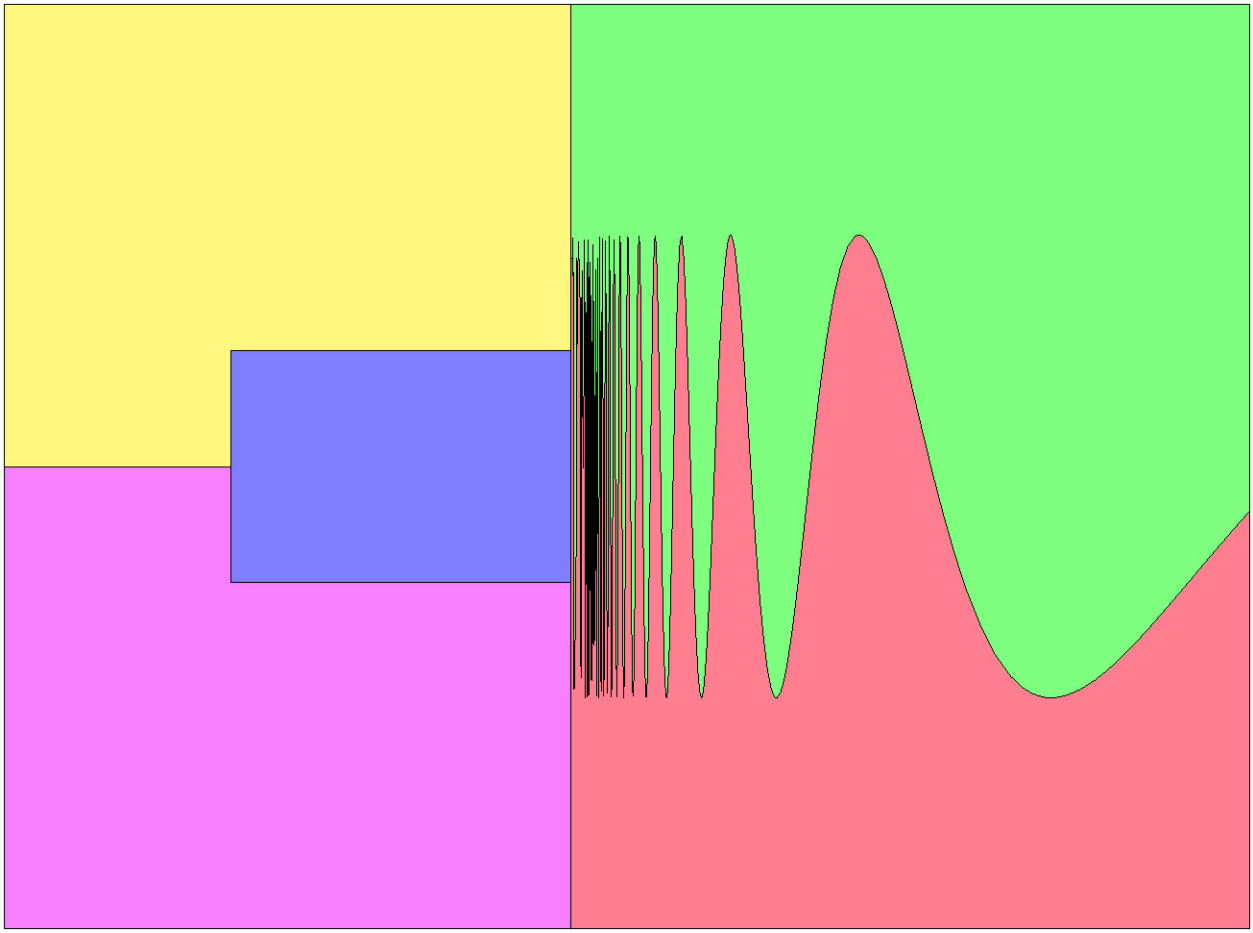


Figure XIII.37: A counterexample to the Four-Color Theorem?

The Hawaiian Earring

XIII.19.5.8. Let X be the union of all circles in $\mathbb{R}^2$ with center $\left(\frac{1}{n}, 0\right)$ and radius $\frac{1}{n}$, for $n \in \mathbb{N}$ (Figure XIII.38).

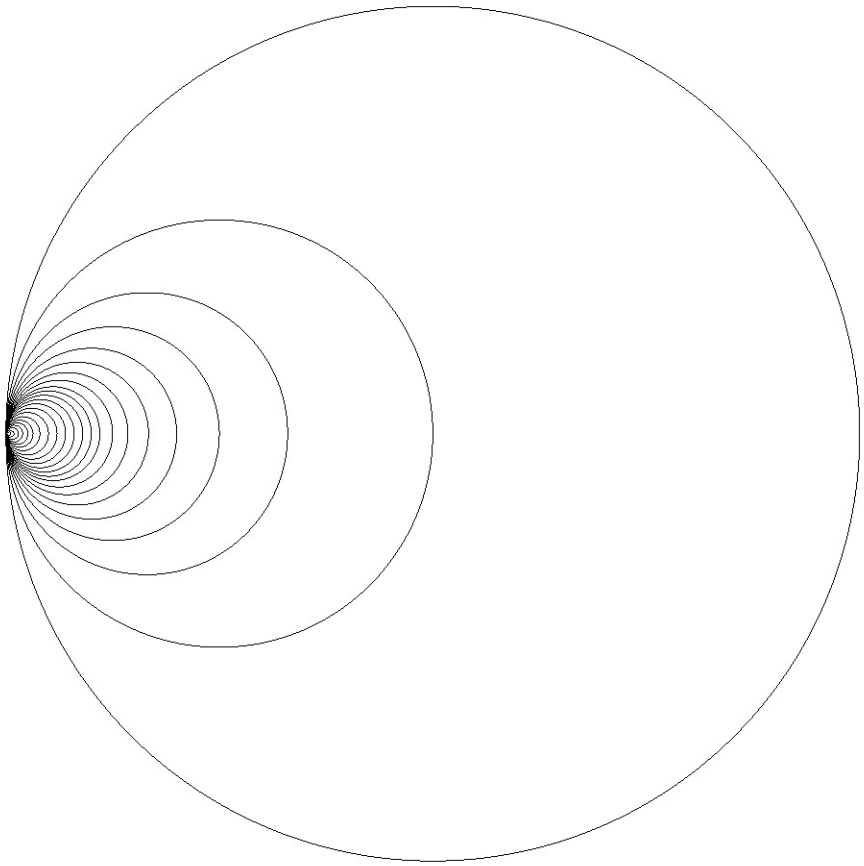


Figure XIII.38: The Hawaiian Earring

The fundamental group of the Hawaiian earring is uncountable: for each subset S of $\mathbb{N}$, there is a loop γ_S which goes around the circle of radius $1/n$ if and only if $n \in S$, and no two of these loops are homotopic.

XIII.19.5.9. Let Y be the union of all circles in $\mathbb{R}^2$ with center $(n, 0)$ and radius n, for $n \in \mathbb{N}$. Then Y is not homeomorphic to X, since it is not compact. The fundamental group of Y is a free group on a countable number of generators. In fact, Y is homotopy equivalent (but not homeomorphic) to an infinite wedge of circles (XIII.9.1.13., XIII.9.5.8.) [collapse a suitable closed neighborhood of $\mathbf{0}$ to a point]. See also XIII.9.5.5..

Subsets Constructed as Intersections

A number of pathological examples can be made by constructing a decreasing sequence of increasingly complicated compact subsets of Euclidean space and taking the intersection. This construction is already familiar in the definition of the Cantor set ().

The Sierpiński Gasket

XIII.19.5.10. This subset of $\mathbb{R}^2$ is constructed as an intersection, and is the first of several straightforward higher-dimensional versions of the Cantor set.

Begin with a (solid) triangle in $\mathbb{R}^2$, say an equilateral triangle. Bisecting each side and connecting the midpoints forms four subtriangles. Remove the (open) middle one. Then repeat the process in each of the three remaining triangles and iterate. The intersection is called the *Sierpiński gasket* (the name is due to B. Mandelbrot; the set is also known as the *Sierpiński triangle* or the *Sierpiński sieve*). It is path-connected, compact, and topologically one-dimensional. See Figure XIII.39 (from http://en.wikipedia.org/wiki/File:Sierpinski_Triangle.svg).

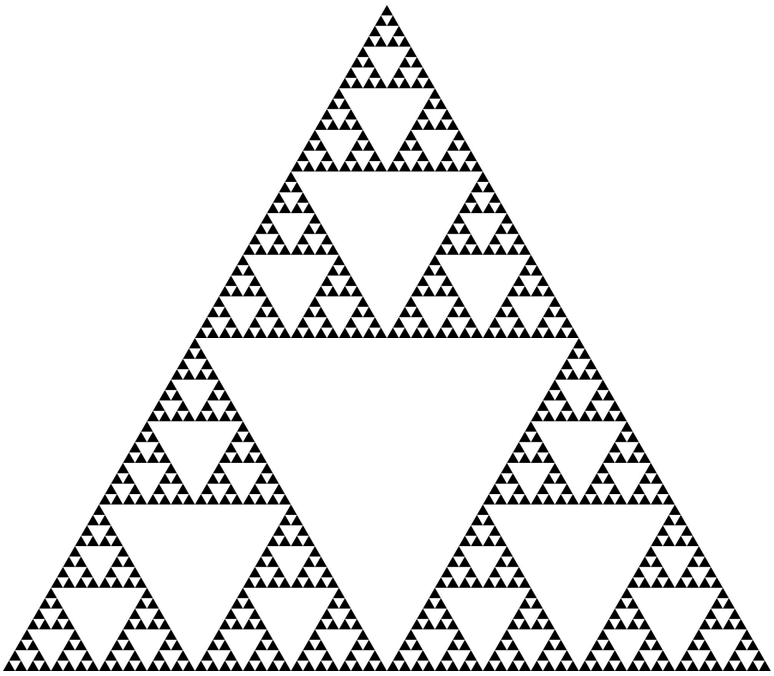


Figure XIII.39: Six steps in the construction of the Sierpiński Gasket

The Sierpiński Carpet

XIII.19.5.11. As a variation, begin with a (solid) square. Divide it into 9 equal-sized squares, and remove the open middle square. Repeat the process in each of the eight remaining squares, and iterate. See

Figure XIII.40 (from http://commons.wikimedia.org/wiki/File:Sierpinski6.png?uselang=fr). This set is called the *Sierpiński carpet*, or *Sierpiński curve*.

Like the Sierpiński gasket, the Sierpiński carpet is topologically one-dimensional.

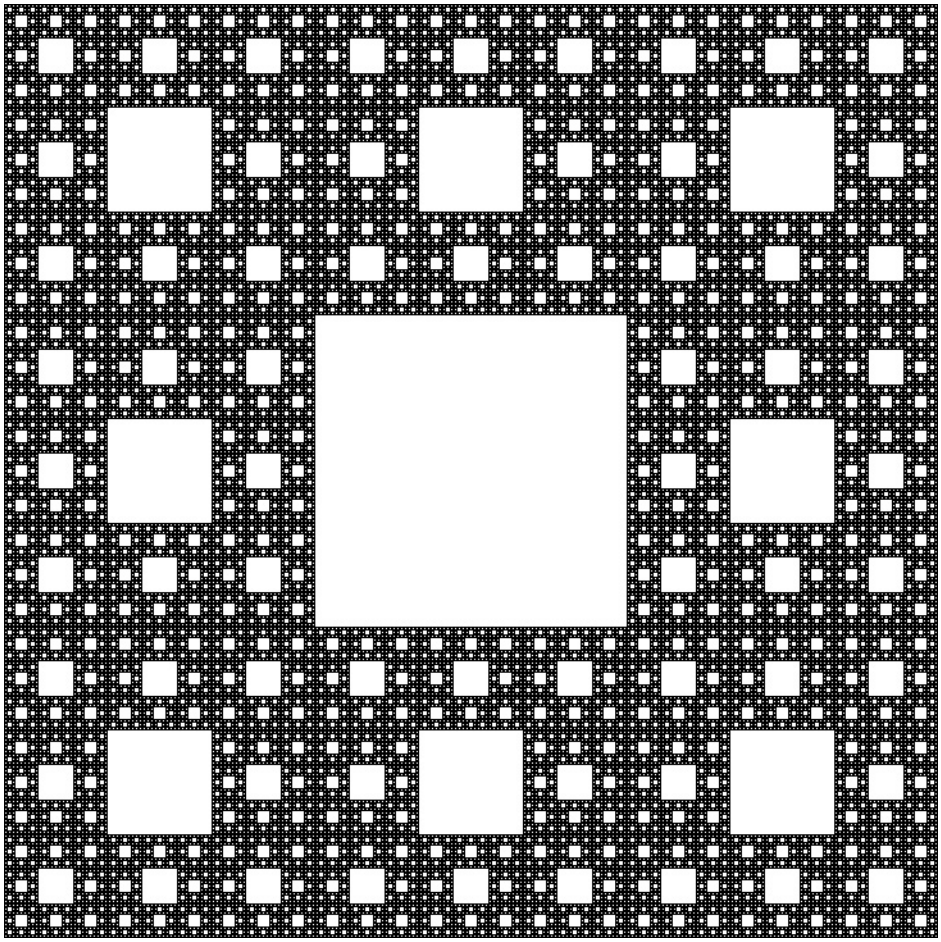


Figure XIII.40: Six steps in the construction of the Sierpiński Carpet

XIII.19.5.12. It is not obvious whether the Sierpiński carpet is homeomorphic to the Sierpiński gasket. In fact, they are not homeomorphic. For each point of the Sierpiński gasket has arbitrarily small open neighborhoods whose boundary consists of finitely many points (no more than 4), while small open neighborhoods of points in the Sierpiński carpet always have uncountably many boundary points. See [Kur68, §51] for a discussion.

XIII.19.5.13. The Sierpiński carpet is universal for one-dimensional compact subsets of the plane: every compact subset of the plane which is topologically one-dimensional is homeomorphic to a subset of

the Sierpiński carpet []. In particular, the Sierpiński gasket is homeomorphic to a subset of the Sierpiński carpet (try to find one!) A pseudo-arc can also be embedded topologically in the Sierpiński carpet.

Swiss Cheese

XIII.19.5.14. The Sierpiński carpet is (topologically) a special case of a general construction: begin with a closed disk in $\mathbb{R}^2$, and remove a sequence of disjoint open disks to obtain a set with empty interior. Such a subset of $\mathbb{R}^2$ is called a *Swiss cheese*. There is a great topological variety among Swiss cheeses; some are similar topologically to the Sierpiński gasket, and some are homeomorphic to the Sierpiński carpet. The one pictured in XIII.41 (from `http://hans.math.upenn.edu/~pstorm/images/round_Sierpinski_carpet.png`) is made in a regular manner and is a fractal; some are much more irregular. As with Cantor sets, Swiss cheeses of positive planar Lebesgue measure can be constructed, as well as ones of zero planar measure; in fact, Swiss cheeses of Hausdorff dimension between 1 and 2 can be constructed, although the exact Hausdorff dimensions are subtle to compute ([Mel66], [Lar66]).

If topological disks are used instead of geometric ones, an even greater variety can be obtained; both the Sierpiński carpet and the Sierpiński gasket are special cases of this construction, as are the Lakes of Wada.

Swiss cheeses are used extensively in the theory of function algebras (); see [FH10] for a good survey, including their topological structure.

One can make higher-dimensional Swiss Cheeses by removing a sequence of open balls from the closed unit ball in $\mathbb{R}^n$.

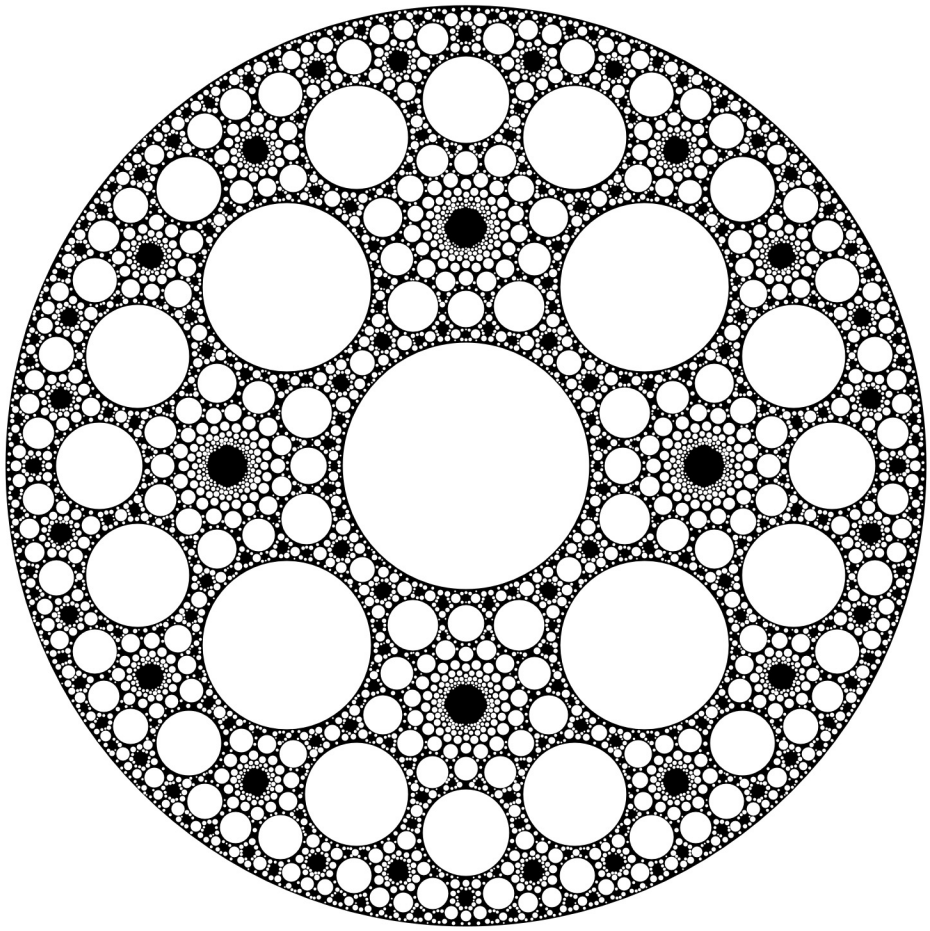


Figure XIII.41: A "Swiss cheese." (Small white disks continue throughout the black spots.)

The Menger Sponge

XIII.19.5.15. There is a three-dimensional analog of the Sierpiński carpet, due to K. Menger [Men26]. Begin with the unit cube in $\mathbb{R}^3$. Dividing each edge in three equal segments leads to a decomposition of the cube into 27 equal-sized subcubes. Remove the central subcube and the central subcubes on each face (i.e. keeping all closed subcubes intersecting an edge of the cube). Seven subcubes are removed in all. Do the same in each of the 20 remaining subcubes and iterate. The intersection is called the *Menger sponge*. Each outside face of the Menger sponge is a copy of the Sierpiński carpet. See Figure XIII.42 (from http://en.wikipedia.org/wiki/File:Menger-Schwamm-farbig.png).

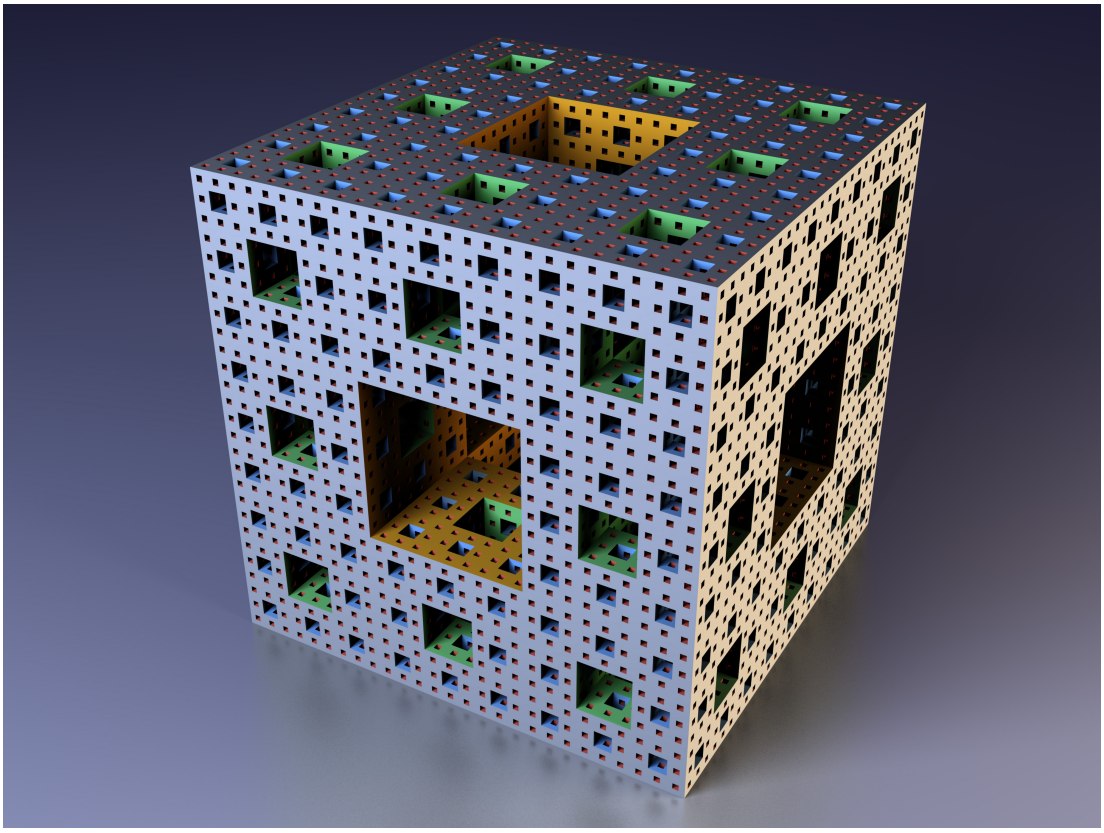


Figure XIII.42: Four steps in the construction of the Menger Sponge

The Menger sponge is compact, connected, locally connected, and topologically one-dimensional. It is universal for one-dimensional separable metrizable spaces: every separable metrizable space of dimension ≤ 1 is homeomorphic to a subset of the Menger sponge [Men26]. Thus the Menger sponge is sometimes called the *Menger Universal Curve* (although it is not a "curve" in any normal sense). It is homogeneous and has a topological characterization [And58].

For any $k, n \in \mathbb{N}_0$, $k < n \leq 2k+1$, there is a similar subset M_k^n of $\mathbb{R}^n$ which is universal for compact subsets of $\mathbb{R}^n$ of dimension $\leq k$ [Šta71]; M_k^{2k+1} is universal for all separable metrizable spaces of dimension

$\leq k$ [Bot63]. M_1^3 is the Menger sponge, M_1^2 the Sierpiński carpet, and M_0^1 the Cantor set. To make M_2^3, begin with the unit cube in $\mathbb{R}^3$, subdivide into 27 subcubes, and just remove the center cube; then iterate.

Sierpiński Pyramids

At least two versions of the Sierpiński Gasket can be made in $\mathbb{R}^3$. Begin with a pyramid with either a triangular or square base. In the triangular case (say, a regular tetrahedron), bisect each edge and keep the four subpyramids whose vertices are one original vertex and the three midpoints of the adjacent edges; then iterate. The intersection is a compact set called the *tetrahedral Sierpiński Pyramid* which is topologically one-dimensional. It has the unusual property that its Hausdorff dimension is an integer (2) which is greater than its topological dimension (see XVII.6.9.3. for another example).

A variation is to begin with a pyramid with square base, bisect the edges, and keep the five subpyramids consisting of the top pyramid with four vertices at the midpoints of the slant edges and the four pyramids with top vertex at one of these midpoints and bottom vertices at a corresponding original vertex, two midpoints of base edges, and the midpoint of the base. See Figure XIII.43, from http://en.wikipedia.org/wiki/File:Sierpinski_pyramid.png.

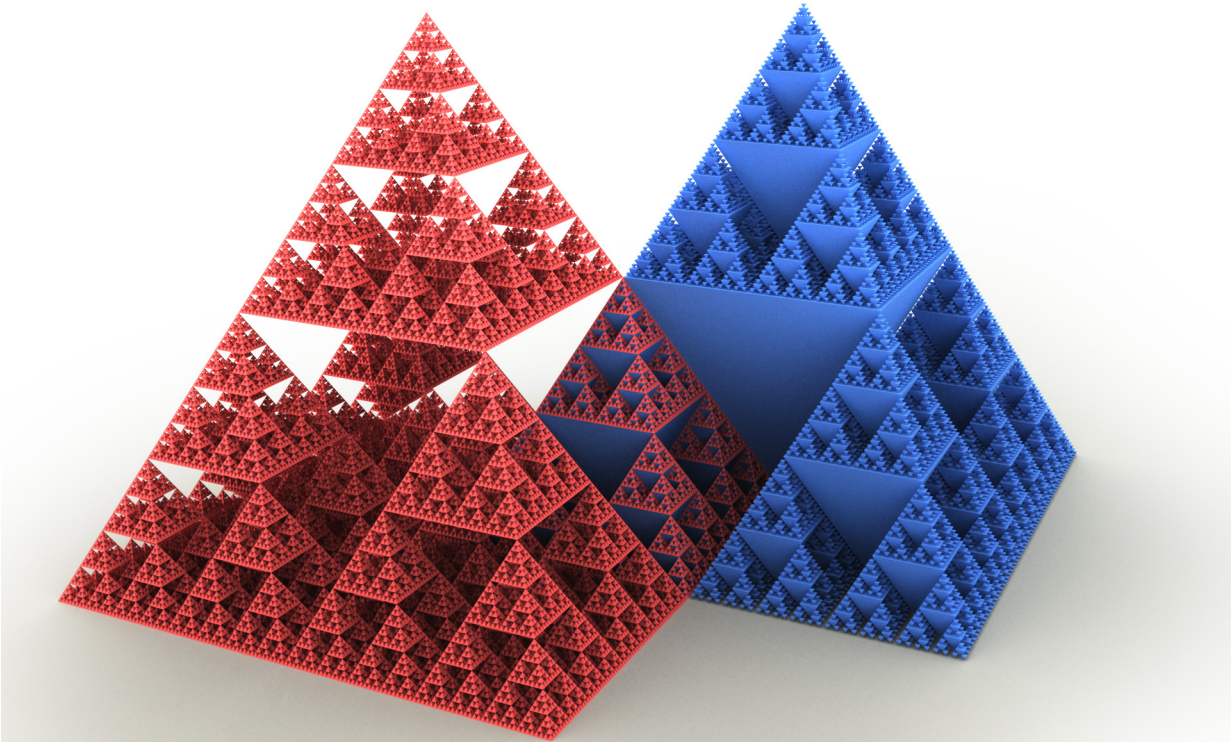


Figure XIII.43: Seven steps in the construction of the Sierpiński Pyramid with square base and its "inverse"

Solenoids

XIII.19.5.16. Let $s = (n_1, n_2, \dots)$ be a sequence of natural numbers with $n_k \geq 2$ for all k. Construct a compact subset of $\mathbb{R}^3$ as follows (in understanding the construction, try taking all n_k to be 2 at first). Let X_0 be a solid torus. Inside X_0 take a thinner solid torus X_1 which "winds around" n_1 times. (Think of wrapping a wire around the "hole" in X_0 n_1 times and connecting the ends of the wire to make a thin solid torus.) See XIII.44 for the case $n_1 = 2$ (from http://en.wikipedia.org/wiki/File:Smale-Williams_Solenoid.png). Now within X_1 take an even thinner solid torus X_2 which wraps around X_1 n_2 times, i.e. every cross-

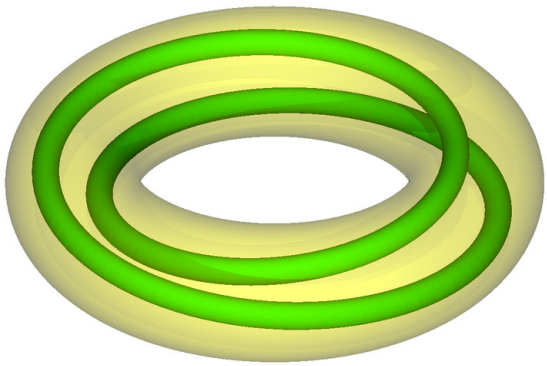


Figure XIII.44: First step in the construction of the solenoid of type $(2, 2, \dots)$

section disk of X_1 will cut X_2 in n_2 cross-sectional disks. Repeat the process countably many times and let $X = \cap_k X_k$. X is called the *solenoid* of type s.

A nice animated construction of the solenoid of type $(2, 2, \dots)$ can be found at http://en.wikipedia.org/wiki/File:Solenoid.gif .

XIII.19.5.17. Solenoids can also be naturally constructed in several other ways. One good way is as inverse limits of circles (XIII.14.4.17.).

XIII.19.5.18. A solenoid is connected but not locally connected: locally, a solenoid looks like $\mathbb{R} \times K$, where K is the Cantor set. In fact, each solenoid can be constructed (up to homeomorphism) from $[0, 1] \times K$ by identifying $\{0\} \times K$ with $\{1\} \times K$ in an appropriate way, e.g. by identifying K with the p-adic integers () and identifying $(1, x)$ with $(0, x + 1)$ (this gives the solenoid of type $(p, p, p, \dots)$). More generally, this can be done using any minimal homeomorphism of K ("flow under a constant function," cf. XIII.9.2.3.), giving additional solenoid-type examples (cf. [**?**]).

XIII.19.5.19. A solenoid is not path-connected. There are uncountably many path components, each of which is dense and a one-to-one continuous image of $\mathbb{R}$. This is seen most easily by viewing the solenoid as an inverse limit as in (), but can also be deduced from the description in XIII.19.5.18. (in this picture, the path components correspond to the cosets of $\mathbb{Z}$ in the p-adic integers).

XIII.19.5.20. Up to homeomorphism, a solenoid depends only on the sequence s. Solenoids corresponding to different sequences are "usually" nonhomeomorphic, but ones corresponding to sequences differing only "slightly" can be homeomorphic.

To give a precise description of the homeomorphism classes, we use the supernatural numbers:

XIII.19.5.21. DEFINITION. A *supernatural number* is an expression of the form

$$\boldsymbol{n} = 2^{e_2}3^{e_3}5^{e_5}\cdots$$

where each prime appears and each exponent e_p is in $\{0\} \cup \mathbb{N} \cup \{\infty\}$.

An ordinary natural number is a supernatural number in which all e_p are finite and all but finitely many are 0. Sometimes we want to exclude the ordinary natural numbers from the supernatural numbers; we call supernatural numbers which are not ordinary natural numbers *infinite*. When writing a supernatural number, we usually leave out all primes with exponent 0; thus if $\boldsymbol{n}$ is the supernatural number with $e_2 = \infty$ and $e_p = 0$ for $p \neq 2$, we write $\boldsymbol{n}$ as 2^∞.

Two supernatural numbers can be multiplied by adding the corresponding exponents. This multiplication extends the ordinary multiplication on $\mathbb{N}$.

We say one supernatural number $\boldsymbol{n}$ *divides* another supernatural number $\boldsymbol{m}$ if $e_p(\boldsymbol{n}) \leq e_p(\boldsymbol{m})$ for all p, i.e. if there is a supernatural number $\boldsymbol{r}$ such that $\boldsymbol{m} = \boldsymbol{nr}$. (Note, however, that $\boldsymbol{r}$ is not necessarily unique, i.e. that there is not a well-defined quotient in general, if the exponents for some p are both infinite.) There is a "largest" supernatural number, the one for which $e_p = \infty$ for all p, and all supernatural numbers divide this one.

Note that the supernatural numbers are not the same thing as the nonstandard natural numbers ${}^*\mathbb{N}$ of ().

XIII.19.5.22. The supernatural numbers are in natural one-one correspondence with the additive subgroups () of $\mathbb{Q}$ containing $\mathbb{Z}$: the group $\mathbb{Z}_{(\boldsymbol{n})}$ is the set of all rational numbers whose denominator divides $\boldsymbol{n}$. For example, $\mathbb{Z}_{(2^\infty)}$ is the set of dyadic rationals (). The group corresponding to the largest supernatural number is $\mathbb{Q}$ itself; $\mathbb{Z}$ corresponds to the supernatural number 1 (i.e. $e_p = 0$ for all p), the smallest supernatural number.

The group $\mathbb{Z}_{(\boldsymbol{n})}$ is cyclic (has a smallest positive element) if and only if $\boldsymbol{n}$ is an ordinary natural number.

XIII.19.5.23. To each sequence (n_k) of natural numbers, there is a naturally associated supernatural number $\prod_k n_k$, for which, for every p, e_p is the sum of the powers of p occurring in the n_k's. This will be an ordinary integer if and only if all but finitely many n_k are 1. Many different sequences can give the same associated supernatural number (for example, reordering the sequence does not change the associated supernatural number, and there are much more drastic changes that work too).

It can be shown without too much difficulty (knowing the first statement of XIII.19.5.20.) that if two sequences give the same supernatural number, the corresponding solenoids are homeomorphic. But it is not quite true that solenoids corresponding to different supernatural numbers are nonhomeomorphic. For a correct statement, we need an equivalence relation on supernatural numbers:

XIII.19.5.24. DEFINITION. Let $\boldsymbol{n}$ and $\boldsymbol{m}$ be supernatural numbers. Then $\boldsymbol{n}$ and $\boldsymbol{m}$ are *equivalent* if there are ordinary natural numbers n and m such that $n\boldsymbol{n} = m\boldsymbol{m}$.

Supernatural numbers $\boldsymbol{n}$ and $\boldsymbol{m}$ are equivalent if and only if only finitely many of their exponents differ, and each by only a finite amount. The supernatural numbers equivalent to any ordinary natural number are precisely all ordinary numbers.

This equivalence is natural because of the following results, the first of which is quite easy:

XIII.19.5.25. Proposition. Two additive subgroups of $\mathbb{Q}$ containing $\mathbb{Z}$ are isomorphic as groups if and only if their associated supernatural numbers are equivalent.

XIII.19.5.26. Theorem. Two solenoids are homeomorphic if and only if their associated supernatural numbers are equivalent.

XIII.19.5.27. Thus the equivalence classes of infinite supernatural numbers exactly parametrize the homeomorphism classes of solenoids. The equivalence class of supernatural numbers can be recovered directly from the intrinsic topology of the solenoid X: the first cohomotopy group $\pi^1(X)$ () is isomorphic to the additive subgroup of $\mathbb{Q}$ associated to the equivalence class of supernatural numbers of X.

XIII.19.5.28. There is another important connection. Any solenoid X can be made into a compact abelian topological group. Its Pontrjagin dual group () is isomorphic to $\pi^1(X)$. The solenoids are the only compact connected one-dimensional Hausdorff spaces besides the circle which can be made into abelian topological groups. See XXIII.5.2.1.(b).

XIII.19.5.29. Solenoids are homogeneous. It was shown in [Hag77] that every homogeneous continuum in which every nondegenerate subcontinuum is an arc is a solenoid (or circle). A solenoid is also indecomposable.

XIII.19.5.30. Solenoids arise naturally as attractors of important nonlinear dynamical systems ().

Knaster's Bucket-Handle

This is a "poor man's solenoid," somewhat easier to visualize and with similar topological properties, at least locally. It is perhaps the simplest example of an indecomposable continuum. It can be constructed as an intersection (see Figure XIII.46), but has a simpler direct definition:

XIII.19.5.31. Definition. The *Knaster bucket-handle* is the union of the following closed half-circles in $\mathbb{R}^2$:

All half-circles in the upper half-plane with center $\left(\frac{1}{2}, 0\right)$ with endpoints in the Cantor set K.

All half-circles in the lower half-plane with center $\left(\frac{5}{2 \cdot 3^n}, 0\right)$ $(n \geq 1)$ passing through a point of $K \cap \left[\frac{2}{3^n}, \frac{1}{3^{n-1}}\right]$.

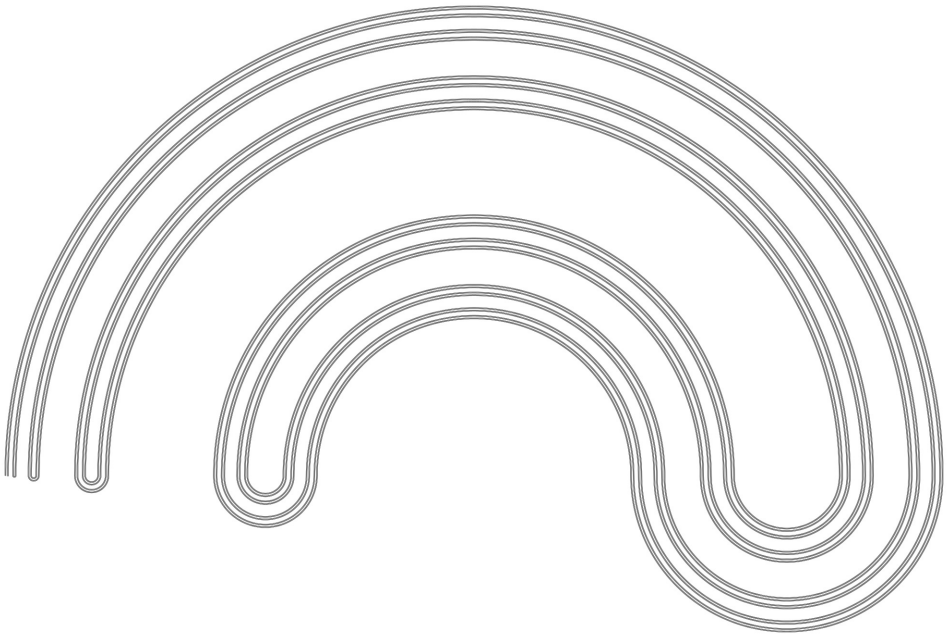


Figure XIII.45: The Knaster bucket-handle

See Figure XIII.45 (from http://upload.wikimedia.org/wikipedia/commons/thumb/3/30/The_Knaster_%22bucket-handle%22_continuum.svg/2000px-The_Knaster_%22bucket-handle%22_continuum.svg.png).

It can be shown that the Knaster bucket-handle is indecomposable (cf. [**?**, 48,V]): although it is connected, it has uncountably many path components (composants), one an injective continuous image of $[0, +\infty)$ and the rest injective continuous images of $\mathbb{R}$. A proper subcontinuum cannot contain any entire path component, so its intersection with any path component is an arc; it follows that it must be contained in a path component, i.e. the complement of any proper subcontinuum is dense.

It turns out there is no Borel subset containing exactly one point from each path component (cf. [**?**]).

There are numerous variations of the construction, with slightly different properties. All these can be described as inverse limits of intervals; see XIII.14.4.18.

Attractors called "horseshoes" () are types of Knaster bucket-handles.

Here is a construction as an intersection, from http://upload.wikimedia.org/wikipedia/commons/thumb/9/98/ContinuBJK.svg/2000px-ContinuBJK.svg.png.



Figure XIII.46: Four steps in the intersection construction of the Knaster bucket-handle

The Pseudo-Arc

XIII.19.5.32. The construction of a pseudo-arc is somewhat involved.

We first need the notion of a *crooked* path. Let R be a (solid) rectangle, subdivided into a strip $R_1, \ldots, R_m$ of subrectangles, a a point in the interior of R_1, and b a point in the interior of R_m:

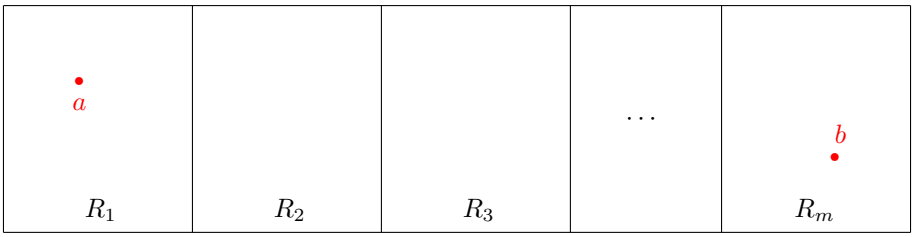


Figure XIII.47: Strip Subdivision of a Rectangle

A path γ from a to b within R is *crooked* with respect to $R_1, \ldots, R_m$ if it has the following property: whenever $t_1, t_2 \in [0,1]$, $t_1 < t_2$, $\gamma(t_1) \in R_i$, and $\gamma(t_2) \in R_j$ with $|i-j| > 2$, there are s_1, s_2 with $t_1 < s_1 < s_2 < t_2$ and

if $i < j$, then $\gamma(s_1) \in R_{j-1}$ and $\gamma(s_2) \in R_{i+1}$

if $j < i$, then $\gamma(s_1) \in R_{j+1}$ and $\gamma(s_2) \in R_{i-1}$.

Instead of a rectangle divided into subrectangles, a chain () of closed balls can be used. A crooked path is defined in the same way.

Thus, for example, if $m = 4$, a crooked path from a to b must first go to R_3, then back to R_2, and then to R_4. So the simplest crooked path looks like this:

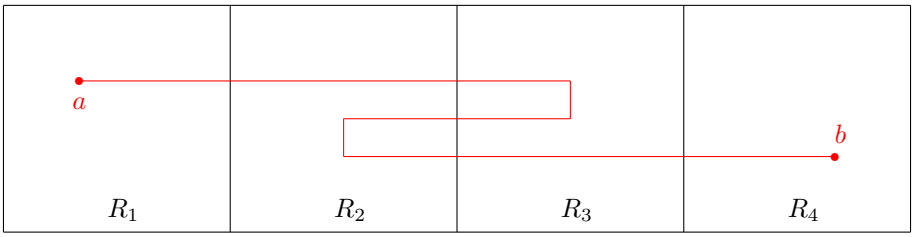


Figure XIII.48: Crooked Path of Length 4

Crooked paths rapidly become more complex as m increases (if $\ell(m)$ is the number of horizontal segments in a minimal crooked path of length m, for $m \geq 4$ we have the recursive formula

$$\ell(m) = 2\ell(m-1) + \ell(m-2)$$

with $\ell(2) = \ell(3) = 1$). For example, if $m = 5$, a crooked path from a to b must first go to R_3, then back to R_2, then R_4, then R_2, then R_4, then R_3, and finally R_5:

A crooked path from a to b in general consists of a crooked path from a to a point in R_{m-1}, followed by a crooked path from that point to a point in R_2, followed by a crooked path from that point to b. Thus

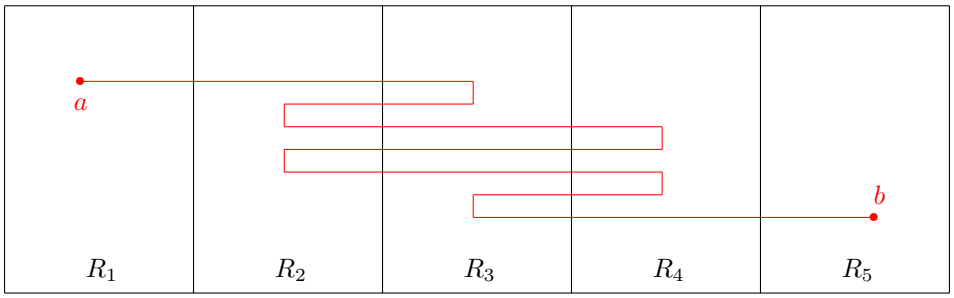


Figure XIII.49: Crooked Path of Length 5

the picture on [HY88, p. 142] is not correct: an actual crooked path is far more complicated (there would be 14,398,739,476,117,879 $\approx 1.4 \times 10^{16}$ horizontal segments for a minimal crooked path of length 45).

To construct a pseudo-arc, fix two points a and b in $\mathbb{R}^2$, and then:

(1) Let X_1 be a narrow (solid) rectangle of width < 1, containing a and b in its interior, and subdivide X_1 into a strip of subrectangles $R_{1,1}, \ldots, R_{1,k_1}$, each of diameter ≤ 1, with a in the interior of $R_{1,1}$ and b in the interior of R_{1,k_1}. Alternatively, let X_1 be the union of a chain of closed balls from a to b.

(2) Take a path from a to b in X_1 which is simple (does not cross itself) and crooked for the $R_{1,j}$.

(3) Fatten up the path within X_1 into a thin strip X_2 homeomorphic to a (solid) rectangle, and subdivide X_2 into a strip of subsets $R_{2,1}, \ldots, R_{2,k_2}$ homeomorphic to subrectangles, of diameter $\leq \frac{1}{2}$, with a in the interior of $R_{2,1}$ and b in the interior of R_{2,k_2}. Alternatively, cover the path with a chain of closed balls, and let X_2 be the union.

(4) Take a simple path from a to b in X_2 which is crooked for the $R_{2,j}$, and iterate the process.

Let $X = \cap_{n=1}^{\infty} X_n$. Then X is a compact subset of $\mathbb{R}^2$ containing a and b. Since each X_n is homeomorphic to a (solid) rectangle, it is connected; thus X is connected by XIII.13.3.6.. X is called a *pseudo-arc*.

A pseudo-arc is compact and connected (a *continuum*), and topologically one-dimensional. It is an *indecomposable continuum*: it cannot be written as a union of two proper subcontinua. In this respect it is enormously different from an ordinary arc. In fact, a pseudo-arc is a *hereditarily indecomposable continuum*: every subcontinuum is also indecomposable. Actually, every subcontinuum of the pseudo-arc with more than one point is homeomorphic to the whole pseudo-arc (an ordinary arc has this property too). And unlike an ordinary arc which has endpoints topologically different from the other points of the arc, a pseudo-arc is *homogeneous*: for any points p and q, there is a homeomorphism of the pseudo-arc to itself sending p to q. These properties are proved in [Bin48], where it is also shown that the pseudo-arc is the *unique* homogeneous hereditarily indecomposable continuum, and also [Bin51a] the unique chainable () hereditarily indecomposable continuum. In particular, up to homeomorphism the pseudo-arc does not depend on the details of the choice of the X_n, so it makes sense topologically to refer to *the* pseudo-arc. The pseudo-arc can be constructed as an inverse limit of intervals (XIII.14.4.18.(ii)).

There is no nonconstant continuous function from an ordinary arc into a pseudo-arc; thus the path components of a pseudo-arc are singletons, i.e. the pseudo-arc is connected but totally path-disconnected.

XIII.19.5.33. Pseudo-arcs may seem pathological, but "almost every" continuum in $\mathbb{R}^2$ is a pseudo-arc [Bin51a], in the sense that if the set $\mathcal{C}$ of continua in $\mathbb{R}^2$ is given the Hausdorff metric () making it into a locally compact metrizable space, the set of continua which are not pseudo-arcs is meager ().

See http://topology.auburn.edu/pm/pseudodraw.pdf for an interesting discussion of the pseudo-arc, including the difficulties inherent in trying to draw a picture of it.

XIII.19.5.34. There is also a similar *pseudo-circle*, described in [Bin51a], whose complement is homeomorphic to the complement of a circle (a bounded component homeomorphic to a disk and an unbounded one). The pseudo-circle is also a hereditarily indecomposable continuum (every proper subcontinuum is a point or a pseudo-arc), and there is also a uniqueness statement up to homeomorphism for pseudo-circles. However, the pseudo-circle is not homogeneous ([**?**], [**?**]). This is a little counterintuitive, since an ordinary circle is homogeneous and an ordinary arc is not; the situation is reversed in the pseudo case. The homogeneity/nonhomogeneity of these spaces is a very subtle matter. There is another similar continuum in the plane, called a *circle of pseudo-arcs* [BJ59], which separates the plane like a circle and is homogeneous (but not indecomposable). There are in fact $2^{\aleph_0}$ topologically distinct hereditarily indecomposable continua in the plane [Bin51a] (there are also hereditarily indecomposable continua of every dimension; cf. [Bin51b], [Kel42]).

There are bizarre plane continua which look nothing like a pseudo-arc: an example is given in [AC59], also constructed as an intersection (or inverse limit), with the property that no two proper subcontinua (containing more than one point) are homeomorphic. In fact, since this continuum has $2^{\aleph_0}$ distinct infinite subcontinua, there are $2^{\aleph_0}$ such continua in the plane, no two of which are homeomorphic. Such a continuum can contain neither an arc nor a pseudo-arc (and hence cannot contain any of the other examples of this section), has empty interior and is topologically one-dimensional, and is totally inhomogeneous (XIII.19.6.11.). Such a space is not indecomposable; in fact, it contains no indecomposable subcontinuum of more than one point.

The Lakes of Wada

XIII.19.5.35. This subset of $\mathbb{R}^2$, first described by K. Yoneyama [] and credited by him to T. Wada, is constructed as an intersection. The pun name "lakes of Wada" may have been first used in [HY88]. A somewhat similar example was previously described by Brouwer [].

A tropical island X_0 named Wada is surrounded by blue ocean. There are two lakes on the island, one with green water and one with brown, with no islands (i.e. the lakes are homeomorphic to open disks). At the first step, canals are dug from the ocean so that every spot of land is within 1 km of blue water. Call the remaining land X_1. In the second step, canals are dug from the green lake to within 1/2 km of every spot of land, so that the green lake plus canals is still homeomorphic to an (open) disk. Call the remaining land X_2. In the third step, canals are dug from the brown lake to within 1/4 km of every spot of land. In the fourth step, the ocean is used again, and then the green lake, then the brown, repeated cyclically. Call the remaining land X after infinitely many steps, i.e. $X = \cap_{n=1}^{\infty} X_n$. X is called the *Land of Wada*.

The construction can (in principle) clearly be continued indefinitely, since each X_n is homeomorphic to a large closed disk with two small open disks removed from its interior (with the boundaries of the three disks pairwise disjoint). At the end, each lake is still homeomorphic to an open disk, and the ocean to the complement in $\mathbb{R}^2$ of a closed disk (if the construction is done on a 2-sphere such as the surface of the earth instead of $\mathbb{R}^2$, the ocean is also an open disk). However, the boundary of each of these three disjoint connected open subsets of $\mathbb{R}^2$ is the set X: every point of X is arbitrarily close to blue, green, and brown water. The set X is compact and connected, and has empty interior (it is topologically one-dimensional, and

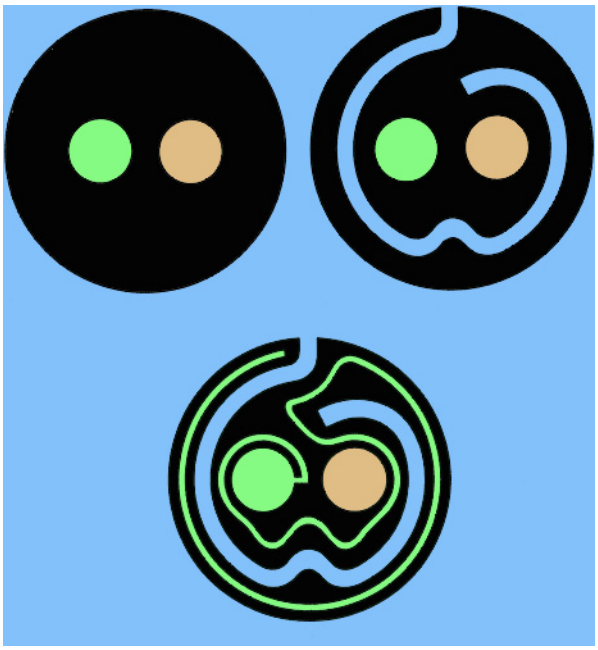


Figure XIII.50: Two steps in the construction of the Lakes of Wada

topologically similar to a pseudo-arc). It is unclear to what extent the homeomorphism type of X depends on the details of the construction of the canals.

Instead of two lakes, we can start with n lakes and end up with $n+1$ disjoint connected open sets in $\mathbb{R}^2$ whose common boundary is the same connected compact set X. With care, the construction can even be done with countably many initial lakes. Thus, if n is any natural number or $\aleph_0$, there are n disjoint open subsets of the 2-sphere, each homeomorphic to an open disk, with dense union, which share a common boundary.

Basins of attraction for nonlinear systems frequently resemble lakes of Wada. These basins of attraction are called *Wada basins*.

The lakes of Wada show again (cf. XIII.19.5.7.) that care must be exercised in formulating the Four-Color Theorem.

Nontrivial Embeddings

There are classes of examples of subsets of $\mathbb{R}^3$ (or $\mathbb{R}^n$) in which the subset itself is a familiar or nicely behaved one, but is embedded in Euclidean space in an unusual or strange way. There are two general subtypes: "tame" embeddings which are locally nice (e.g. smooth embeddings of manifolds), and "wild" embeddings which are not locally nice.

Knots

This is the simplest instance of interesting tame embeddings, consisting of topological circles embedded in $\mathbb{R}^3$ in a nontrivial way.

XIII.19.5.36. DEFINITION. A *knot* is a smooth embedding of S^1 into $\mathbb{R}^3$.

Knot theory is a fascinating and active branch of modern topology; see Figure XIII.51, from http://www.math.unl.edu/~mbrittenham2/ldt/table9.gif, for some examples of knots (some of the pictures are of *links*, made up of more than one circle), and Figure XIII.52 for a nice depiction of the simplest nontrivial knot, the trefoil knot (in a mathematical knot, the blue tube has no thickness). There are many other knot images and animations on the internet. See e.g. [] for more information on knot theory.

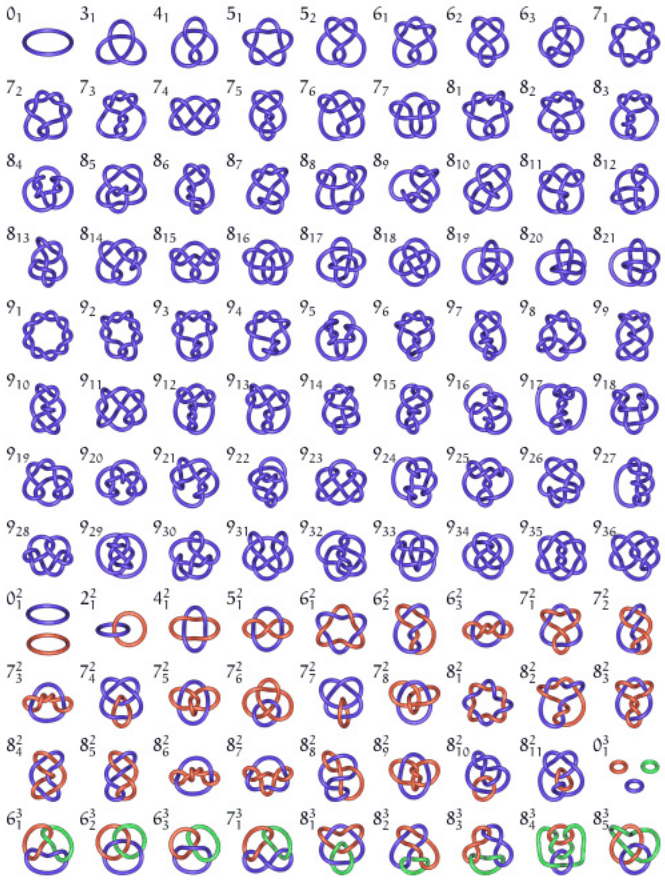


Figure XIII.51: Some Knots and Links



Figure XIII.52: A Trefoil Knot. Artwork by Jim Belk

Higher-Dimensional Knots

XIII.19.5.37. One might also wonder about subsets of $\mathbb{R}^n$ (or S^n) homeomorphic to S^m for $1 < m < n-1$. Such subsets always have connected complements, but the ways they can be embedded (even leaving out "wild" embeddings, e.g. restricting to smooth or piecewise-linear embeddings) are very complicated.

Wild Arcs

XIII.19.5.38. An *arc* in a topological space X is a copy of $[0,1]$ in X (a subspace of X homeomorphic to $[0,1]$). Even an arc in $\mathbb{R}^3$ can be wildly embedded. The simplest and best-known example is due to Fox and Artin [FA48] (Figure XIII.53, from http://www.sciencedirect.com/science/article/pii/S0960077901002429, originally from [HY88]):

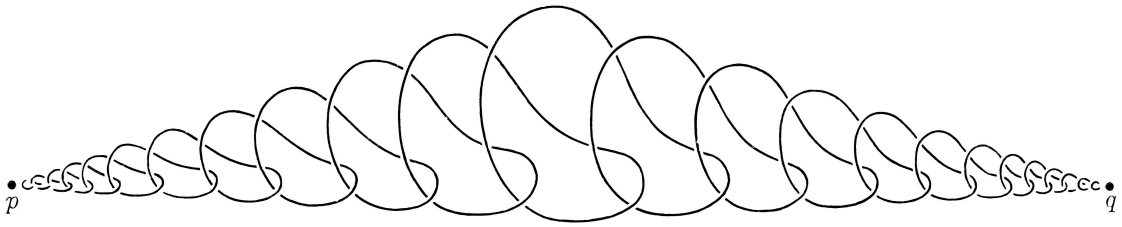


Figure XIII.53: The Fox-Artin Wild Arc

This embedding is wild at the two endpoints. The complement of the arc is not simply connected.

XIII.19.5.39. By successively replacing small subarcs of this arc with a small copy of the whole wild arc, a sequence of embeddings of $[0,1]$ into $\mathbb{R}^3$ can be obtained which are wild at many points. If done carefully, this sequence of embeddings can be made to converge uniformly on $[0,1]$ to an embedding (arc) which is wild at every point.

See [Bor47] for a different kind of pathological arc in $\mathbb{R}^3$.

Alexander's Horned Sphere

XIII.19.5.40. The Alexander horned sphere is a subset of $\mathbb{R}^3$ homeomorphic to S^2 whose exterior is not homeomorphic to the exterior of the unit sphere in $\mathbb{R}^3$ (it is not simply connected ()). The union of the horned sphere and its interior is homeomorphic to B^3, and the interior is homeomorphic to the open unit ball. By extending the horns into the ball instead of outside, a homeomorph of S^2 in $\mathbb{R}^3$ can be made whose interior is not topologically a ball, and by making horns both inside and outside both the interior and exterior can be made pathological.

There are many pictures and animations of the horned sphere on the internet (Figure XIII.55 is from http://conan777.files.wordpress.com/2010/04/hornedsphere3.jpg), although I still think the best picture is the classic one in [HY88] (Figure XIII.54, from http://www.sciencedirect.com/science/article/pii/S0960077901002429). For a beautiful animation of the Alexander Horned Sphere, see https://www.youtube.com/watch?v=Pe2mnrLUYFU.

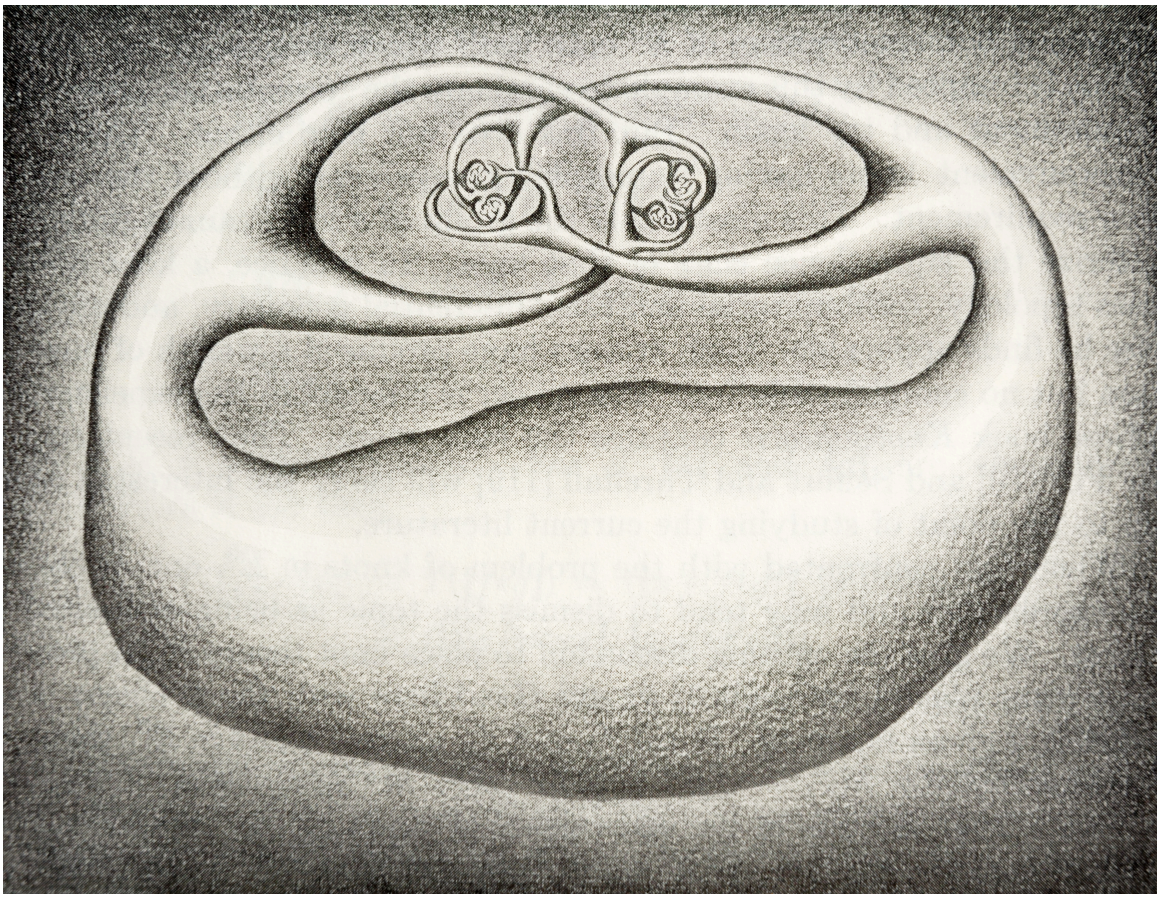


Figure XIII.54: The Alexander Horned Sphere

XIII.19.5.41. The Alexander Horned Sphere is wild at a Cantor set of points. By successively adding small pairs of horns densely, a sequence of embeddings of S^2 into $\mathbb{R}^3$ can be obtained which converges uniformly to an embedding which is wild at every point (cf. XIII.19.5.39.). The horns can be added outside, inside, or both inside and outside.

XIII.19.5.42. There are other rather different wild embeddings of S^2 into $\mathbb{R}^3$. A simple one is to fatten up the Fox-Artin wild arc (Figure XIII.53) into a tube tapering to a point at the ends. A variation, also discussed in [FA48], is shown in Figure XIII.56, from [GH78] (the knots repeat and approach p, q, r, s). See also [Rus73]. Actually, the first wild sphere was obtained by L. Antoine (who was blind!) [**?**], who also constructed Antoine's necklace (XIII.19.5.43.).

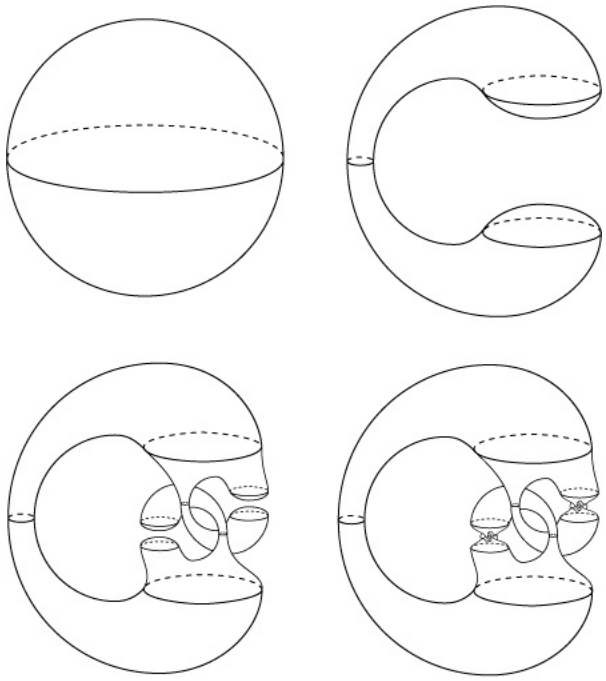


Figure XIII.55: Construction of an Alexander horned sphere

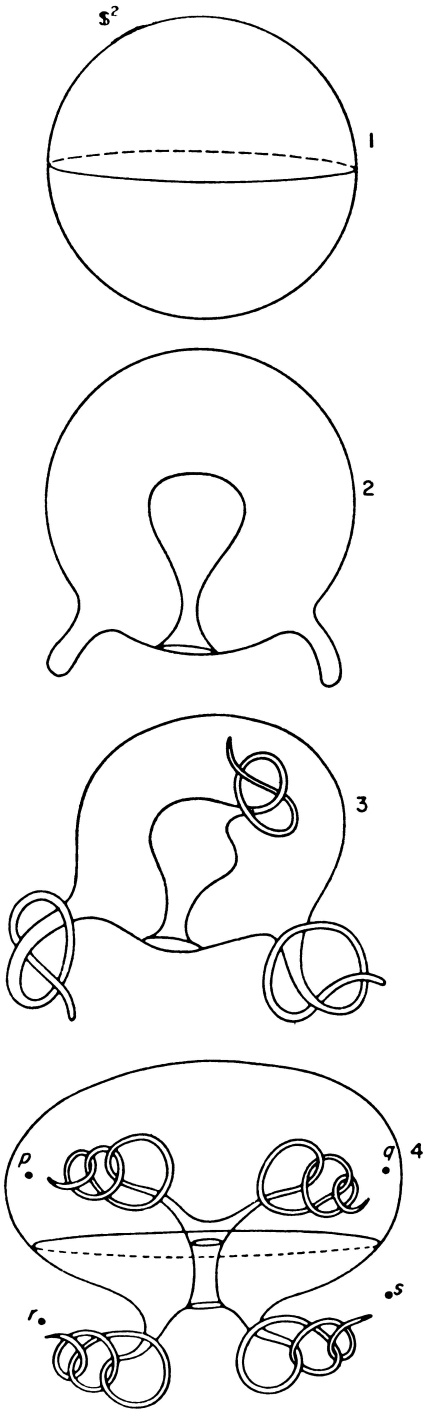


Figure XIII.56: A Few Steps in the Construction of a Wild Sphere

Antoine's Necklace

XIII.19.5.43. This is another subset of $\mathbb{R}^3$ defined by an intersection, a Cantor set wildly embedded in $\mathbb{R}^3$. Begin with a solid torus X_1 in $\mathbb{R}^3$. Inside take a chain X_2 of smaller interlocked solid tori as in Figure XIII.57 (the number of interlocked tori is unimportant as long as it is at least three, although the final embedding of the Cantor set depends on the number of tori at each stage [She68]):

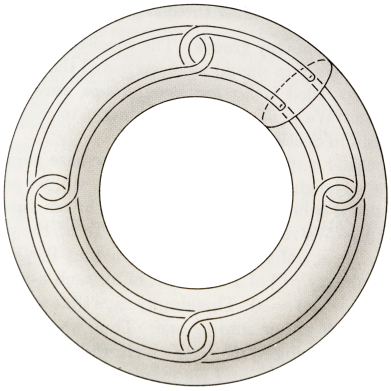


Figure XIII.57: First step in the construction of Antoine's Necklace

Inside each of these smaller solid tori put a similar chain of even smaller interlocked solid tori, and iterate in such a way as to make the diameters of the solid tori at the n'th stage X_n go to zero as $n \to \infty$. The intersection X is *Antoine's Necklace*.

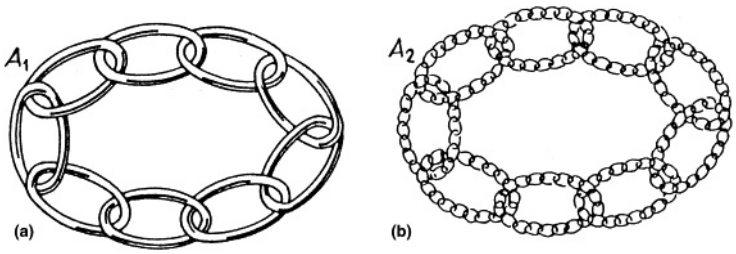


Figure XIII.58: Two steps in the construction of Antoine's Necklace

Figure XIII.59 shows three stages of the construction.

The connected components of X are contained in the connected components of X_n for each n; since the diameters of these connected components approach zero as $n \to \infty$, the connected components of X have diameter zero, i.e. are single points. Thus X is totally disconnected. However, it is clear that X can have no isolated points, since every point of X is arbitrarily close to points from different components of X_n for large enough n. Thus X is homeomorphic to the Cantor set K ().

However, the complement of X in $\mathbb{R}^3$ is not simply connected. If γ is a loop around the "hole" in X_1, then γ cannot be shrunk to a point in the complement of X_n for any n (this seems "obvious", but some

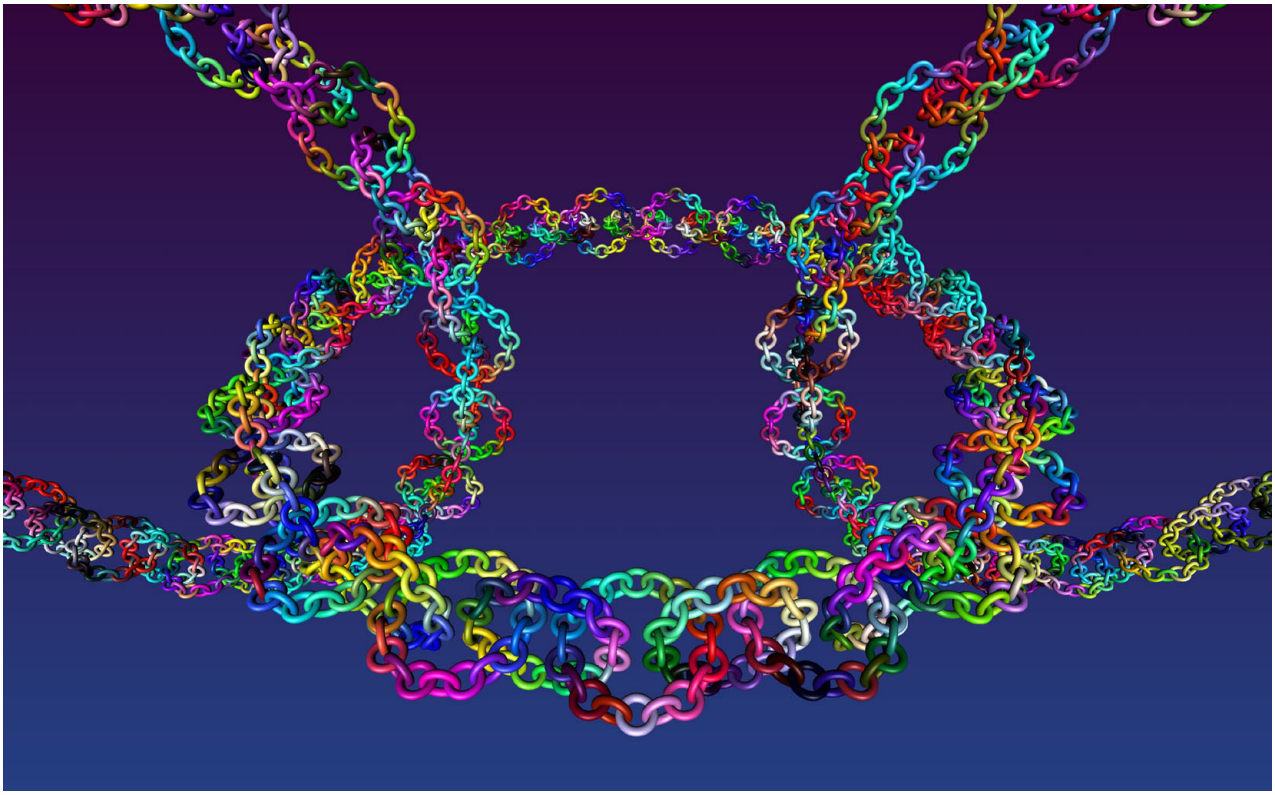


Figure XIII.59: Three steps in the construction of Antoine's Necklace

nontrivial topology is needed to prove it). If it could be shrunk to a point in the complement of X, the homotopy would trace out a continuous image of a disk, a compact set Y in $\mathbb{R}^3$. This set would be disjoint from X_n for sufficiently large n by a simple compactness argument (the complements of the X_n give an open cover of Y). Thus γ could be shrunk to a point in the complement of some X_n, a contradiction.

So X is embedded in $\mathbb{R}^3$ in a very different way than the ordinary Cantor set, regarded as a subset of the x-axis in $\mathbb{R}^3$. The construction and its variations are described in [BM88], where the set is described as a necklace that "cannot fall apart" despite having no string.

It also follows from results in [BM88] that any uncountable compact space which can be embedded in a surface in $\mathbb{R}^3$ (e.g. any uncountable compact subset of $\mathbb{R}^2$) can be embedded in $\mathbb{R}^3$ in uncountably many essentially different ways.

It is known that any subset Z of $\mathbb{R}^2$ homeomorphic to the ordinary Cantor set K (sitting in the x-axis) is embedded in the same way as K: there is a homeomorphism of $\mathbb{R}^2$ onto itself carrying Z onto K [CV98]. Compare with XVII.6.9.3..

See [Rus92] for some generalizations of Antoine's necklace and wild embeddings of spheres in $\mathbb{R}^n$. See also Exercise XIII.19.6.4..

Other Examples

XIII.19.5.44. These examples only scratch the surface of the possibilities. Many other examples can be found in books like [SS95], [Bor67], [Edg08], and [Fal14].

When noncompact subsets of Euclidean space are considered, the possibilities become far greater and more bizarre. For example:

XIII.19.5.45. PROPOSITION. [Kur66, p. 263] Up to homeomorphism, there are $2^{2^{\aleph_0}}$ distinct subsets of $\mathbb{R}$, or of any separable metric space of cardinality $2^{\aleph_0}$.

PROOF: If A is a subset of $\mathbb{R}$, then A is separable (in the relative topology). There are only $2^{\aleph_0}$ continuous functions from any separable space to $\mathbb{R}$, since such a function is completely determined by its values on a countable dense set. Thus there can only be $2^{\aleph_0}$ subsets of $\mathbb{R}$ homeomorphic to A. Since there are $2^{2^{\aleph_0}}$ subsets of $\mathbb{R}$, there must be $2^{2^{\aleph_0}}$ homeomorphism classes. 👍

Since there are only $2^{\aleph_0}$ Borel sets in $\mathbb{R}$ (), "most" of these sets are not even homeomorphic to Borel sets. (Actually, any subset of $\mathbb{R}$ homeomorphic to a Borel set is a Borel set (XIII.10.3.8.).)

XIII.19.6. Exercises

XIII.19.6.1. Let $\boldsymbol{k}$ and $X_{\boldsymbol{k}}$ be as in XIII.19.5.1.. Suppose $\boldsymbol{k}'$ is another sequence of natural numbers and $\phi : X_{\boldsymbol{k}} \to X_{\boldsymbol{k}'}$ is a homeomorphism.

(a) For various points p in $X_{\boldsymbol{k}}$, consider the number and homeomorphism type of connected components of $X_{\boldsymbol{k}} \setminus \{p\}$. Observe that this data must be the same for $X_{\boldsymbol{k}'} \setminus \{\phi(p)\}$.

(b) Use (a) to show that necessarily $\phi\left(\left(\frac{1}{n}, 0\right)\right) = \left(\frac{1}{n}, 0\right)$ for all n.

(c) Conclude that $\boldsymbol{k} = \boldsymbol{k}'$.

(d) If the sequences $\boldsymbol{k}$ and $\boldsymbol{k}'$ are strictly increasing and distinct, there is a simpler proof that $X_{\boldsymbol{k}}$ and $X_{\boldsymbol{k}'}$ are not homeomorphic, since for each $n \geq 2$ there is a unique point $p \in X_{\boldsymbol{k}}$ such that $X_{\boldsymbol{k}} \setminus \{p\}$ has exactly $k_n + 2$ components. Show that there are $2^{\aleph_0}$ such sequences [they are in one-one correspondence with the infinite subsets of $\mathbb{N}$].

XIII.19.6.2. This problem is an elaboration of XIII.19.5.4.. Recall the notation of (). For a compact space X, let X_∞ be the stable perfect subset (K in the case $X = X_{\boldsymbol{m}}$ for some $\boldsymbol{m}$), $S^{(k)}(X)$ the set of isolated points of the k'th derived set $X^{(k)}$, and

$$B^{(k)}(X) = \left[\overline{S^{(k)}(X)} \setminus S^{(k)}(X)\right] \cap X_\infty \ .$$

A homeomorphism from X to a space Y must send X_∞ to Y_∞ and $B^{(k)}(X)$ to $B^{(k)}(Y)$ for each k. Show that $B^{(k)}(X_{\boldsymbol{m}}) \neq B^{(k+1)}(X_{\boldsymbol{m}})$ if and only if $k = m_n$ for some n. Conclude that the $X_{\boldsymbol{m}}$ are mutually nonhomeomorphic for distinct $\boldsymbol{m}$.

XIII.19.6.3. (a) Let X be a compact connected subset of $\mathbb{R}^2$. Show that if $\mathbb{R}^2 \setminus X$ is connected, it is homeomorphic to $\mathbb{R}^2 \setminus \{0\}$. [Use XIII.23.10.1., and show the boundary there can be taken to be a single curve.] Contrast with XIII.20.5.14..

(b) Show more generally that if X and Y are compact subsets of $\mathbb{R}^2$ such that $U = \mathbb{R}^2 \setminus X$ and $V = \mathbb{R}^2 \setminus Y$ are connected, then U and V are homeomorphic if and only if the component spaces X_c and Y_c (XIII.13.9.15.) are homeomorphic. This generalizes both (a) and XIII.20.5.14..

(c) Let X be a compact connected subset of $\mathbb{R}^2$. Show that the unbounded component of $\mathbb{R}^2 \setminus X$ is homeomorphic to $\mathbb{R}^2 \setminus \{0\}$, and that each bounded component is homeomorphic to an open disk. [For the unbounded component, apply (a) to the union of X and the bounded components of $\mathbb{R}^2 \setminus X$. For bounded components, invert in a small circle within the component.] Make an analogous statement about complements of compact connected subsets of S^2. Compare with XII.8.6.6..

These results are special to $\mathbb{R}^2$ (analogous results hold more simply in $\mathbb{R}^1$): several examples from this section show that nothing like this holds in $\mathbb{R}^3$.

XIII.19.6.4. Modify the construction of Antoine's Necklace as follows.

(a) Consider the following linked set of solid tori (picture from [Gar01]; the surfaces should not touch):



Figure XIII.60: Variation of Borromean Rings

The number n of solid tori is not important as long as it is at least 3 (although, as with Antoine's Necklace, the number probably affects the final embedding); the case $n = 3$ is the usual Borromean Rings (Figure XIII.51 6_2^3). Show that the entire set of tori is linked nontrivially, that any two of the solid tori are not linked, and that if any one of the tori is removed the remaining ones are unlinked.

(b) Replace each of these solid tori by a set of smaller solid tori contained in it, linked in the same way. Iterate the construction so that the diameters of the solid tori go to zero, and take the intersection X.

(c) Show that X is homeomorphic to the Cantor set.

(d) Show that the complement of X is not simply connected; thus X is nontrivially (wildly) embedded in $\mathbb{R}^3$.

(e) If T is one of the solid tori in the construction, then $X \cap T$ is a clopen set in X. Show that every clopen set in X is a finite disjoint union of such sets.

(f) Two disjoint compact subsets A and B of $\mathbb{R}^n$ are *unlinked* (or *separated by spheres*) if there is a homeomorphism of $\mathbb{R}^n$ sending A and B into disjoint open balls. Show that if A and B are disjoint clopen sets in X, then there is a partition of A into clopen sets C and D such that $\{B, C, D\}$ are pairwise unlinked in $\mathbb{R}^3$. Conclude that X is not embedded in $\mathbb{R}^3$ in the same way as Antoine's Necklace (any version).

(g) What happens if we use other linkings of the embedded solid tori at each stage, and perhaps vary the linkings from stage to stage? We could, for example, at infinitely many stages knot the solid tori, or replace each solid torus with a thinner solid torus wrapping twice around inside it as in the construction of the solenoid (Figure XIII.44). See Figure XIII.61 (from T. Gray, *The Elements*), and [She68], [Sko86], and [SS13], and the references therein.



Figure XIII.61: Another link which can be iterated.

XIII.19.6.5. Knotted Solenoids. Let S be a solenoid (XIII.19.5.16.).

(a) By construction, S is a subset of a solid torus T. If K is a knot, fatten up K into a solid torus T_K and let ϕ be a homeomorphism of T onto T_K. Let $S_K = \phi(S)$. S_K can be regarded as a knotted version of S. Show that if K_1 and K_2 are knots, show that S_{K_1} and S_{K_2} are embedded in $\mathbb{R}^3$ in the same way if and only if $K_1 \cong K_2$.

(b) If K is a knot, then K can be embedded nontrivially in a solid torus with winding number greater than 1 (write the knot as the closure of a braid; cf. ()). By fattening up K inside the solid torus, we obtain a

knotted solid torus. Another knot can be embedded in this in the same way. Iterating this construction and taking an intersection, an "infinitely knotted" solenoid is obtained. Classify the solenoid obtained via the winding numbers and show that any solenoid can be embedded in $\mathbb{R}^3$ in $2^{\aleph_0}$ essentially different ways.

XIII.19.6.6. (a) Let S be a solenoid embedded in $\mathbb{R}^3$ in the standard (unknotted) way, with $G = \pi^1(S)$ the corresponding subgroup of $\mathbb{Q}$. Show that the fundamental group of $\mathbb{R}^3 \setminus S$ is G. Note that $\mathbb{R}^3 \setminus S$ is a connected open subset of $\mathbb{R}^3$, hence a connected 3-manifold.

(b) Compute the fundamental groups of the complements of the knotted solenoids of XIII.19.6.5., which are also connected 3-manifolds.

XIII.19.6.7. Let G be any countable group. Show that there is a connected open subset of $\mathbb{R}^5$ (hence a connected 5-manifold) with fundamental group G. [See http://mathoverflow.net/questions/192230/is-there-a-manifold-with-fundamental-group-mathbbq for an outline of the construction.] Conversely, the fundamental group of any manifold is countable ().

XIII.19.6.8. **The Whitehead Continuum and the Whitehead Manifold.** [**?**] Here is an interesting construction reminiscent of the construction of solenoids. Let T_0 be a solid (unknotted) torus in $\mathbb{R}^3$, and T_1 a thinner solid torus embedded in T_0 as in Figure XIII.62.

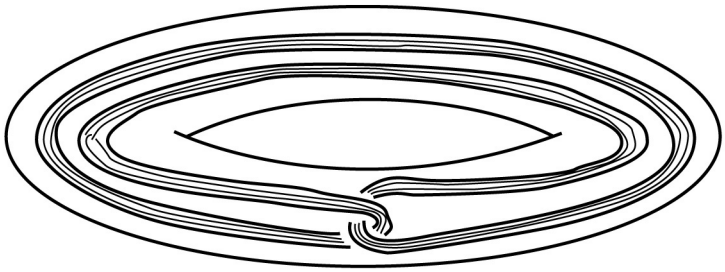


Figure XIII.62: First two steps for the Whitehead continuum

For each k let T_{k+1} be a thin solid torus embedded in T_k in the same way. Note that each T_k is unknotted in $\mathbb{R}^3$.

Let $W = \cap_{k=1}^{\infty} T_k$. Since each T_k is compact and connected, W is also compact and connected. W is called the *Whitehead continuum* (the exact space depends on the details of the construction).

(a) If L is a meridian circle for T_0, show that

(i) L and the longitude axis circle of T_1 form a nontrivial link in $\mathbb{R}^3$, called the *Whitehead link* (XIII.63).

(ii) L is not null-homotopic in $T_0 \setminus T_1$ (or even in $\mathbb{R}^3 \setminus T_1$).

(iii) L is null-homotopic in $S^3 \setminus T_1$, where S^3 is the one-point compactification of $\mathbb{R}^3$.

(b) Discuss other properties of W. Is it path-connected? Simply connected?



Figure XIII.63: The Whitehead link

(c) The complement of W is more interesting. Regard the construction as being done in the one-point compactification S^3 of $\mathbb{R}^3$, and let $E = S^3 \setminus W$. The complement of a solid torus in S^3 is another solid torus; thus E is an increasing union of solid tori. Show that each is embedded in the next as T_1 is embedded in T_0. Show that E is a connected, noncompact 3-manifold, and that E is simply connected [use (a)(iii).] Conclude from () that E is contractible (or give a direct argument). E is called the *Whitehead manifold*. (Unlike for W, up to homeomorphism E does not depend on the details of the construction.)

(d) Show that no neighborhood of ∞ in the one-point compactification $E^\dagger$ of E is simply connected [$E^\dagger$ is homeomorphic to the quotient S^3/W of S^3 obtained by collapsing W to a point. Use (a)(ii).] Thus $E^\dagger$ is not a manifold. Conclude that E is not homeomorphic to $\mathbb{R}^3$.

(e) Show that $E \times \mathbb{R}$ is homeomorphic to $\mathbb{R}^4$, as is $(\mathbb{R}^3/W) \times \mathbb{R}$. (However, contrary to statements in some references like [CWY10], $(S^3/W) \times \mathbb{R} = E^\dagger \times \mathbb{R}$ is not homeomorphic to $\mathbb{R}^4$ since it has two ends and is not contractible.)

(f) If p is a point of W, then $E \subseteq S^3 \setminus \{p\} \cong \mathbb{R}^3$, and thus E is homeomorphic to an open subset of $\mathbb{R}^3$.

(g) Some more properties of E: E is the union of two open subsets, each homeomorphic to $\mathbb{R}^3$, whose intersection is also homeomorphic to $\mathbb{R}^3$ [Gab11]; E does not have any orientation-reversing diffeomorphism [AW07]; E does not admit a complete metric with strictly positive scalar curvature [CWY10], unlike $\mathbb{R}^3$, which does have such a metric, e.g. by identifying it with

$$\{(x, y, z, w) : w = x^2 + y^2 + z^2\} \subseteq \mathbb{R}^4 \ .$$

(h) Let $\boldsymbol{n} = (n_1, n_2, \dots)$ be a sequence in $\mathbb{N}$. Vary the construction of W and E by wrapping T_k around T_{k-1} n_k times before linking as in Figure XIII.62. Let $W_{\boldsymbol{n}}$ be the intersection and $E_{\boldsymbol{n}} = S^3 \setminus W_{\boldsymbol{n}}$. (Thus the Whitehead continuum is $W_{\boldsymbol{n}}$ where $n_k = 1$ for all k.) Show that $W_{\boldsymbol{n}}$ and $E_{\boldsymbol{n}}$ have the same properties as before. Show that $2^{\aleph_0}$ distinct homeomorphism classes of $E_{\boldsymbol{n}}$ are obtained. Thus there are $2^{\aleph_0}$ topologically distinct contractible open subsets of $\mathbb{R}^3$. See [McM62] for details.

In contrast, every simply connected open subset (in particular, every contractible open subset) of $\mathbb{R}^2$ is homeomorphic to $\mathbb{R}^2$ (). In fact, the only simply connected 2-manifolds are $\mathbb{R}^2$ and S^2.

Not every contractible 3-manifold is of the form of (h); cf. [McM62], [TW97]. There does not appear to be any reasonable classification of contractible 3-manifolds. The already known existence of exotic contractible 3-manifolds seems to have been overlooked in [BT82, p. 147].

XIII.19.6.9. The House with Two Rooms. Let X be the subset of $\mathbb{R}^3$ pictured in Figure XIII.64 (from [Hat02]).

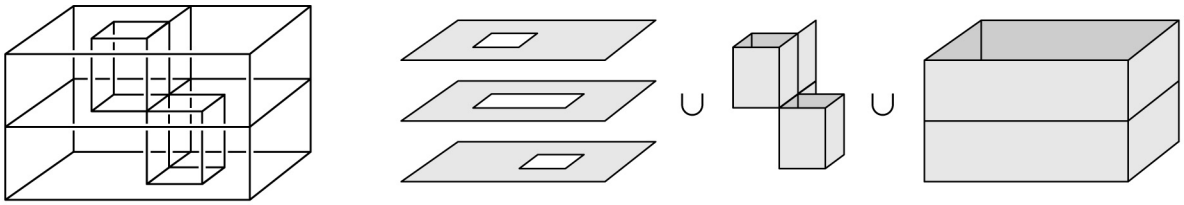


Figure XIII.64: The House With Two Rooms

(a) X is a finite union of planar rectangles, and thus is a two-dimensional (compact) polyhedron in $\mathbb{R}^3$.

(b) Let Y be the rectangular solid in $\mathbb{R}^3$ generated by X (convex hull of X). Show that there is a deformation retraction of Y onto X. [Push down the upper hole and expand the indentation to fill the bottom room, and similarly push up the lower hole.] Conclude () that X is an absolute retract, and in particular is contractible.

(c) Consider the loop in X which is the perimeter of the hole in the upper surface. Show how to contract this loop to a point in X.

(d) Describe an explicit contraction of X. (This is not easy! See e.g. https://sketchesoftopology.wordpress.com/2010/06/23/the-deformation-retraction-of-bings-house/.)

This example is due to Bing [Bin64]. There are several slight variations in various references, all of which have the same essential features.

XIII.19.6.10. The Rational Ruler. Construct a subset X of the closed unit square $[0,1]^2 \subseteq \mathbb{R}^2$ as follows. Begin with the horizontal segment $[0,1]$ on the x-axis, and for each $r \in \mathbb{Q} \cap [0,1]$, $r = \frac{p}{q}$ in lowest terms, add the closed vertical segment between $(r,0)$ and $\left(r, \frac{1}{q}\right)$.

(a) Show that X is a compact subset of $\mathbb{R}^2$.

(b) Construct a retraction of $[0,1]^2$ onto X as follows. Set

$$X_n = X \cup \left\{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq \frac{1}{n}\right\}$$

(note that $X = \cap_n X_n$) and construct a uniformly convergent sequence of retractions from $[0,1]^2$ onto the X_n. Conclude () that X is an absolute retract.

XIII.19.6.11. Say a topological space X is *totally inhomogeneous* if the only homeomorphism from X to X is the identity. Construct $2^{\aleph_0}$ pairwise nonhomeomorphic totally inhomogeneous dendrites in $\mathbb{R}^2$ as follows. Let $(k_1, k_2, \dots)$ be a sequence of distinct natural numbers. As in the construction of the rational ruler (XIII.19.6.10.), begin with $[0,1]$ on the x-axis, and at the n'th rational in $(0,1)$ add k_n whiskers of length $\frac{1}{n}$ (no whiskers at 0 or 1). Replace each whisker by a small copy of the whole set, and iterate. If done carefully, a sequence of retractions from a containing disk can be made to converge uniformly.

XIII.19.6.12. ([Lub53]; cf. [Bor67, p. 150-151]) Let $n \in \mathbb{N}$. Adapt the Lakes of Wada construction to construct in $\mathbb{R}^3$ a connected two-dimensional absolute neighborhood retract X_n such that the complement of X is dense in $\mathbb{R}^3$ and has exactly n components, the boundary of each of which is X. (X thus has a locally nice structure, a least in a homotopy sense, unlike the examples in $\mathbb{R}^2$ like the Lakes of Wada which necessarily are locally quite pathological [**?**].)

XIII.19.6.13. Show that there is no reasonable analog of the four-color theorem in $\mathbb{R}^3$, even for polyhedra. [Lay n boards with rectangular cross-section side-by-side on the xy-plane in $\mathbb{R}^3$ with long dimension parallel to the x-axis, and on top of them lay n boards with long dimension parallel to the y-axis. Glue the k'th horizontal board to the k'th vertical board to make a single cross-shaped polyhedron for each k. See http://demonstrations.wolfram.com/AnInfinitelyColorableSetOf3DRegions/. Another more informal argument I heard somewhere: "Imagine a bowl of spaghetti where every noodle touches all the others."] There are meaningful higher-dimensional analogs of the graph-theoretic versions of the four-color theorem, however.

XIII.19.6.14. The Hahn-Mazurkiewicz Theorem says that a Hausdorff topological space is a continuous image of $[0,1]$ if and only if it is compact, metrizable, connected, and locally connected (and thus automatically path-connected and locally path-connected). Thus the Sierpiński gasket, Sierpiński carpet, and Menger sponge are all continuous images of $[0,1]$. Find explicit continuous surjections. [The construction of a space-filling curve () may provide some insight.]

XIII.19.6.15. **The Infinite Comb.** Consider the compact subset of $\mathbb{R}^2$ consisting of the line segments from 0 to 1 on the x and y axes and the vertical segments from 0 to 1 with x-coordinate $\frac{1}{n}$ for each n (the exact location of the x-coordinates of the vertical segments are unimportant as long as they form a sequence of numbers in $(0,1]$ converging to 0). See Figure XIII.65.

(a) Show that the infinite comb is contractible.

(b) Show that the infinite comb is not locally connected at points on the positive y-axis. Thus it is not strongly contractible to any of these points. It is strongly contractible to any other point.

(c) Figures XIII.66 and XIII.67 show a number of variations of the infinite comb. Which of these are contractible? Which are strongly contractible, to which points? Which pairs of these are homeomorphic? (The first space in XIII.66 is called *Zeno's maze* in [Pug15], but I have not found this name in other references.) Show that all have trivial fundamental group, i.e. are simply connected.

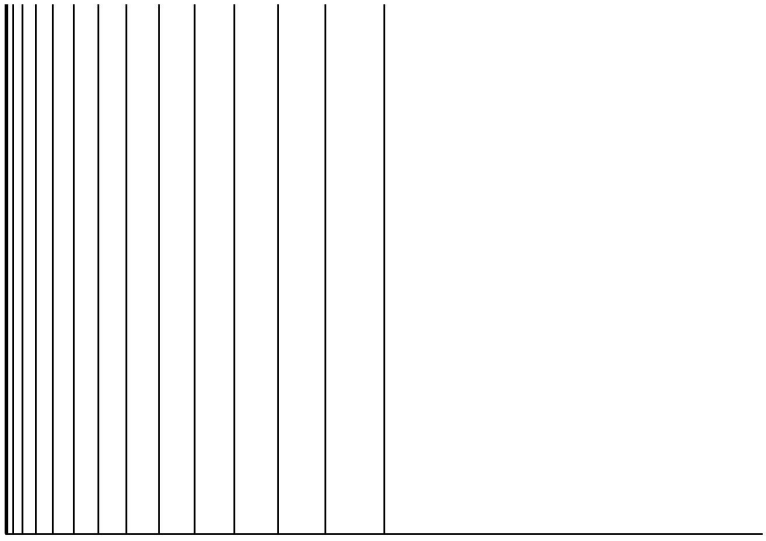


Figure XIII.65: The Infinite Comb

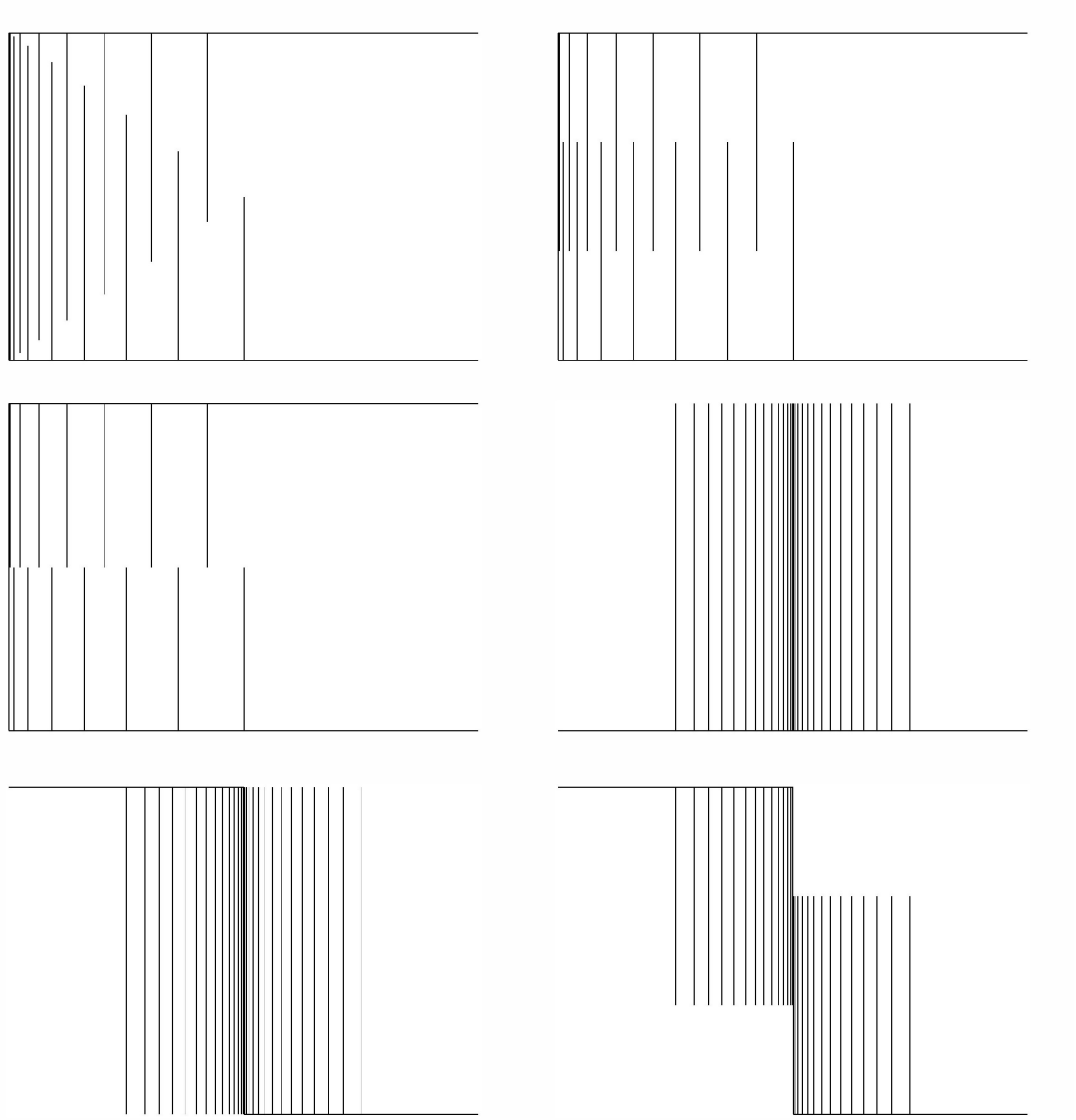


Figure XIII.66: Infinite Comb Variations I

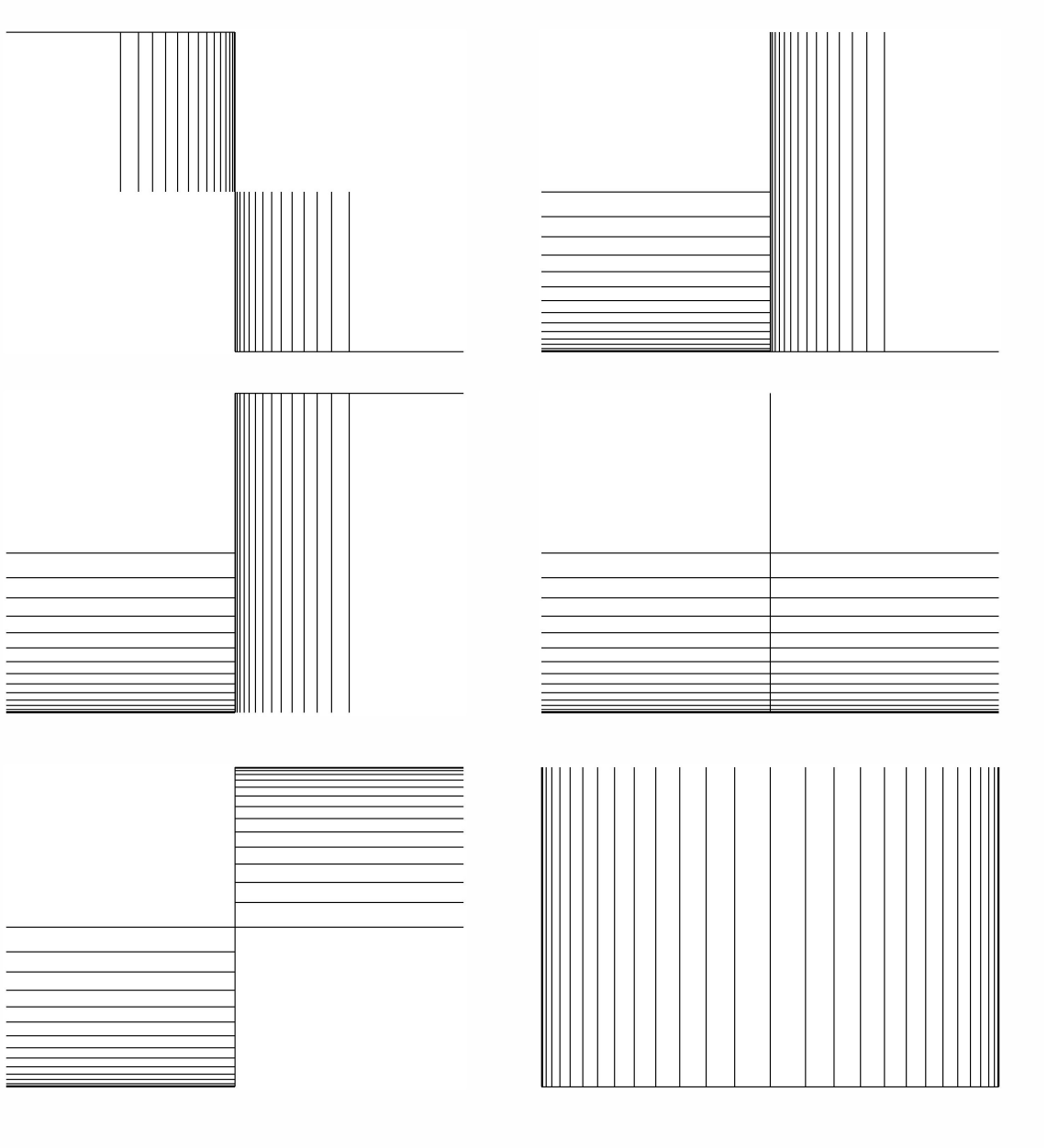


Figure XIII.67: Infinite Comb Variations II

XIII.19.6.16. The Infinite Ladder. A variation of the infinite comb is to add the horizontal segment from 0 to 1 with y-coordinate 1 (Figure XIII.68). This space is somewhat similar to the Hawaiian Earring, but with a crucial difference: show that the fundamental group is countable.

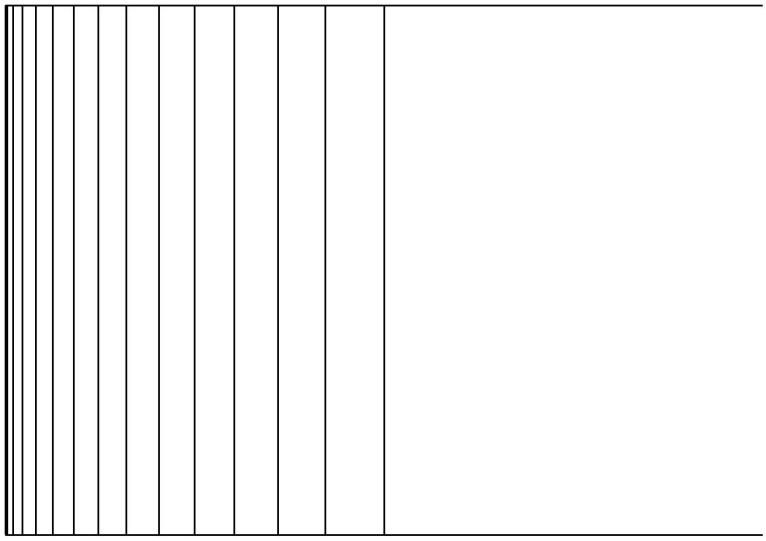


Figure XIII.68: The Infinite Ladder

XIII.20. Manifolds

There are two types of mathematical objects commonly called manifolds: *topological manifolds* and *differentiable*, or *smooth, manifolds* (there are also *PL manifolds*, which are something in between). There is also a notion of "manifold with boundary," both topological and smooth. The term "manifold" is also used in a few other related and even entirely unrelated senses in parts of mathematics. One must be aware of context whenever the term "manifold" is used. Smooth manifolds are special types of topological manifolds with specified extra structure (smooth manifolds come in a number of flavors). Both types of manifolds occur in analysis. In this section, we describe the basics of both theories; each theory is very extensive and goes far beyond what we do here.

XIII.20.1. Topological Manifolds

Topological manifolds are generalizations of curves and surfaces. The property of a curve or surface we use is that near each point, a small piece of the curve or surface looks like an open ball in Euclidean space (of dimension 1 or 2 respectively). Our generalization will allow for higher-dimensional objects also, and will be intrinsic, not requiring our space to be embedded in a Euclidean space.

XIII.20.1.1. Definition. A topological space X is *locally Euclidean* if every point $p \in X$ has an open neighborhood which is homeomorphic to an open ball in $\mathbb{R}^n$ for some $n \geq 0$ (which may vary from point to point), where $\mathbb{R}^0$ is a single point. Such a neighborhood is called a *coordinate patch* and a homeomorphism with an open ball in $\mathbb{R}^n$ is a *coordinate chart*, and the n is called the *local dimension* of X at p. X is *n-dimensional* if it is has the same local dimension n at every point.

Roughly speaking, a topological manifold will just be a topological space which is locally Euclidean. But it turns out we should make a few additional technical restrictions to rule out pathologies that can occur in bizarre examples.

XIII.20.1.2. Some initial remarks about the definition:

(i) The local dimension is well defined by the Invariance of Domain Theorem ().

(ii) All points within a coordinate patch have the same local dimension; thus the local dimension is locally constant (justifying the terminology "local dimension.") Hence the local dimension is constant on components of X (). In particular, a connected locally Euclidean space has the same local dimension at every point.

(iii) An n-dimensional metrizable locally Euclidean topological space has topological dimension n (), so the terminology "n-dimensional" is consistent. The topological dimension of a general metrizable locally Euclidean space is the maximum or supremum of the local dimensions at all points. (This can fail in the nonmetrizable case, cf. [Fed95].)

(iv) Since an open ball in $\mathbb{R}^n$ is first countable, even second countable, a locally Euclidean space is first countable. But it need not be second countable (XIII.20.1.9., XIII.20.5.5.). A locally Euclidean space is locally second countable (XIII.1.3.7.).

(v) Since open balls in $\mathbb{R}^n$ are locally compact, a (Hausdorff) locally Euclidean space is locally compact () and hence completely regular ().

(vi) Since an open ball in $\mathbb{R}^n$ is path-connected, a locally Euclidean space is locally path-connected; hence the components and path components coincide and are clopen sets ().

(vii) An open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$. Thus the definition can be equivalently rephrased: X is locally Euclidean if every point $p \in X$ has an open neighborhood which is homeomorphic to $\mathbb{R}^n$ for some n. This is the form given in many references. However, it seems preferable to write the definition as we have, since in most examples there are more natural homeomorphisms of point neighborhoods with open balls in $\mathbb{R}^n$ than with $\mathbb{R}^n$ itself. For example, it is immediate from our definition that any open set in $\mathbb{R}^n$ is locally Euclidean (in the relative topology).

(viii) Since an open ball in $\mathbb{R}^n$ is T_1 (even T_2), a locally Euclidean topological space is locally T_1 (even locally T_2), hence T_1 (XIII.7.3.5.). But it need not be Hausdorff (XIII.20.1.3.(v)). This is one of the pathologies we want to exclude in the definition of a topological manifold.

XIII.20.1.3. Examples. (i) Any open set in $\mathbb{R}^n$ is an n-dimensional locally Euclidean space with the relative topology.

(ii) A curve in $\mathbb{R}^n$ is a 1-dimensional locally Euclidean space (with the relative topology), provided it has no endpoints and does not cross or loop back onto itself (more precisely, if it is homeomorphic to an open interval or a circle, or a separated disjoint union of such curves). A simple curve (arc) with endpoints is an example of a manifold with boundary (). "Curves" such as a figure 8, a figure 6, or a "Warsaw circle" (XIII.19.5.6.) are not locally Euclidean and are not manifolds with boundary.

(iii) A surface in $\mathbb{R}^3$ is a 2-dimensional locally Euclidean space with the same restrictions as in (ii), which are more complicated to state precisely. Examples include a sphere, torus, a cylinder without its end circles, or a Möbius strip without its edge. A cylinder with end circles included, or a Möbius strip with its edge included, is a manifold with boundary.

(iv) The unit sphere in $\mathbb{R}^{n+1}$ is an n-dimensional locally Euclidean space for any n. The closed unit ball is not locally Euclidean, but is a manifold with boundary.

(v) A locally Euclidean space need not be Hausdorff. Here is the standard example. Let X be $\mathbb{R} \cup \{0'\}$, with the topology consisting of all open subsets of $\mathbb{R}$ and all sets consisting of $0'$ along with an open neighborhood or deleted open neighborhood of 0. Since $\mathbb{R}$ is an open subset of X, every point of $\mathbb{R}$ has a locally Euclidean neighborhood; and $0'$ does also, e.g. $\{0'\} \cup (\mathbb{R} \setminus \{0\})$. Thus X is locally Euclidean (and 1-dimensional). But 0 and $0'$ do not have disjoint neighborhoods. See Exercise XIII.20.5.15. for a more interesting example.

There are far more complicated examples; see Exercises XIII.20.5.5. and XIII.20.5.8..

XIII.20.1.4. Theorem. Let M be a topological space which is Hausdorff, locally compact, and locally second countable, in which every component is open (e.g. Hausdorff and locally Euclidean). The following are equivalent:

(i) M is paracompact.

(ii) M is metrizable.

(iii) Every component of M is Lindelöf.

(iv) Every component of M is σ-compact.

(v) Every component of M is second countable.

(vi) Every component of M is a Polish space.

The following are equivalent, and imply (i)–(vi):

(vii) M is Lindelöf.

(viii) M is σ-compact.

(ix) M is second countable.

(x) M is a Polish space.

If M satisfies (i)–(vi), then M satisfies (vii)–(x) if and only if M has only countably many components. In particular, if M is connected, then (i)–(x) are equivalent.

PROOF: Since the components of M are clopen (), obviously (vii)–(x) each imply that M has only countably components; and if M has only countably components, then (vii) $\iff$ (iii), (viii) $\iff$ (iv), (ix) $\iff$ (v), and (x) $\iff$ (vi). So we need only prove that (i)–(vi) are equivalent.

Since M is locally compact, (iii) $\iff$ (iv) by XIII.11.8.11., and by XIII.11.8.13., (v) and (vi) are equivalent and imply (iv) and that every component of M is metrizable. Since the components of M are clopen, M is metrizable if and only if each component is metrizable (), so (v) $\Rightarrow$ (ii). (ii) $\Rightarrow$ (i) by XIII.11.10.4.. M is paracompact if and only if each component of M is paracompact (XIII.11.10.10.), and if a component is paracompact it is σ-compact by XIII.11.10.20.. Thus (i) $\Rightarrow$ (iv). And since M is locally second countable, (iii) $\Rightarrow$ (v) by XIII.1.3.8.. 👍

XIII.20.1.5. There are connected Hausdorff locally Euclidean topological spaces not satisfying the conditions of Theorem XIII.20.1.4.. Such a space can be normal (Exercise XIII.20.5.5.) or nonnormal (Exercises XIII.20.5.8.–XIII.20.5.9.). Separability is also not equivalent to the conditions of the theorem; such examples can be separable (Exercise XIII.20.5.9.) or nonseparable (Exercises XIII.20.5.5. and XIII.20.5.8.). Conditions (vii)–(x) imply separability, and are equivalent to separability in the presence of (i)-(vi).

XIII.20.1.6. There is no universal consensus on the proper definition of a topological manifold. All of the following are reasonable candidates which have in fact all appeared in various references. They are arranged roughly, but not exactly, in increasingly restrictive order. Throughout, X denotes a topological space. ✋

(i) X is locally Euclidean.

(ii) X is Hausdorff and locally Euclidean.

(iii) X is Hausdorff, locally Euclidean, and satisfies conditions (i)–(vi) of XIII.20.1.4..

(iv) X is Hausdorff, locally Euclidean, and satisfies conditions (i)–(x) of XIII.20.1.4..

(i)′ X is locally Euclidean and has constant local dimension.

(ii)′ X is Hausdorff, locally Euclidean, and has constant local dimension.

(iii)′ X is Hausdorff, locally Euclidean, satisfies conditions (i)–(vi) of XIII.20.1.4., and has constant local dimension.

(iv)′ X is Hausdorff, locally Euclidean, satisfies conditions (i)–(x) of XIII.20.1.4., and has constant local dimension.

(i)″ X is connected and locally Euclidean.

(ii)″ X is Hausdorff, connected, and locally Euclidean.

(iii)″ X is Hausdorff, connected, locally Euclidean, and satisfies conditions (i)–(vi) of XIII.20.1.4..

(iv)″ X is Hausdorff, connected, locally Euclidean, and satisfies conditions (i)–(x) of XIII.20.1.4..

(Note that (iii)″ and (iv)″ are identical.)

We will take (iii) as our definition, which is probably the most common choice. Reasonable arguments can be made for all of the versions (although the arguments for (i)–(i)″ are weak; manifolds should at least be Hausdorff). Almost all spaces arising in practice which one would want to call manifolds satisfy (iv)″ or at least (iv)′. However, almost all properties of such spaces hold equally well for ones satisfying (iii), with identical proofs; practically all interesting things about manifolds happen separately in each component, so it doesn't really matter how many components there are as long as each individual component is well behaved. Definition (iii) can also be given very cleanly:

XIII.20.1.7. Definition. A topological space X is a *(topological) manifold* if it is metrizable and locally Euclidean.

A space satisfying (iv)″ is then just a connected manifold, one satisfying (iv) is a separable manifold, and one satisfying (iii)′ is an n-dimensional manifold (usually just called an n-manifold) for some n. It seems rather artificial to allow manifolds with nonconstant local dimension, and they rarely arise in practice; we can restrict to the constant local dimension case by just using the term "n-manifold".

XIII.20.1.8. A compact Hausdorff locally Euclidean space automatically satisfies (viii), hence (i)–(x), and in particular is a manifold.

XIII.20.1.9. The case $n = 0$ is allowed in the definition of a manifold. A 0-manifold is just a set with the discrete topology.

XIII.20.2. Topological Manifolds With Boundary

We would like to extend the notion of manifold to include such examples as curves with endpoints, surfaces with boundary curves, and closed balls in $\mathbb{R}^n$. These will be examples of "manifolds with boundary."

The basic tool is the notion of a half-space:

XIII.20.2.1. Definition. The *n-dimensional half-space* ($n \geq 1$) is the subset

$$\mathbb{H}^n = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_n \geq 0\}$$

with the relative topology from $\mathbb{R}^n$. Its boundary $\partial\mathbb{H}^n$ is the set of n-tuples with last coordinate zero. We also define $\mathbb{H}^0$ to be a single point, with $\partial\mathbb{H}^0 = \emptyset$.

XIII.20.2.2. DEFINITION. A *(topological) manifold with boundary* is a metrizable space X in which every point has an open neighborhood homeomorphic to an open set in $\mathbb{H}^n$ for some $n \geq 0$ (which may vary from point to point.) Such an open set is called a *coordinate patch*, and the homeomorphism is called a *coordinate chart*. If the n is constant, X is called an *n-manifold with boundary*. A point of X with a neighborhood homeomorphic to $\mathbb{R}^n$ (or to an open set in $\mathbb{R}^n$) is called an *interior point* of X, and a point which maps to a point in $\partial\mathbb{H}^n$ under a coordinate chart is called a *boundary point* of X. Denote by ∂X the set of boundary points of X, called the *boundary* of X.

Note that the use of "interior point" and "boundary point" here is not the same as the general use in topology: such points may not be interior or boundary points respectively in the topological sense. Indeed, since a manifold with boundary is not naturally embedded as a subset of a larger topological space, topological interior and boundary do not really even make sense.

XIII.20.2.3. Just as with manifolds, manifolds with boundary are locally compact and locally path-connected, and the components and path components coincide and are clopen subsets. Each component is second countable and σ-compact, and a Polish space.

XIII.20.2.4. EXAMPLES. (i) $\mathbb{H}^n$ is an n-manifold with boundary. The boundary points are the points of $\partial\mathbb{H}^n$.

(ii) Any (topological) manifold X is a manifold with boundary, with $\partial X = \emptyset$. (So the term "manifold with boundary" does not imply that the boundary is nonempty.)

(iii) Since $\mathbb{H}^1$ is a ray, a closed or half-open interval I in $\mathbb{R}$ is a 1-manifold with boundary ∂I consisting of the endpoint(s). In fact, a complete list (up to homeomorphism) of connected 1-manifolds with boundary is: an open interval, a half-open interval, a closed interval, and a circle.

(iv) A closed disk D in $\mathbb{R}^2$ is a 2-manifold with boundary, with ∂D the boundary circle. More generally, a closed ball in $\mathbb{R}^n$ is an n-manifold with boundary.

(v) A Möbius strip M is a 2-manifold with boundary, with ∂M a circle.

(vi) A figure 8 or figure 6 is not a manifold with boundary, since in each case there is one point which does not have any neighborhood homeomorphic to an open set in $\mathbb{H}^1$ (or $\mathbb{H}^n$ for any n).

(vii) An open disk with a relatively open subset of the bounding circle added (i.e. a relatively open subset of a closed disk) is a manifold with boundary; more generally, an open subset of any manifold with boundary is a manifold with boundary. But an open disk with a subset of the bounding circle added which is not relatively open is not a manifold with boundary, although this is not entirely obvious.

The next result may seem obvious, but a rigorous proof is rather involved on one point.

XIII.20.2.5. THEOREM. Let X be an n-manifold with boundary. Then

(i) The interior and the boundary of X are disjoint. ∂X is closed in X, and the set of interior points is dense and open.

(ii) The interior of X is an n-manifold, and ∂X (if nonempty) is an $(n-1)$-manifold.

It is obvious that the set of interior points is dense and open, and that the interior is an n-manifold. (Under any coordinate chart, points going to a point of $\mathbb{H}^n \setminus \partial\mathbb{H}^n$ are interior points.) The issue in proving

that the interior and boundary are disjoint is to show that a point of $\partial\mathbb{H}^n$ does not have an open neighborhood in $\mathbb{H}^n$ homeomorphic to an open set in $\mathbb{R}^n$ (or in $\mathbb{R}^m$ for any m). This is nontrivial to prove directly, but is an immediate consequence of the Invariance of Domain Theorem (). The simplest direct proof, which uses some algebraic topology, is that a point of $\partial\mathbb{H}^n$ has a deleted neighborhood in $\mathbb{H}^n$ which is contractible, while no point of $\mathbb{R}^n$ has a contractible deleted neighborhood. It follows from the disjointness that ∂X is closed. Showing that ∂X is an $(n-1)$-manifold is straightforward, since the points of ∂X in a coordinate patch around a point of ∂X form a set homeomorphic to a relatively open set in $\partial\mathbb{H}^n$, which is homeomorphic to $\mathbb{R}^{n-1}$.

XIII.20.2.6. Convention: In geometry/topology, the term "manifold" is often used to mean "manifold with boundary" (with, of course, the possibility that the boundary is empty). In particular, the term "compact manifold" normally means "manifold with boundary which is compact" (i.e. the whole space is compact, not just the boundary). To refer to a manifold in the sense of XIII.20.1.7. which is compact, the term *closed manifold* (which has no real logical meaning) is normally used. The term *open manifold* is also sometimes used to mean a manifold which is noncompact and has empty boundary. The same conventions are used for smooth manifolds.

In some references, a manifold with empty boundary is called *unbounded*. However, this term seems much too fraught with potential misinterpretation to be appropriate, especially in an analysis book. I also have a very hard time calling a circle or sphere "unbounded."

In this book, we will strive for minimum ambiguity at the possible cost of excessive pedantry on these terms: we will always use the phrase "manifold with boundary" when a boundary is allowed, and we will avoid using "compact manifold", writing "compact manifold with boundary" if a boundary is allowed and "closed manifold" otherwise. (There is still some ambiguity: is a "closed submanifold" of a manifold a submanifold which is a closed subset of the manifold, or a submanifold which is compact? We will be explicit about which is meant when this occurs.) And we will not use the term "unbounded" in the context of manifolds, except when they are explicit subsets of Euclidean space, where the term "unbounded" has its usual meaning for subsets (i.e. not contained in a ball of finite radius). We will avoid the term "open manifold."

XIII.20.3. New Manifolds From Old

XIII.20.4. Classification of Topological Manifolds

Number of Manifolds

XIII.20.4.1. How many topologically distinct topological manifolds (with or without boundary) are there? Since according to our definition any discrete space is a manifold, the question is meaningless; we should restrict to connected manifolds, or at least to manifolds with only countably many components, i.e. separable manifolds.

Any separable manifold with boundary is a Polish space, and there are $2^{\aleph_0}$ topologically distinct Polish spaces (), so there are at most $2^{\aleph_0}$ topologically distinct separable manifolds with boundary. On the other hand, if $n > 1$ there are $2^{\aleph_0}$ topologically distinct connected open subsets of $\mathbb{R}^n$ (XIII.20.5.14.), each of which is a separable connected n-manifold, so up to homeomorphism there are exactly $2^{\aleph_0}$ separable manifolds, and exactly $2^{\aleph_0}$ separable manifolds with boundary. The same is true if "separable" is replaced by "connected."

There are $2^{2^{\aleph_0}}$ topologically distinct 2-dimensional connected Hausdorff locally Euclidean spaces, both with and without boundary (Exercise XIII.20.5.10.).

The question is more interesting and difficult in the compact case (a compact manifold with boundary is automatically separable). In fact:

XIII.20.4.2. THEOREM. There are only countably many homeomorphism classes of compact manifolds with boundary.

For each $n > 1$, there are infinitely many topologically distinct connected closed n-manifolds (). See [CK70] for a proof of XIII.20.4.2..

XIII.20.5. Exercises

XIII.20.5.1. (a) Show that an open set in a locally Euclidean space is locally Euclidean in the relative topology. Show that an open set in a topological manifold is a topological manifold in the relative topology.

(b) Show that a finite cartesian product of locally Euclidean spaces is locally Euclidean. Show that a finite cartesian product of topological manifolds is a topological manifold. What about products of topological manifolds with boundary? [Consider $\mathbb{H}^n \times \mathbb{H}^m$.]

XIII.20.5.2. Let X be a locally Euclidean space, and U and V open subsets of X. Suppose ϕ is a homeomorphism from U to V. Let X/ϕ be the topological space obtained from X by identifying U and V via ϕ, with the quotient topology.

(a) Show that X/ϕ is locally Euclidean.

(b) Show that the example of XIII.20.1.3.(v) is obtained this way by letting X be two disjoint copies of $\mathbb{R}$ and U and V the two copies of $\mathbb{R} \setminus \{0\}$.

(c) Under what conditions will X/ϕ be Hausdorff?

(d) Construct a circle from $\mathbb{R}$ by this method, and a Möbius strip from $\mathbb{R}^2$.

(e) Generalize to a family of open subsets $\{U_i : i \in I\}$, with specified homeomorphisms ϕ_i of U_i with a fixed space U_0.

XIII.20.5.3. Let X and Y be topological manifolds with boundary. Let ϕ be a homeomorphism from ∂X to ∂Y (existence of such a ϕ entails that X and Y have the same dimension, at least locally), and $X \cup_\phi Y$ the topological space obtained from the disjoint union of X and Y by identifying ∂X and ∂Y by ϕ. (We say X and Y are *glued* along the boundary via ϕ).

(a) Show that $X \cup_\phi Y$ is a topological manifold.

(b) What manifold results from gluing two disks along the boundary, which is a circle? What if one or both disks are replaced by a Möbius strip, whose boundary is also a circle?

(c) The boundary of a cylinder is two disjoint circles. What if two cylinders are glued together along the boundary circles? Does the resulting manifold depend on the homeomorphisms chosen?

(d) Show that if X and Y are smooth manifolds with boundary, and ϕ is a diffeomorphism, then $X \cup_\phi Y$ is a smooth manifold.

XIII.20.5.4. Show that $[0,1) \times (0,1)$, $[0,1) \times [0,1)$, $[0,1) \times [0,1]$, and a closed disk with a single boundary point removed are all homeomorphic to $\mathbb{H}^2$ and hence are topological 2-manifolds with boundary.

XIII.20.5.5. The Long Line. Denote by ω_1 the ordered set of ordinals less than the first uncountable ordinal (). Give $L_0 := \omega_1 \times [0,1)$ the lexicographic ordering, and the order topology (). L_0 is called the *long ray*.

(a) If $a, b \in L_0$, $a < b$, then the interval $[a,b]$ is homeomorphic to a closed bounded interval in $\mathbb{R}$. [Use the fact that any countable successor ordinal is order-isomorphic to a closed subset of $[0,1]$ ().] Hence the open interval (a,b) is homeomorphic to $\mathbb{R}$. Show that L_0 is order-complete and that it has no countable cofinal subset. Show that L_0 is countably compact and sequentially compact, and that every real-valued continuous function on L_0 is eventually constant.

(b) L_0 is not locally Euclidean since it has an endpoint (smallest point). If this is removed, a 1-dimensional locally Euclidean space L_+ is obtained which is called the *open long ray* or *half-long line*. L_+ is normal, even completely normal (and in particular Hausdorff), as is any space with the order topology (XIII.16.1.1.), and is connected since it is order-complete. L_+ has a countable coinitial set but no countable cofinal set.

(c) Show that the open cover $\{(a,b) : a, b \in L_+, a < b\}$ of L_0 has no locally finite refinement. Thus L_+ is not paracompact. Show directly that L_+ does not satisfy any of the other conditions of XIII.20.1.4.. L_+ is also not separable. L_+ is not countably compact or sequentially compact.

(d) Let $-L_0$ be a copy of L_0 with ordering reversed, and L be the union of $-L_0$ and L_0 with the endpoint of $-L_0$ identified with the initial point of L_0, with the order topology (the points of $-L_0$ smaller than the points of L_0). Then L is a 1-dimensional locally Euclidean space which is connected and normal, with neither a countable coinitial set nor a countable cofinal set, does not satisfy any of the conditions of XIII.20.1.4., and is not separable. L is called the *long line*. If $a, b \in L$, $a < b$, then the subset $[a,b]$ of L is homeomorphic to $[0,1]$, and (a,b) is homeomorphic to $\mathbb{R}$. Show that L is countably compact and sequentially compact, and that every real-valued continuous function on L is eventually constant in each direction.

Note that terminology in references is inconsistent: some authors call L_0 or L_+ the "long line."

(e) Show that L, L_+, and L_0 are not contractible.

(f) Show that if M is a connected 1-dimensional Hausdorff locally Euclidean space, then M is homeomorphic to exactly one of the following: $\mathbb{R}$, L, L_+, or a circle S^1.

XIII.20.5.6. Let L and L_0 be as in XIII.20.5.5.. Show that the Stone-Čech compactification () of L_0 coincides with the one-point compactification (). Show that the Stone-Čech compactification of L consists of adding two points to L, one at each "end."

XIII.20.5.7. [**?**, Appendix A] Consider the following Cartesian products of the spaces from XIII.20.5.5.(f):

$L \times L$, the *big plane.*

$L \times L_+$, the *big half-plane.*

$L_+ \times L_+$, the *big quadrant.*

$L \times \mathbb{R}$, the *wide plane* or *long strip.*

$L_+ \times \mathbb{R}$, the *half-long strip.*

$L \times S^1$, the *long cylinder.*

$L_+ \times S^1$, the *half-long cylinder.*

(a) Show that each is a connected 2-dimensional Hausdorff locally Euclidean space not satisfying the conditions of XIII.20.1.4..

(b) There are four obvious copies of L in $L \times L$, the "x-axis", the "y-axis", and the "diagonals." Show that the diagonals are embedded differently than the axes. [Show that any copy of ω_1 in $L \times L$ which hits the x-axis in uncountably many points has only countably many points off the x-axis, but this is false for the diagonals.] The same is true in $L_0 \times L_0$, where L_0 is the long ray (there is only one diagonal here). Use this to show that that no two of the above spaces are homeomorphic (it is hardest to distinguish between the big half-plane, the big quadrant, and the half-long strip; cf. XIII.20.5.4.). [For another approach, use (e).]

(c) Identify the initial circle of $L_0 \times S^1$ to a point. Show that a sequentially compact connected 2-dimensional Hausdorff locally Euclidean space is obtained not satisfying the conditions of XIII.20.1.4., sometimes called the *big disk*. Show that the big disk is not homeomorphic to any of the spaces of (a) (it is hardest to distinguish it from the big plane).

(d) Let E be the "octant" $\{(x, y) \in L_0 \times L_0 : y \leq x\}$. E has two edges, the x-axis A and the diagonal D. Let σ be a finite or infinite sequence of A's and D's. Let M_σ be the space obtained from the finite disjoint union of E's, the same number as the number of terms in σ, by taking the n'th term of σ as the first edge of the n'th copy of E and the other edge as the second edge, and identifying the first edge of the $(n+1)$'st copy of E with the second edge of the n'th copy. Then delete the first edge of the first copy of E and, if σ is finite, the second edge of the last copy. As a variation if σ is finite, identify the second edge of the last copy of E with the initial edge of the first copy to obtain a space M'_σ. All these spaces are connected 2-dimensional locally Euclidean Hausdorff spaces, and the M'_σ are sequentially compact. Show that all M_σ and M'_σ are mutually nonhomeomorphic, except that if σ is finite, σ can be replaced by its complementary sequence obtained by interchanging A's and D's and reversing order, and in the case of M'_σ, σ can be cyclically permuted. We have:

> The big plane is M'_σ, where $\sigma = (A, D, A, D, A, D, A, D)$ or (D, A, D, A, D, A, D, A).
>
> The big half-plane is M_σ, where $\sigma = (A, D, A, D)$.
>
> The big quadrant is M_σ, where $\sigma = (A, D)$.

The big disk can be regarded as M'_σ, where σ is the empty sequence. The "open octant" is the space M_σ, where $\sigma = (A)$ or (D). The M'_σ construction works even for infinite sequences, even (countably) transfinitely [Nyi84].

(e) Describe the Stone-Čech compactification of each of the above spaces.

XIII.20.5.8. The Prüfer Manifold. This is an example of a 2-dimensional connected locally Euclidean space not satisfying the conditions of XIII.20.1.4..

This space was first described by T. Radó in 1925 [], and credited to his teacher, H. Prüfer. We describe a version of the construction from [**?**]; see also [**?**, Appendix A]. There are actually two rather different spaces constructed this way; only the first, described in this problem, is called the "Prüfer manifold," and the second, described in Exercise XIII.20.5.9., is a variation of a space (called the *Moore manifold*) constructed by R. L. Moore []. See [Roy88, p. 208-209] for a description of Moore's construction. To add to the terminological confusion, all these spaces are examples of a class of spaces called *Moore spaces*. A variation of the Prüfer manifold is described in [CV98, p. 268-272].

Although these examples are customarily called "manifolds", they are not manifolds according to the definition in this book since they are not metrizable (in fact, not even normal).

Begin with an uncountable collection $\{\mathbb{R}^2_a : a \in \mathbb{R}\}$ of copies of the plane, and an additional copy $\mathbb{R}^2_+$ of the open upper half plane. Take the separated union U of the $\mathbb{R}^2_a$ and $\mathbb{R}^2_+$; then U is a 2-manifold. Identify the open upper half plane of each $\mathbb{R}^2_a$ with $\mathbb{R}^2_+$ by

$$(x, y) \in \mathbb{R}^2_a \ (y > 0) \iff (xy + a, y) \in \mathbb{R}^2_+$$

and let P be the quotient space. P is called the *Prüfer manifold*.

(a) Show that P is a connected 2-dimensional locally Euclidean space (cf. Exercise XIII.20.5.2.(a)).

(b) Show that P is Hausdorff (cf. Exercise XIII.20.5.2.(c)).

(c) Let Z be the union of the copies of the x-axis in $\mathbb{R}^2_a$ for all a. Show that in the relative topology Z is a separated union of uncountably many copies of $\mathbb{R}$: the x-axis of each $\mathbb{R}^2_a$ is clopen in Z. Thus Z and hence P is not second countable, so P does not satisfy the conditions of XIII.20.1.4..

(d) Let A be the subset of Z consisting of the points on the x-axis of $\mathbb{R}^2_a$ for a rational, and $B = Z \setminus A$ the set of points on the x-axis of $\mathbb{R}^2_a$ for a irrational. Then A and B are disjoint closed subsets of P. Use the Baire Category Theorem () to prove that A and B do not have disjoint neighborhoods in P. Thus P is not normal.

(e) Show that P is not separable. [Consider points in the open lower half-spaces of the $\mathbb{R}^2_a$.]

(f) Show that P is contractible (cf. [**?**, Appendix A]).

XIII.20.5.9. This problem is a continuation of XIII.20.5.8..

(a) Let P_0 be the union of Z and the upper half plane of P (i.e. the image of $\mathbb{R}^2_+$ in P). Then P_0 is a "2-dimensional locally Euclidean space with boundary." P_0 is connected and separable since $\mathbb{R}^2_+$ is dense in P_0, but $\partial P_0 = Z$ consists of a separated union of uncountably many copies of $\mathbb{R}$.

(b) Let Q be obtained by gluing two copies of P_0 along the boundary. Then Q is a 2-dimensional connected Hausdorff locally Euclidean space which is separable but does not satisfy the conditions of XIII.20.1.4..

(c) Q can alternately be constructed from P by identifying for each a the open lower half plane of $\mathbb{R}^2_a$ with a new copy $\mathbb{R}^2_-$ of the open lower half plane by

$$(x, y) \in \mathbb{R}^2_a \ (y < 0) \iff (xy + a, y) \in \mathbb{R}^2_-$$

and taking the quotient topology.

(d) The closed subsets A and B of Z in XIII.20.5.8.(d) also do not have disjoint neighborhoods in Q, so Q is not normal.

(e) Q is not simply connected: the fundamental group of Q is quite complicated (uncountable).

(f) Another example can be constructed from P_0 by identifying, for each a, the positive and negative x-axes of $\mathbb{R}^2_a$ by folding over at 0. A separable locally Euclidean space M is obtained. This is the *Moore manifold*. The Moore manifold is contractible (), hence simply connected.

XIII.20.5.10. If $A \subseteq \mathbb{R}$, modify the construction of P, P_0, Q, and M by attaching $\mathbb{R}^2_a$ if and only if $a \in A$. Call the resulting spaces P_A, P^0_A, Q_A, M_A.

(a) Show that if A is countable, P_A and M_A are homeomorphic to $\mathbb{R}^2$. If $card(A) = n$, $1 \leq n \leq \aleph_0$, Q_A is homeomorphic to a plane with a discrete set of points of cardinality $n - 1$ removed (where $\aleph_0 - 1 = \aleph_0$).

(b) [Bla] Show that if A is a subset of $\mathbb{R}$, there at most $2^{\aleph_0}$ subsets B of $\mathbb{R}$ such that the spaces P_A^0 and P_B^0 are homeomorphic. [Regard all the P_B^0 as subsets of P_0. A homeomorphism from P_A^0 to P_B^0 can be regarded as a continuous function from P_A^0 to P_0. But since P_A^0 is separable and P_0 has cardinality $2^{\aleph_0}$, there are only $2^{\aleph_0}$ continuous functions from P_A^0 to P_0. cf. XIII.19.5.45..] So there are $2^{2^{\aleph_0}}$ topologically distinct P_A^0. Show similarly that there are $2^{2^{\aleph_0}}$ topologically distinct M_A. The P_A^0 and M_A are all separable and contractible. There are also $2^{2^{\aleph_0}}$ topologically distinct P_A and Q_A. The M_A and P_A constructions can be combined into a unified family.

A family of $2^{2^{\aleph_0}}$ mutually nonhomeomorphic simply connected 2-dimensional Hausdorff locally Euclidean spaces, as well as a family of $2^{\aleph_1}$ mutually nonhomeomorphic sequentially compact simply connected 2-dimensional Hausdorff locally Euclidean spaces (the maximum number possible), is constructed in [Nyi84], which is a thorough study of nonmetrizable "manifolds." However, these examples are not separable or contractible.

XIII.20.5.11. (a) A compact Hausdorff locally Euclidean space is metrizable, hence a manifold (XIII.20.1.4.).

(b) If X is Hausdorff and locally Euclidean, it is locally compact, hence has a one-point compactification $X^\dagger$. If X is nonmetrizable, then $X^\dagger$ is not locally Euclidean at infinity, and hence the point at infinity is topologically different from all other points. (Actually $X^\dagger$ is rarely locally Euclidean at infinity even if X is metrizable, hence a manifold.)

(c) If X and Y are nonmetrizable Hausdorff locally Euclidean spaces and $X^\dagger$ and $Y^\dagger$ are homeomorphic, then X and Y are homeomorphic.

(d) For $A \subseteq \mathbb{R}$, let M_A be the space of XIII.20.5.10.. Then there are $2^{2^{\aleph_0}}$ topologically distinct spaces among the $(M_A)^\dagger$. Each of these is a separable connected (even path-connected) compact Hausdorff space.

XIII.20.5.12. ([Kne60]; cf. [Spi79, p. 643]) Construct a three-dimensional version of Q by taking a copy $\mathbb{R}^3_a$ for each $a \in \mathbb{R}$, and identifying $(x, y, z) \in \mathbb{R}^3_a$, $y \neq 0$, with the point in a fixed copy of $\mathbb{R}^3$ with coordinates $(a + yx, y, z + a)$ if $y > 0$ and $(a + yx, y, z - a)$ if $y < 0$.

(a) Show that the quotient Q^3 is a connected separable Hausdorff locally Euclidean space with local dimension 3.

(b) Show that the surfaces $z = c$ give a foliation () of Q^3 with only one leaf, which is two-dimensional. Thus there is a stronger topology on Q^3 under which it is a connected Hausdorff locally Euclidean space of local dimension 2 (not separable).

(c) For an arbitrary subset A of $\mathbb{R}$, similarly construct Q^3_A. Show there are $2^{2^{\aleph_0}}$ distinct homeomorphism classes. Describe Q^3_A if A is finite or countable. There is still a similar foliation of Q^3_A, but it has multiple leaves in general.

XIII.20.5.13. [Spi79, Problem A-8] Show that a connected Hausdorff locally Euclidean space which is not a single point has cardinality $2^{\aleph_0}$.

XIII.20.5.14. [Spi79, Problem 1-24] (a) Let X and Y be totally disconnected closed subsets of $\mathbb{R}^2$. Show that if $\mathbb{R}^2 \setminus X$ and $\mathbb{R}^2 \setminus Y$ are homeomorphic, then X and Y are homeomorphic. [See XIII.19.6.3..]

(b) Conclude from XIII.19.5.4. that there are $2^{\aleph_0}$ topologically distinct connected open subsets of $\mathbb{R}^2$. [Note that if A is any proper subset of the x-axis in $\mathbb{R}^2$, then $\mathbb{R}^2 \setminus A$ is path-connected.]

(c) There are also $2^{\aleph_0}$ topologically distinct connected relatively open subsets of the closed unit disk in $\mathbb{R}^2$, each containing the entire boundary circle.

(d) Show that these spaces remain topologically distinct even if the Cartesian product with $\mathbb{R}^m$ for any m is taken. Thus there are $2^{\aleph_0}$ topologically distinct connected open sets in $\mathbb{R}^n$ for any $n > 1$; in fact, see XIII.19.6.6.–XIII.19.6.8.(h). (There is only one nonempty connected open set in $\mathbb{R}$ up to homeomorphism, and only countably many open sets in $\mathbb{R}$ in all up to homeomorphism, since every open set is a countable disjoint union of open intervals ().)

XIII.20.5.15. [Hat02, 1.3.25] **An Interesting Non-Hausdorff Locally Euclidean Space.** Let T be the invertible linear transformation on $\mathbb{R}^2$ given by $T(x, y) = (2x, y/2)$. Define an equivalence relation on $E = \mathbb{R}^2 \setminus \{0\}$ by $\boldsymbol{v} \sim \boldsymbol{w}$ if $T^n \boldsymbol{v} = \boldsymbol{w}$ for some $n \in \mathbb{Z}$. Let M be the quotient of E by this equivalence relation, and $\pi : E \to M$ the quotient map.

(a) Show that each equivalence class is closed and discrete in E, and that M is T_1 and second countable.

(b) Show that M is the disjoint union of four cylinders C_1, C_2, C_3, C_4, each homeomorphic to $S^1 \times \mathbb{R}_+$, where C_j is the image of the open j'th quadrant in E, and four circles X_+, X_-, Y_+, Y_-, where X_+ is the image of the positive x-axis, X_- the image of the negative x-axis, and $Y_\pm$ the images of the positive and negative y-axis. $C_1 - C_4$ are open in M and $X_\pm, Y_\pm$ are closed in M. The "horizontal" circle in C_j at height $|r|$ is the image of the part of the hyperbola $xy = r$ in the j'th quadrant.

(c) The set $\{(x, y) : 1 \leq x < 2\}$ maps one-one onto $C_1 \cup X_+ \cup C_4$ under π, and the image in M is open and homeomorphic to $S^1 \times \mathbb{R}$. Similarly, $\{(x, y) : -2 \leq x < -1\}$ maps one-one onto $C_2 \cup X_- \cup C_3$, $\{(x, y) : 1 \leq y < 2\}$ maps one-one onto $C_1 \cup Y_+ \cup C_2$ and $\{(x, y) : -2 \leq y < -1\}$ maps one-one onto $C_3 \cup Y_- \cup C_4$. These cylinders are open in M. Conclude that M is locally Euclidean.

(d) The closure of C_1 is $C_1 \cup X_+ \cup Y_+$. Show that if $p \in X_+$ and $q \in Y_+$, then p and q do not have disjoint neighborhoods. Thus M is not Hausdorff. Similarly, points in $X_\pm$ and $Y_\pm$ do not have disjoint neighborhoods. (However, X_+ and X_- have disjoint neighborhoods, as do Y_+ and Y_-.) Describe the closure of $C_2 - C_4$.

(d) Let S be the image of a vertical line segment $\{(x, y) : x = x_0, 0 < y \leq y_0\}$. Show that the closure of S consists of S, one point of X_+, and all of Y_+. Find a sequence in S which converges to a point in X_+ and has every point of Y_+ as a cluster point, and a subsequence which converges to any arbitrarily given point of Y_+. Do the same for other vertical and horizontal line segments.

(e) Show that π is a covering map () from E to M.

(f) Let N be the space obtained from M by additionally identifying diametrically opposed points of E (N is the quotient of E by the action of the subgroup G of $SL_2(\mathbb{R})$ generated by T and $-I$, which is isomorphic to $\mathbb{Z} \times \mathbb{Z}_2$). Then N consists of two open cylinders C_+ (the image of C_1 and C_3) and C_- (the image of C_2 and C_4), and two closed circles X (the image of X_+ and X_-) and Y (the image of Y_+ and Y_-). $C_+ \cup X \cup C_-$ and $C_+ \cup Y \cup C_-$ are open subsets of N homeomorphic to $S^1 \times \mathbb{R}$; thus N is a cylinder $S^1 \times \mathbb{R}$ with the circle $S^1 \times \{0\}$ "doubled." If C_+ is identified with $S^1 \times \mathbb{R}_+$ by taking the images of the hyperbolas $xy = r$ as horizontal circles and images of rays from the origin as vertical lines (one only needs rays of slope between $1/2$ and 2), then points of X are obtained by approaching 0 along the image of a vertical ray in the right half-plane, i.e. a line spiraling around the cylinder at a certain rate, and points of Y are obtained by approaching along the image of a horizontal ray, i.e. a line spiraling around the cylinder at the same rate but in the opposite direction. The quotient map from E to N is also a covering map, i.e. G acts properly discontinuously on E ().

XIII.20.5.16. Show that a metrizable topological space X is a manifold with boundary if and only if every point of X has a closed neighborhood homeomorphic to a closed ball in a Euclidean space.

XIII.21. The Brouwer Fixed-Point Theorem

The Brouwer Fixed-Point Theorem, in its various equivalent formulations, is arguably the most important fact about the topology of Euclidean space (although the Heine-Borel Theorem () is arguably more fundamental), and indeed one of the most important results in all of topology. It essentially makes possible notions of "higher-order connectedness." Among other things, the entire subject of algebraic topology is at least implicitly based on this result. A coherent theory of dimension for topological spaces, and many applications in analysis, also depend on this theorem.

XIII.21.1. Statements of the Results

XIII.21.1.1. We use the following notation. Let B^n be the closed unit ball and S^{n-1} the closed unit sphere in $\mathbb{R}^n$, i.e.

$$B^n = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq 1\}$$
$$S^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\} \ .$$

In writing these, we implicitly assume $n \in \mathbb{N}$ (although they also make sense if $n = 0$: B^0 is a single point and $S^{-1} = \emptyset$). The term "map" will mean "continuous function."

XIII.21.1.2. Theorem. [Brouwer Fixed-Point Theorem] Every continuous function $f : B^n \to B^n$ has a fixed point, i.e. there is a point $\boldsymbol{x} \in B^n$ with $f(\boldsymbol{x}) = \boldsymbol{x}$.

XIII.21.1.3. Theorem. [No-Retraction Theorem] There is no retraction from B^n to S^{n-1}.

XIII.21.1.4. Theorem. [No-Contraction Theorem] The sphere S^{n-1} is not contractible, i.e. the identity map on S^{n-1} is not homotopic to a constant map.

XIII.21.1.5. We will show by elementary arguments that these three statements are equivalent. Thus to prove all the theorems it suffices to prove any one of them. We will also obtain some additional equivalent formulations of the theorem.

XIII.21.1.6. If $n = 1$, the statements are quite simple. $B^1 = [-1, 1]$ and $S^0 = \{-1, 1\}$. Thus the Brouwer Fixed-Point Theorem for $n = 1$ is (), essentially the Intermediate Value Theorem of calculus. The No-Retraction Theorem is also the Intermediate Value Theorem, or the fact that a connected space cannot be mapped continuously onto a disconnected space. The No-Contraction Theorem is obvious, since S^0 is not even connected (cf. ()).

If $n = 2$, the statements are also fairly elementary to prove; see XIII.21.2.. But if $n \geq 3$, all (known) proofs are fairly involved and/or use considerable machinery of some sort. There are many quite varied proofs known, and which one is the "best" is a quite subjective matter.

To prove that XIII.21.1.2.–XIII.21.1.4. are equivalent, it is convenient to show the negations of the statements are equivalent (although this seems a little perverse, since the ultimate result is that these statements are all false):

XIII.21.1.7. THEOREM. Let $n \in \mathbb{N}$. The following statements are equivalent:

(i) There is a continuous function $f : B^n \to B^n$ with no fixed point.

(ii) There is a retraction $r : B^n \to S^{n-1}$.

(iii) S^{n-1} is contractible.

XIII.21.1.8. To prove (ii) $\Rightarrow$ (iii), note that there is a contraction of B^n. The simplest one is given by the function $g(t, \boldsymbol{x}) = t\boldsymbol{x}$ for $0 \leq t \leq 1$, $\boldsymbol{x} \in B^n$. If there were a retraction $r : B^n \to S^{n-1}$, the function $h(t, \boldsymbol{x}) = r(t\boldsymbol{x})$ $(0 \leq t \leq 1,\ \boldsymbol{x} \in S^{n-1})$ would give a contraction of S^{n-1} (to the point $\boldsymbol{x}_0 = r(\mathbf{0})$). [Note that essentially the same argument shows that if X is a contractible space and r is a retraction of X onto a subspace Y, then Y is also contractible.]

XIII.21.1.9. For the converse (iii) $\Rightarrow$ (ii), suppose $h : [0, 1] \times S^{n-1} \to S^{n-1}$ is a contraction of S^{n-1} with $h(1, \boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in S^{n-1}$ and $h(0, \boldsymbol{x}) = \boldsymbol{x}_0$ for all $\boldsymbol{x} \in S^{n-1}$, where $\boldsymbol{x}_0$ is a fixed element of S^{n-1}. Define $r : B^n \to S^{n-1}$ by $r(\boldsymbol{x}) = h\left(\|\boldsymbol{x}\|, \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)$ for $\boldsymbol{x} \neq \mathbf{0}$ and $r(\mathbf{0}) = \boldsymbol{x}_0$. It is easily checked that r is continuous (only continuity at $\mathbf{0}$ is an issue), and is therefore a retraction from B^n to S^{n-1}.

Thus XIII.21.1.7. (ii) and (iii) are equivalent.

XIII.21.1.10. To prove (ii) $\Rightarrow$ (i), let r be a retraction of B^n to S^{n-1}. Then the map $f : B^n \to B^n$ defined by $f(\boldsymbol{x}) = -r(\boldsymbol{x})$ has no fixed point.

XIII.21.1.11. The proof of (i) $\Rightarrow$ (ii) is the hardest to write carefully, but the argument is geometrically simple and appealing. Suppose $f : B^n \to B^n$ has no fixed point. For each $\boldsymbol{x} \in B^n$, the line segment from $f(\boldsymbol{x})$ to $\boldsymbol{x}$ is nontrivial, and can be extended until it hits S^{n-1} in a point we call $r(\boldsymbol{x})$:

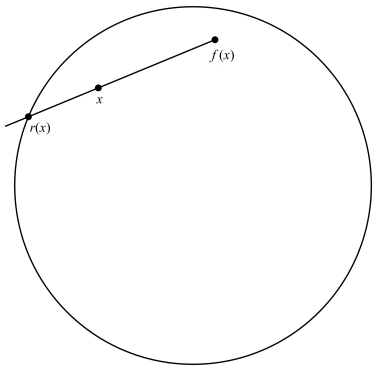


Figure XIII.69: Definition of the retraction

The function thus defined will be a retraction from B^n to S^{n-1}.

The only issue is whether r is continuous. This is geometrically plausible, and can be established carefully as follows. The vector $r(\boldsymbol{x})$ is of the form $f(\boldsymbol{x}) + t(\boldsymbol{x} - f(\boldsymbol{x}))$ for some $t \geq 1$. We can determine t explicitly using inner products and the quadratic formula:

$$\langle f(\boldsymbol{x}) + t(\boldsymbol{x} - f(\boldsymbol{x})), f(\boldsymbol{x}) + t(\boldsymbol{x} - f(\boldsymbol{x}))\rangle = 1$$

$$\langle f(\boldsymbol{x}), f(\boldsymbol{x})\rangle + 2t\langle \boldsymbol{x} - f(\boldsymbol{x}), f(\boldsymbol{x})\rangle + t^2\langle \boldsymbol{x} - f(\boldsymbol{x}), \boldsymbol{x} - f(\boldsymbol{x})\rangle - 1 = 0$$

$$\|\boldsymbol{x} - f(\boldsymbol{x})\|^2 t^2 - 2\langle f(\boldsymbol{x}) - \boldsymbol{x}, f(\boldsymbol{x})\rangle t + (\|f(\boldsymbol{x})\|^2 - 1) = 0$$

$$t = \frac{\langle f(\boldsymbol{x}) - \boldsymbol{x}, f(\boldsymbol{x})\rangle + \sqrt{\langle f(\boldsymbol{x}) - \boldsymbol{x}, f(\boldsymbol{x})\rangle^2 + \|\boldsymbol{x} - f(\boldsymbol{x})\|^2(1 - \|f(\boldsymbol{x})\|^2)}}{\|\boldsymbol{x} - f(\boldsymbol{x})\|^2}$$

which is a continuous function of $\boldsymbol{x}$ since the denominator is never zero.

This completes the proof of XIII.21.1.7..

Essential Maps

XIII.21.1.12. If X is a topological space of "dimension" less than n (we have not yet carefully defined what this means), and f is a map from X into $\mathbb{R}^n$ (or B^n), the range of f should be a set of dimension $< n$ and should thus have dense complement; so it should be possible to perturb f to a uniformly close function g such that the range of g does not include some specified point, e.g. the origin. Actually, the existence of such pathological functions as space-filling curves () casts doubt on the intuition behind this argument, but it turns out rather remarkably that the conclusion is correct.

Another argument leading to the conclusion is that a system of n equations in fewer than n variables "usually" does not have a solution. In special cases there might happen to be a solution, but it should be possible to perturb such a system to a "nearby" system with no solution. On the other hand, it is essentially the content of the Inverse Function Theorem () that if we have a "smooth" system of n equations in n variables which has a locally unique solution, then any "nearby" smooth system also has a (locally unique) solution.

XIII.21.1.13. Definition. Let X be a topological space. A map (continuous function) $f : X \to B^n$ is *inessential* if for every $\epsilon > 0$ there is a map $g : X \to B^n$ with $\|f - g\|_\infty = \sup\{\|f(x) - g(x)\| : x \in X\} < \epsilon$ and with $\mathbf{0} \notin g(X)$. A map f from X to B^n is *essential* if it is not inessential, i.e. every map sufficiently uniformly close to f has $\mathbf{0}$ in the range.

XIII.21.1.14. This use of the term *essential* is somewhat nonstandard, although it is used in a similar sense in [Pea75]. In topology, it is usually said that a map $f : X \to Y$ is essential if it is not homotopic to a constant map. Thus any map from a topological space to a contractible space such as B^n would be inessential. We will use the term in our sense only in relation to maps into B^n.

XIII.21.1.15. The definition of an essential map makes use of the standard metric on B^n. But the definition would be unchanged if any other metric giving the standard topology were used instead, since any two metrics giving the same compact topology are uniformly equivalent (). Thus the property of being essential or inessential is a purely topological property.

We then obtain another statement equivalent to the Brouwer Fixed-Point Theorem:

XIII.21.1.16. THEOREM. The identity map from B^n to B^n is essential.

XIII.21.1.17. To prove that this statement is equivalent to the No-Retraction Theorem, and thus to the Brouwer Fixed-Point Theorem, suppose that the identity map f on B^n is inessential, and fix $g : B^n \to B^n$ with $\|f - g\| < \frac{1}{2}$ and $g(\boldsymbol{x}) \neq \mathbf{0}$ for all $\boldsymbol{x} \in B^n$. We will gradually modify g on the annulus $\{\boldsymbol{x} \in B^n : \frac{1}{2} \leq \|\boldsymbol{x}\| \leq 1\}$ to a function $h : B^n \to B^n$ with $h(\boldsymbol{x}) \neq \mathbf{0}$ for all $\boldsymbol{x} \in B^n$ and $h(\boldsymbol{x}) = \boldsymbol{x}$ for $\boldsymbol{x} \in S^{n-1}$. We will take $h(\boldsymbol{x})$ to be a point on the line segment between $\boldsymbol{x}$ and $g(\boldsymbol{x})$, closer to $\boldsymbol{x}$ as $\|\boldsymbol{x}\|$ approaches 1. The precise formula is

$$h(\boldsymbol{x}) = \begin{cases} g(\boldsymbol{x}) & \text{if} \quad \|\boldsymbol{x}\| \leq \frac{1}{2} \\ (2\|\boldsymbol{x}\| - 1)\boldsymbol{x} + (2 - 2\|\boldsymbol{x}\|)g(\boldsymbol{x}) & \text{if} \quad \frac{1}{2} \leq \|\boldsymbol{x}\| \leq 1 \end{cases} .$$

It is clear that h is a continuous function from B^n to B^n and that $h(\boldsymbol{x}) = \boldsymbol{x}$ if $\|\boldsymbol{x}\| = 1$. To show that $h(\boldsymbol{x}) \neq \mathbf{0}$ for all $\boldsymbol{x}$, it suffices to note that if $\|\boldsymbol{x}\| \geq \frac{1}{2}$, then $\mathbf{0}$ is not on the line segment between $\boldsymbol{x}$ and $g(\boldsymbol{x})$ since $\|\boldsymbol{x} - g(\boldsymbol{x})\| < \frac{1}{2}$.

There is a retraction r from $B^n \setminus \{\mathbf{0}\}$ to S^{n-1}, e.g. radial projection; the map $r \circ h$ would then be a retraction from B^n to S^{n-1}. Thus by the No-Retraction Theorem, f cannot be inessential.

Conversely, suppose there is a retraction r from B^n to S^{n-1}. By scaling, for any $\epsilon > 0$ there is a retraction r_ϵ from $B^n_\epsilon = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq \epsilon\}$ onto $S^{n-1}_\epsilon = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = \epsilon\}$. Define $g_\epsilon : B^n \to B^n$ by

$$g_\epsilon(\boldsymbol{x}) = \begin{cases} r_\epsilon(\boldsymbol{x}) & \text{if} \quad \|\boldsymbol{x}\| \leq \epsilon \\ \boldsymbol{x} & \text{if} \quad \epsilon \leq \|\boldsymbol{x}\| \leq 1 \end{cases} .$$

Then $\|f - g_\epsilon\| \leq 2\epsilon$ and $g_\epsilon(\boldsymbol{x}) \neq \mathbf{0}$ for all $\boldsymbol{x} \in B^n$, so f is inessential. Thus XIII.21.1.16. implies the No-Retraction Theorem.

Partitions and the Lebesgue Covering Theorem

XIII.21.1.18. We now discuss two more equivalent versions of the Brouwer Fixed-Point Theorem. The statements are slightly more technical, and are phrased in terms of the n-cube

$$I^n = [-1, 1]^n = \{(x_1, \ldots, x_n) : -1 \leq x_k \leq 1\}$$

which is obviously homeomorphic to B^n. We will let A_k and B_k $(1 \leq k \leq n)$ be the opposite k-faces of I^n:

$$A_k = \{(x_1, \ldots, x_n) \in I^n : x_k = -1\}$$

$$B_k = \{(x_1, \ldots, x_n) \in I^n : x_k = 1\}$$

which are obviously disjoint closed subsets of I^n. The union of all the A_k and B_k is the boundary ∂I^n of I^n, which is homeomorphic to S^{n-1}.

Recall the definition of a partition (XIII.7.6.4.).

XIII.21.1.19. THEOREM. [FACE PARTITION THEOREM] For $1 \leq k \leq n$, let P_k be a partition between A_k and B_k in I^n. Then

$$P_1 \cap \cdots \cap P_n \neq \emptyset .$$

XIII.21.1.20. We show that the Face Partition Theorem for $n \in \mathbb{N}$ is equivalent to the No-Retraction Theorem for the same n, and hence to the Brouwer Fixed-Point Theorem. It is again convenient to prove that the negations of the statements are equivalent.

Suppose there is a retraction r from I^n to ∂I^n. Let $(r_1, \ldots, r_n)$ be the coordinate functions of r, and set

$$P_k = r_k^{-1}(\{0\})$$

$$U_k = r_k^{-1}([-1, 0))$$

$$V_k = r_k^{-1}((0, 1])$$

Then P_k is a partition between A_k and B_k for each k, and clearly $P_1 \cap \cdots \cap P_n = \emptyset$, so the statement of the Face Partition Theorem is false.

Conversely, let P_k be a partition between A_k and B_k for each k, with associated open sets U_k containing A_k and V_k containing B_k, such that $P_1 \cap \cdots \cap P_n = \emptyset$. For each k, define $f_k : I^n \to [-1, 1]$ by

$$f_k(\boldsymbol{x}) = \begin{cases} \frac{\rho(\boldsymbol{x}, P_k)}{\rho(\boldsymbol{x}, P_k) + \rho(\boldsymbol{x}, A_k)} & \text{if} \quad \boldsymbol{x} \in U_k \\ 0 & \text{if} \quad \boldsymbol{x} \in P_k \\ -\frac{\rho(\boldsymbol{x}, P_k)}{\rho(\boldsymbol{x}, P_k) + \rho(\boldsymbol{x}, B_k)} & \text{if} \quad \boldsymbol{x} \in V_k \end{cases}$$

where ρ is the standard metric on I^n. Then f_k is continuous. If $f : I^n \to I^n$ is the function with coordinate functions $(f_1, \ldots, f_n)$, then $f(\boldsymbol{x}) \neq \mathbf{0}$ for all $\boldsymbol{x} \in I^n$. Let r be a retraction of $I^n \setminus \{\mathbf{0}\}$ onto ∂I^n, and let $g = r \circ f$. It is easy to check that $g(A_k) \subseteq B_k$ and $g(B_k) \subseteq A_k$. It follows that g cannot have a fixed point, so the statement of the Brouwer Fixed-Point Theorem is false.

XIII.21.1.21. Theorem. [Lebesgue Covering Theorem] Let $\mathcal{F}$ be a finite collection of closed sets covering I^n, with the property that no $F \in \mathcal{F}$ meets both A_k and B_k for any k. Then there are distinct $F_1, \ldots, F_{n+1}$ in $\mathcal{F}$ with

$$F_1 \cap \cdots \cap F_{n+1} \neq \emptyset$$

(i.e. $\mathcal{F}$ has order $\geq n$ ()).

XIII.21.1.22. It is easy to show that the Lebesgue Covering Theorem implies the Brouwer Fixed-Point Theorem. Again we show the contrapositive. Suppose r is a retraction of I^n onto ∂I^n. Let

$$\mathcal{F} = \{r^{-1}(A_k), r^{-1}(B_k) : 1 \leq k \leq n\}$$

which is a finite cover of I^n by $2n$ sets satisfying the hypotheses of the Lebesgue Covering Theorem, but no $n + 1$ of the sets can have nonempty intersection.

To show the Face Partition Theorem, and hence the Brouwer Fixed-Point Theorem, implies the Lebesgue Covering Theorem is more complicated, and requires a lemma (cf. [Eng95, Lemma 1.2.9]):

XIII.21.1.23. LEMMA. Let X be a metrizable space, or more generally a completely normal space (), and E a closed subset of X. Let Y and Z be disjoint closed subsets of X. Let $\tilde{P}$ be a partition between $Y \cap E$ and $Z \cap E$ in E. Then there is a partition P between Y and Z in X with $P \cap E \subseteq \tilde{P}$.

PROOF: There are relatively open subsets $\tilde{U}$, $\tilde{V}$ of E such that $Y \cap E \subseteq \tilde{U}$, $Z \cap E \subseteq \tilde{V}$, $\tilde{U} \cap \tilde{V} = \emptyset$, and $E \setminus (\tilde{U} \cup \tilde{V}) = \tilde{P}$. The closure of $Z \cup \tilde{V}$ (in X) is $Z \cup [\tilde{V}]^-$, and $Y \cap [\tilde{V}]^- = (Y \cap E) \cap [\tilde{V}]^- = \emptyset$ since $[\tilde{V}]^- \subseteq E$ because E is closed, and $Y \cap E \subseteq \tilde{U}$, which is relatively open in E. Thus Y is disjoint from the closure of $Z \cup \tilde{V}$. We also have that $\tilde{U}$ is disjoint from both Z and $[\tilde{V}]^-$. So $Y \cup \tilde{U}$ is disjoint from the closure of $Z \cup \tilde{V}$. By a symmetric argument $Z \cup \tilde{V}$ is disjoint from the closure of $Y \cup \tilde{U}$. Thus by () there are disjoint open sets U and V in X with $Y \cup \tilde{U} \subseteq U$ and $Z \cup \tilde{V} \subseteq V$. Set $P = X \setminus (U \cup V)$. Then P is a partition between Y and Z in X, and $P \cap E$ is disjoint from both $\tilde{U}$ and $\tilde{V}$, i.e. $P \cap E \subseteq \tilde{P}$. 👍

XIII.21.1.24. We now show that the Face Partition Theorem for an $n \in \mathbb{N}$ implies the Lebesgue Covering Theorem for the same n (cf. [Eng95, Theorem 1.8.15]). Let $\mathcal{F}$ be a finite collection of closed subsets of I^n satisfying the hypotheses of the Lebesgue Covering Theorem. For $1 \leq k \leq n$ set

$$\mathcal{F}_k = \{F \in \mathcal{F} : A_k \cap F \neq \emptyset\}$$

$$\mathcal{E}_1 = \mathcal{F}_1, \quad \mathcal{E}_{k+1} = \mathcal{F}_{k+1} \setminus (\mathcal{E}_1 \cup \cdots \cup \mathcal{E}_k) \text{ for } 1 \leq k \leq n$$

$$E_k = \cup\{F : F \in \mathcal{E}_k\} \text{ for } 1 \leq k \leq n+1 \ .$$

Then the E_k are closed, and we have

$$A_k \subseteq I^n \setminus (E_{k+1} \cup \cdots \cup E_{n+1})$$

$$B_k \subseteq I^n \setminus E_k \ .$$

Set $\tilde{P}_0 = I^n$, and for $1 \leq k \leq n$ set

$$\tilde{P}_k = E_1 \cap \cdots \cap E_k \cap (E_{k+1} \cup \cdots \cup E_{n+1})$$

$$\tilde{U}_k = \tilde{P}_{k-1} \setminus (E_{k+1} \cup \cdots \cup E_{n+1})$$

$$\tilde{V}_k = \tilde{P}_{k-1} \setminus (E_1 \cap \cdots \cap E_k) \ .$$

Then, for $1 \leq k \leq n$, $\tilde{P}_k$ is closed, $\tilde{P}_k \subseteq \tilde{P}_{k-1}$, $A_k \cap \tilde{P}_{k-1} \subseteq \tilde{U}_k$, $B_k \cap \tilde{P}_{k-1} \subseteq \tilde{V}_k$, $\tilde{U}_k \cap \tilde{V}_k = \emptyset$, and

$$\tilde{P}_{k-1} \setminus (\tilde{U}_k \cup \tilde{V}_k) = \tilde{P}_{k-1} \cap \tilde{P}_k = \tilde{P}_k$$

so $\tilde{P}_k$ is a partition between $A_k \cap \tilde{P}_{k-1}$ and $B_k \cap \tilde{P}_{k-1}$ in $\tilde{P}_{k-1}$. Apply Lemma XIII.21.1.23. successively to obtain for each k a partition P_k between A and B in I^n with $P_k \cap \tilde{P}_{k-1} \subseteq \tilde{P}_k$. Thus by the Face Partition Theorem

$$\emptyset \neq P_1 \cap \cdots \cap P_n \subseteq \tilde{P}_n = E_1 \cap \cdots \cap E_{n+1}.$$

If $\boldsymbol{x} \in E_1 \cap \cdots \cap E_{n+1}$, then for $1 \leq k \leq n+1$ there is an $F_k \in \mathcal{E}_k$ with $\boldsymbol{x} \in F_k \subseteq E_k$; then $F_1 \cap \cdots \cap F_{n+1} \neq \emptyset$, and the F_k are all distinct since each $F \in \mathcal{F}$ is in exactly one $\mathcal{E}_k$.

XIII.21.1.25. The Face Partition Theorem and the Lebesgue Covering Theorem are versions of the statements $ind(\mathbb{R}^n) = n$ () and $dim(\mathbb{R}^n) = n$ () respectively; cf. [Eng95].

XIII.21.2. The Case $n = 2$

The Brouwer Fixed-Point Theorem for $n = 2$ is not as easy as the case $n = 1$ (XIII.21.1.6.), but is relatively simple.

XIII.21.2.1. The argument uses the map $p : \mathbb{R} \to S^1$ defined by

$$p(t) = (\cos 2\pi t, \sin 2\pi t) \ .$$

This map is surjective, but far from injective; however, if $s_0 \in \mathbb{R}$ and $\boldsymbol{z}_0 = p(s_0) \in S^1$, the restriction of p to the open interval $\left(s_0 - \frac{1}{2}, s_0 + \frac{1}{2}\right)$ is a homeomorphism from $\left(s_0 - \frac{1}{2}, s_0 + \frac{1}{2}\right)$ onto $S^1 \setminus \{-\boldsymbol{z}_0\}$. We will write $p_{s_0}^{-1}$ for the inverse map from $S^1 \setminus \{-\boldsymbol{z}_0\}$ onto $\left(s_0 - \frac{1}{2}, s_0 + \frac{1}{2}\right)$.

XIII.21.2.2. Lemma. [Path-Lifting Lemma] Let $\gamma : [0,1] \to S^1$ be a path (continuous function), with $\gamma(0) = \boldsymbol{z}_0$. If $s_0 \in \mathbb{R}$ is fixed with $p(s_0) = \boldsymbol{z}_0$, then there is a unique path $\tilde{\gamma} : [0,1] \to \mathbb{R}$ with $\tilde{\gamma}(0) = s_0$ and $\gamma = p \circ \tilde{\gamma}$.

Proof: A partition $\mathcal{P} = \{0 = t_0, t_1, \ldots, t_n = 1\}$ of $[0,1]$ will be called *sufficiently fine* if, for each k, $\|\gamma(t) - \gamma(s)\| < 2$ for all $t, s \in [t_{k-1}, t_k]$. There is a sufficiently fine partition since γ is uniformly continuous (). Suppose $\mathcal{P}$ is a sufficiently fine partition. There is a unique path $\beta : [0, t_1] \to \mathbb{R}$ with $\beta(0) = s_0$ and $\gamma|_{[0,t_1]} = p \circ \beta$, namely $\beta(t) = p_{s_0}^{-1}(\gamma(t))$ for $0 \le t \le t_1$ [to show uniqueness, note that the range of β must be contained in a connected subset of $p^{-1}(\gamma([0,t_1])) \subseteq p^{-1}(S^1 \setminus \{-\boldsymbol{z}_0\})$, so $\beta([0,t_1]) \subseteq \left(s_0 - \frac{1}{2}, s_0 + \frac{1}{2}\right)$.] In particular, $s_1 = \beta(t_1)$ is uniquely determined by γ and the choice of s_0. Define $\tilde{\gamma}(t)$ for $t \in [0,t_1]$ to be $\beta(t)$.

Now let $\boldsymbol{z}_1 = p(s_1) = \gamma(t_1)$, and repeat the construction to define $\tilde{\gamma}$ on $[t_1, t_2]$, and inductively extend to all of $[0,1]$. We then have $\tilde{\gamma}(0) = s_0$ and $\gamma = p \circ \tilde{\gamma}$.

To show uniqueness, suppose $\bar{\gamma}$ is another lift of γ, i.e. $\bar{\gamma}(0) = s_0$ and $\gamma = p \circ \bar{\gamma}$ (e.g. $\bar{\gamma}$ could be a lift defined by the same process using a different sufficiently fine partition). By the uniqueness shown above, we must have $\bar{\gamma} = \tilde{\gamma}$ on $[0,t_1]$, then on $[t_1,t_2]$, and ultimately on all of $[0,1]$. 👍

XIII.21.2.3. The lift $\tilde{\gamma}$ depends on the choice of s_0. But if the s_0 is changed, necessarily by addition of an integer n, then the lift is simply translated by n (the translation by n is obviously a lift with the right properties, so must be the unique one).

XIII.21.2.4. We will want a similar result for homotopies, i.e. maps from $[0,1]^2$ to S^1. If h is a map from $[0,1]^2$ to S^1, and $\boldsymbol{z}_0 = h(\mathbf{0})$, and $s_0 \in \mathbb{R}$ with $p(s_0) = \boldsymbol{z}_0$, then by XIII.21.2.2. there is a unique lifting $\tilde{h} : [0,1]^2 \to \mathbb{R}$ with $\tilde{h}(\mathbf{0}) = s_0$ which is continuous on each line segment in $[0,1]^2$ beginning at $\mathbf{0}$, defined for each $\boldsymbol{x} \in [0,1]^2$ by lifting the path $\gamma(t) = h(t\boldsymbol{x})$. However, it is not obvious that this map is continuous on $[0,1]^2$ (this is roughly the same question as whether the same value for $\tilde{h}(\boldsymbol{x})$ is obtained by lifting any other path in $[0,1]^2$ from $\mathbf{0}$ to $\boldsymbol{x}$). The lifted map $\tilde{h}$ is continuous in a neighborhood of $\mathbf{0}$ since all the lifts are done using $p_{s_0}^{-1}$, which is continuous; it is possible, but tricky, to expand successively through larger regions to the whole square (cf. ()).

We will take an alternate approach, first due to A. W. Tucker [Tuc46] (cf. [Spa81]).

XIII.21.2.5. There is a "multiplication" in S^1: if $\boldsymbol{z} = (\cos 2\pi t, \sin 2\pi t)$, write z symbolically as $e^{2\pi it}$ (this is actually more than symbolic; cf. ()). The t is not unique, but for a given $\boldsymbol{z}$ any two such t differ by an integer. If $\boldsymbol{w} = e^{2\pi is}$, define $\boldsymbol{zw} = e^{2\pi i(t+s)}$. Although t and s are only determined up to adding an integer, and thus $t+s$ is only determined up to an integer, the point $e^{2\pi i(t+s)}$ is well defined in S^1. This multiplication is associative and commutative. The point $\mathbf{1} = (1,0) = p(0)$ is an identity for this multiplication, and the point $\boldsymbol{z}^{-1} = e^{-2\pi it}$ is an inverse for $\boldsymbol{z} = e^{2\pi it}$. Multiplication is continuous as a function from $S^1 \times S^1$ to S^1, inversion (which geometrically is just reflection across the x-axis) is also continuous, and p has the property that $p(t+s) = p(t)p(s)$ for any $t, s \in \mathbb{R}$ (p is a *homomorphism*).

XIII.21.2.6. Now suppose $h : [0,1]^2 \to S^1$ is a continuous function, with $h(\mathbf{0}) = \boldsymbol{z}_0$, and $s_0 \in \mathbb{R}$ with $p(s_0) = \boldsymbol{z}_0$. By uniform continuity there is an $\epsilon > 0$ such that $\|h(\boldsymbol{x}) - h(\boldsymbol{y})\| < 2$ whenever $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^2$ and $\|\boldsymbol{x} - \boldsymbol{y}\| < \epsilon$. Fix $n \in \mathbb{N}$ with $n > \frac{\sqrt{2}}{\epsilon}$; then $\frac{\|\boldsymbol{x}\|}{n} < \epsilon$ for all $\boldsymbol{x} \in [0,1]^2$. For $\boldsymbol{x} \in [0,1]^2$ and $1 \le k \le n$, define

$$g_k(\boldsymbol{x}) = h\left(\frac{k}{n}\boldsymbol{x}\right) h\left(\frac{k-1}{n}\boldsymbol{x}\right)^{-1} .$$

Then g_k is a continuous function from $[0,1]^2$ to S^1, and

$$h(\boldsymbol{x}) = \boldsymbol{z}_0 g_1(\boldsymbol{x}) g_2(\boldsymbol{x}) \cdots g_n(\boldsymbol{x}) .$$

The range of g_k is contained in $S^1 \setminus \{-\mathbf{1}\}$: since $\left\|\frac{k}{n}\boldsymbol{x} - \frac{k-1}{n}\boldsymbol{x}\right\| < \epsilon$,

$$\left\| h\left(\frac{k}{n}\boldsymbol{x}\right) - h\left(\frac{k-1}{n}\boldsymbol{x}\right)\right\| = \left\| h\left(\frac{k}{n}\boldsymbol{x}\right) h\left(\frac{k-1}{n}\boldsymbol{x}\right)^{-1} - \mathbf{1}\right\| < 2 .$$

So there is a lift $\tilde{g}_k : [0,1]^2 \to \mathbb{R}$ defined by

$$\tilde{g}_k(x) = p_0^{-1}(g_k(\boldsymbol{x}))$$

which is continuous. Set

$$\tilde{h}(\boldsymbol{x}) = s_0 + \tilde{g}_1(\boldsymbol{x}) + \cdots + \tilde{g}_n(\boldsymbol{x}) .$$

then $\tilde{h}$ is a finite sum of continuous functions, hence continuous, and is a lift of h. By the uniqueness described in XIII.21.2.4., this is the unique continuous lift of h with $h(\mathbf{0}) = s_0$. Thus we have proved:

XIII.21.2.7. Lemma. [Homotopy Lifting Lemma] Let $h : [0,1]^2 \to S^1$ be a continuous function, with $h(\mathbf{0}) = \boldsymbol{z}_0$, and $s_0 \in \mathbb{R}$ with $p(s_0) = \boldsymbol{z}_0$. Then there is a unique continuous function $\tilde{h} : [0,1]^2 \to \mathbb{R}$ such that $\tilde{h}(\mathbf{0}) = s_0$ and $h = p \circ \tilde{h}$.

Note that the proof actually shows that the same is true for functions from $[0,1]^n$ to S^1 for any n.

XIII.21.2.8. A *loop* in S^1 is a path $\gamma : [0,1] \to S^1$ with $\gamma(1) = \gamma(0)$. A loop in S^1 is effectively the same thing as a continuous function from S^1 to S^1: the restriction q of p to $[0,1)$ is a bijection, and although the inverse map q^{-1} is not continuous, if γ is a loop in S^1 the map $f = \gamma \circ q^{-1}$ is a continuous function from S^1 to S^1, and conversely if $f : S^1 \to S^1$ is continuous, $\gamma = f \circ q$ (extended by setting $\gamma(1) = \gamma(0)$) is a loop in S^1.

XIII.21.2.9. If γ is a loop in S^1 with $\gamma(0) = \gamma(1) = \boldsymbol{z}_0$, and $s_0 \in \mathbb{R}$ with $p(s_0) = \boldsymbol{z}_0$, and $\tilde{\gamma}$ is the lift of γ to $\mathbb{R}$, then it is *not* generally true that $\tilde{\gamma}(1) = \tilde{\gamma}(0)$. But $\tilde{\gamma}(1) - \tilde{\gamma}(0)$ is an integer which does not depend on the choice of s_0 by XIII.21.2.3., called the *index* or *degree* of γ. Thus each continuous function from S^1 to S^1 has a well-defined index.

XIII.21.2.10. PROPOSITION. Homotopic maps from S^1 to S^1 have the same index.

PROOF: Let f_0 and f_1 be homotopic maps from S^1 to S^1 via maps f_s, and γ_s the loop corresponding to f_s. Then $h(t, s) = \gamma_s(t)$ is a continuous map from $[0, 1]^2$ to S^1. Take a lift $\tilde{h}$ of h to a continuous function from $[0, 1]^2$ to $\mathbb{R}$ (XIII.21.2.7.); then

$$\phi(s) = \tilde{h}(1, s) - \tilde{h}(0, s)$$

is the index of f_s, and is a continuous function of s. Since it takes only integer values, it must be constant, i.e. the index of f_0 equals the index of f_1. 👍

XIII.21.2.11. THEOREM. The circle S^1 is not contractible; thus the Brouwer Fixed-Point Theorem holds for $n = 2$.

PROOF: The identity map on S^1 has index 1, and a constant map has index 0, so they cannot be homotopic. 👍

XIII.21.2.12. The converse of XIII.21.2.10. is true (): maps from S^1 to S^1 with the same index are homotopic; so we get a parametrization of the homotopy classes of maps from S^1 to S^1. It is easy to get a loop with arbitrary index: set $\gamma_n(t) = p(nt)$. Thus the index map from homotopy classes of maps to $\mathbb{Z}$ is a bijection. This set of homotopy classes can be made into a group (the *fundamental group*) $\pi_1(S^1)$ via concatenation, and $\pi_1(S^1) \cong \mathbb{Z}$. See () for the general theory of fundamental groups.

XIII.21.2.13. The map $p : \mathbb{R} \to S^1$ is called the *universal covering map*, and $\mathbb{R}$ is the universal cover of S^1. See () for the general theory of covering maps.

XIII.21.3. The General Case

If $n \geq 3$, the Brouwer Fixed-Point Theorem for n is more difficult than for $n = 2$ since there is no map from $\mathbb{R}^{n-1}$ onto S^{n-1} analogous to the map p (S^{n-1} is simply connected ()). So one must proceed differently. One method of proof amounts to again defining a degree (index) which is an integer for every map from S^{n-1} to S^{n-1}, and proving that homotopic maps have the same degree, and that the identity map has degree 1 and a constant map has degree 0.

The proof must have a combinatorial component of some sort. The standard argument requires working with triangulations () and can be phrased in several ways which are all more or less equivalent:

(1) The most direct argument involves a combinatorial result called *Sperner's Lemma*. See e.g. (), (), [Spa81, p. 151], ... for expositions of this argument.

(2) One can develop abstract homology theory up through the Mayer-Vietoris sequence, and use this sequence to give a proof of the degree formula and therefore the No-Retraction Theorem by induction. See e.g. (), (), (), ... for this approach.

(3) The homology argument of (2) can be written out directly in the case of spheres without explicit use of the term "homology." See [Dug78] or () for this approach.

The Sperner's Lemma argument actually proves the following more general result (cf. [Spa81, p. 151]):

XIII.21.3.1. THEOREM. Let X be a compact manifold with boundary which can be triangulated (e.g. a smooth compact manifold with boundary). Then there is no retraction of X onto ∂X.

As a variation of this result, the following version follows immediately from the Brouwer Fixed-Point Theorem:

XIII.21.3.2. THEOREM. Let U be a bounded open set in $\mathbb{R}^n$. Then there is no retraction from $\bar{U}$ onto ∂U.

PROOF: Suppose r is a retraction of $\bar{U}$ onto ∂U. Fix $\boldsymbol{a} \in U$, and m so that $B = \bar{B}_m(\boldsymbol{a})$ contains $\bar{U}$. Let s be radial retraction of $B \setminus \{\boldsymbol{a}\}$ onto $\partial B \cong S^{n-1}$. Define, for $\boldsymbol{x} \in B$,

$$f(\boldsymbol{x}) = \begin{cases} s(r(\boldsymbol{x})) & \text{if} \quad \boldsymbol{x} \in \bar{U} \\ s(\boldsymbol{x}) & \text{if} \quad \boldsymbol{x} \in B \setminus \bar{U} \end{cases} .$$

It is easy to check that f is continuous, since $\partial(B \setminus \bar{U}) \subseteq \partial U \cup \partial B$. Then f is a retraction of B onto ∂B, contradicting the No-Retraction Theorem. 👍

XIII.21.3.3. This result can fail if U is not bounded, e.g. if U is the open upper half-plane in $\mathbb{R}^2$.

XIII.21.3.4. We will not give a direct proof of the Brouwer Fixed-Point Theorem for $n \geq 3$. We will deduce the theorem from some analysis results, including Sard's Theorem and the Implicit Function Theorem ().

XIII.21.4. Exercises

XIII.21.4.1. **Fundamental Theorem of Algebra.** Here is a winding-number proof of the Fundamental Theorem of Algebra (XII.1.4.2., XII.9.9.3., XII.9.9.4.); cf. [vM89]. Suppose

$$f(z) = a_0 + a_1 z + \cdots + a_n z^n$$

is a nonconstant polynomial of degree n (i.e. $a_n \neq 0$) with complex coefficients, with no roots in $\mathbb{C}$. Then $a_0 \neq 0$.

(a) Set

$$g(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n \ .$$

Show that $g(z) = z^n f\left(\frac{1}{z}\right)$ for $z \neq 0$ and $g(0) = a_n$. Thus g also has no roots in $\mathbb{C}$.
(b) Let r be radial retraction of $\mathbb{C} \setminus \{0\}$ onto S^1, and define $h : [0,1] \times S^1 \to S^1$ by

$$h(t,z) = r\left(\frac{f(tz)}{g\left(\frac{t}{z}\right)}\right) \ .$$

Show that h is a homotopy between a constant function and $\phi(z) = z^n$, a contradiction since $n > 0$.

XIII.21.5. Proof of the No-Retraction Theorem

In this section, we give two nice analysis proofs of the No-Retraction Theorem, which give a proof of the general Brouwer Fixed-Point Theorem. The first proof was found by MILNOR [Mil78] and simplified by ROGERS [Rog80]. The second proof is due to HIRSCH [Hir94].

We first reduce the problem to the smooth case. For the first proof, we will only need the following result (and not even all of it) for $\mathcal{C}^1$, but it is no harder to prove for $\mathcal{C}^\infty$, and the second proof needs the $\mathcal{C}^\infty$ result. We use the notation B^n for the closed unit ball in $\mathbb{R}^n$, S^{n-1} its boundary, the $(n-1)$-sphere, and $\mathbb{D}^n$ the open unit ball in $\mathbb{R}^n$.

XIII.21.5.1. LEMMA. If there is a (continuous) retraction of B^n onto S^{n-1}, then there is a $\mathcal{C}^\infty$ retraction. In fact, there is a C^∞ retraction which is radial projection on a neighborhood of S^{n-1}.

PROOF: Let s be a retraction. Since B^n is compact, s is uniformly continuous, so there is a δ, $0 < \delta < \frac{1}{4}$, such that $\|s(\boldsymbol{x}) - s(\boldsymbol{y})\| < \frac{1}{4}$ whenever $\boldsymbol{x}, \boldsymbol{y} \in B^n$ and $\|\boldsymbol{x} - \boldsymbol{y}\| < \delta$. Then if $\boldsymbol{x} \in B^n$ and $\|\boldsymbol{x}\| > 1 - \delta$, we have

$$\begin{aligned}\|s(\boldsymbol{x}) - \boldsymbol{x}\| &\le \left\|s(\boldsymbol{x}) - s\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right)\right\| + \left\|s\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right) - \boldsymbol{x}\right\| \\ &= \left\|s(\boldsymbol{x}) - \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right\| + \left\|\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \boldsymbol{x}\right\| < \frac{1}{4} + \delta < \frac{1}{2}\ .\end{aligned}$$

By the General Weierstrass Approximation Theorem (XX.10.3.3.), or (), there is a $\mathcal{C}^\infty$ function $f : B^n \to \mathbb{R}^n$ (in fact the coordinate functions can be polynomials) with $\|f - s\|_\infty < \frac{1}{4}$. Then, for $\boldsymbol{x} \in B^n$, $\|\boldsymbol{x}\| > 1 - \delta$, we have

$$\|f(\boldsymbol{x}) - \boldsymbol{x}\| \le \|f(\boldsymbol{x}) - s(\boldsymbol{x})\| + \|s(\boldsymbol{x}) - \boldsymbol{x}\| < \frac{1}{4} + \frac{1}{2} = \frac{3}{4}\ .$$

Let h be a $\mathcal{C}^\infty$ function from $[0,1]$ to $[0,1]$ which is identically 0 on $[0, 1-\delta]$ and identically 1 on $\left[1 - \frac{\delta}{2}, 1\right]$ (). For $\boldsymbol{x} \in B^n$, define

$$g(\boldsymbol{x}) = h(\|\boldsymbol{x}\|)\boldsymbol{x} + (1 - h(\|\boldsymbol{x}\|)f(\boldsymbol{x})\ .$$

Then g is $\mathcal{C}^\infty$, $g(\boldsymbol{x}) = \boldsymbol{x}$ if $\|x\| \ge 1 - \frac{\delta}{2}$, and $g(\boldsymbol{x}) = f(\boldsymbol{x})$ if $\|x\| \le 1 - \delta$.

If $\|x\| \le 1 - \delta$, we have $\|g(\boldsymbol{x}) - s(\boldsymbol{x})\| = \|f(\boldsymbol{x}) - s(\boldsymbol{x})\| < \frac{1}{4}$, and since $\|s(\boldsymbol{x})\| = 1$, we have $\|g(\boldsymbol{x})\| > \frac{3}{4}$, and in particular $g(\boldsymbol{x}) \ne \boldsymbol{0}$. And if $\|x\| > 1 - \delta$, then $g(\boldsymbol{x})$ lies on the line segment between $\boldsymbol{x}$ and $f(\boldsymbol{x})$; since

$$\|f(\boldsymbol{x}) - \boldsymbol{x}\| < \frac{3}{4} < 1 - \delta < \|\boldsymbol{x}\|$$

we again have that $g(\boldsymbol{x}) \ne \boldsymbol{0}$. Thus

$$r(\boldsymbol{x}) = \frac{g(\boldsymbol{x})}{\|g(\boldsymbol{x})\|}$$

is a $\mathcal{C}^\infty$ retraction of B^n onto S^{n-1} (note that $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|$ is $\mathcal{C}^\infty$ on $\mathbb{R}^n \setminus \{\boldsymbol{0}\}$). We have $r(\boldsymbol{x}) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ if $\|\boldsymbol{x}\| \ge 1 - \frac{\delta}{2}$. 👍

Thus the No-Retraction Theorem reduces to:

XIII.21.5.2. THEOREM. Let $n \in \mathbb{N}$. Then there is no $\mathcal{C}^\infty$ retraction from B^n to S^{n-1} which is radial projection on a neighborhood of S^{n-1}.

The First Proof

XIII.21.5.3. To prove the theorem, suppose r is a $\mathcal{C}^1$ retraction. Set $g(\boldsymbol{x}) = r(\boldsymbol{x}) - \boldsymbol{x}$ and, for $0 \leq t \leq 1$ set

$$f_t(\boldsymbol{x}) = (1-t)\boldsymbol{x} + tr(\boldsymbol{x}) = \boldsymbol{x} + tg(\boldsymbol{x})$$

for $\boldsymbol{x} \in B^n$. Then g and each f_t is C^1, $f_t(\boldsymbol{x}) \in B^n$ for all t and $\boldsymbol{x} \in B^n$, and $f_t(\boldsymbol{x}) = \boldsymbol{x}$ for all t and $\boldsymbol{x} \in S^{n-1}$.

Since B^n is compact, Dg is bounded on B^n, and hence there is a constant $K(>0)$ such that

$$\|g(\boldsymbol{x}) - g(\boldsymbol{y})\| \leq K\|\boldsymbol{x} - \boldsymbol{y}\|$$

for all $\boldsymbol{x}, \boldsymbol{y} \in B^n$. If $0 \leq t < \frac{1}{K}$ and $\boldsymbol{x}, \boldsymbol{y} \in B^n$ with $f_t(\boldsymbol{x}) = f_t(\boldsymbol{y})$, then

$$\boldsymbol{x} - \boldsymbol{y} = t(g(\boldsymbol{x}) - g(\boldsymbol{y}))$$

$$\|\boldsymbol{x} - \boldsymbol{y}\| = t\|g(\boldsymbol{x}) - g(\boldsymbol{y})\| \leq Kt\|\boldsymbol{x} - \boldsymbol{y}\|$$

and since $Kt < 1$, $\boldsymbol{x} = \boldsymbol{y}$. Thus f_t is one-to-one if $0 \leq t < \frac{1}{K}$.

We have $Df_t(\boldsymbol{x}) = I + tDg(\boldsymbol{x})$ for all t and $\boldsymbol{x}$. Since Dg is bounded on B^n, there is a constant $t_0 > 0$, which we may take to be $< \frac{1}{K}$, such that $Det(Df_t(\boldsymbol{x})) > 0$ for all t, $0 \leq t \leq t_0$, and all $\boldsymbol{x} \in B^n$.

By the Inverse Function Theorem, f_t is an injective open mapping on $\mathbb{D}^n$ for $0 \leq t \leq t_0$. Let $U_t = f_t(\mathbb{D}^n)$; then U_t is an open set in $\mathbb{R}^n$ contained in B^n, hence $U_t \subseteq \mathbb{D}^n$.

XIII.21.5.4. LEMMA. Let f be a continuous function from B^n to B^n. Suppose $f(S^{n-1}) \subseteq S^{n-1}$ and that $f|_{\mathbb{D}^n}$ is an open mapping (as a map to $\mathbb{R}^n$). Then f maps $\mathbb{D}^n$ onto $\mathbb{D}^n$ and B^n onto B^n.

PROOF: Since f is an open mapping on $\mathbb{D}^n$, we have $f(\mathbb{D}^n) \subseteq \mathbb{D}^n$. If $f(\mathbb{D}^n)$ is a proper subset of $\mathbb{D}^n$, there is a $\boldsymbol{y} \in \mathbb{D}^n$ which is in $\overline{f(\mathbb{D}^n)} \setminus f(\mathbb{D}^n)$. But $f(B^n)$ is a compact set containing $f(\mathbb{D}^n)$, so $\boldsymbol{y} \in f(B^n)$, and hence $\boldsymbol{y} = f(\boldsymbol{x})$ for some $\boldsymbol{x} \in B^n \setminus \mathbb{D}^n = S^{n-1}$. But $f(S^{n-1}) \subseteq S^{n-1}$, a contradiction. Thus $f(\mathbb{D}^n) = \mathbb{D}^n$. Then $f(B^n)$ is a closed set containing $\mathbb{D}^n$, so $f(B^n) = B^n$.

XIII.21.5.5. Continuing with the proof of the theorem, we conclude from the lemma that $U_t = \mathbb{D}^n$ and $f_t(B^n) = B^n$ for $0 \leq t \leq t_0$.

Now consider the function

$$\phi(t) = \int_{B^n} Det(Df_t(\boldsymbol{x}))\, d\boldsymbol{x} = \int_{B^n} Det(I + tDg(\boldsymbol{x}))\, d\boldsymbol{x}\ .$$

This $\phi(t)$ is a polynomial (of degree $\leq n$) in t. If $0 \leq t \leq t_0$, then f_t is a $\mathcal{C}^1$ diffeomorphism from B^n to itself, and thus by the change-of-variables theorem () we have $\phi(t) = Vol(B^n)$ for $0 \leq t \leq t_0$. Thus ϕ is constant on $[0, t_0]$, and therefore constant everywhere, and nonzero.

However, f_1 maps B^n into S^{n-1} and thus has rank $\leq n-1$ at every point, so $Det(Df_1(\boldsymbol{x}))$ is zero for all $\boldsymbol{x}$, and hence $\phi(1) = 0$, a contradiction.

This completes the proof of Theorem XIII.21.5.2..

The Second Proof

The second proof is much shorter (given XIII.21.5.1.), but uses Sard's Theorem, and thus is not as elementary as the first proof.

XIII.21.5.6. Let r be a C^∞ retraction from B^n to S^{n-1} which is radial projection on a neighborhood of S^{n-1}. By Sard's Theorem (the version in XIII.23.6.3. suffices), there is a regular value $\boldsymbol{y} \in S^{n-1}$ (in fact a dense set of regular values, but we only need one!) Let $X = r^{-1}(\{\boldsymbol{y}\})$. Then X consists of regular points of r, hence is a one-dimensional compact manifold with boundary consisting of a radial line segment in from $\boldsymbol{y}$ and possibly some additional points in $\mathbb{D}^n$ (cf. VIII.5.8.1.), and $X \cap \mathbb{D}^n$ is a manifold without boundary, i.e. $\partial X = \{\boldsymbol{y}\}$. But a compact 1-manifold with boundary is a finite separated union of copies of $[0,1]$ and S^1, and thus has an *even* number of boundary points, a contradiction.

This completes the proof of Theorem XIII.21.5.2.. 

XIII.21.6. The Fixed-Point Theorem for Cubes

From the Brouwer Fixed-Point Theorem we can deduce that a continuous function from any cube to itself has a fixed point.

XIII.21.6.1. DEFINITION. Let X be a topological space. Then X has the *fixed-point property* if every map (continuous function) ϕ from X to X has a fixed point (i.e. there is an $x \in X$ with $\phi(x) = x$).

This definition is mostly useful only for compact Hausdorff spaces; it is difficult for a space which is not a compact Hausdorff space to have the fixed-point property. It is rare even for compact metrizable spaces: for example, a two-point space or circle does not have the property.

XIII.21.6.2. Recall that a *cube* is a product $[0,1]^I$ of copies of $[0,1]$, where I is an index set. If I is a finite set with n elements, $[0,1]^I$ is homeomorphic to B^n, hence has the fixed-point property by the Brouwer Fixed-Point Theorem.

XIII.21.6.3. THEOREM. Every cube has the fixed-point property.

PROOF: Let $X = [0,1]^I$ be a cube. For each finite set $F \subseteq I$, let X_F be the subset of X for which all coordinates not in F are 0. Then $X_F \cong [0,1]^F$. The X_F form a directed set of subspaces of X whose union is dense.

Let $\phi : X \to X$ be a map. For each finite $F \subseteq I$, define $\phi_F : X_F \to X_F$ by $\phi_F(x) = \pi_F(\phi(x))$, where π_F is the projection from X to X_F (setting coordinates outside F to 0). Since X_F has the fixed-point property, ϕ_F has a fixed point; let x_F be a fixed point of ϕ_F for each finite F.

Then (x_F) is a net in X, indexed by finite subsets of I directed by inclusion. Since X is compact, this net has a convergent subnet $(x_j)_{j \in J}$, where each x_j is x_{F_j} for some finite $F_j \subseteq I$. Say $x_j \to x \in X$. Then $\phi(x_j) \to \phi(x)$ since ϕ is continuous.

Fix a coordinate $i \in I$. Then $i \in F_j$ for all sufficiently large j. Let $\pi_i : X \to [0,1]$ be the i'th coordinate function (projection onto the i'th coordinate). Then $\pi_i(x_j) \to \pi_i(x)$ and $\pi_i(\phi(x_j)) \to \pi_i(\phi(x))$. If j is large enough that $i \in F_j$, then $\pi_i(\phi(x_j)) = \pi_i(\phi_{F_j}(x_j))$ by definition of ϕ_{F_j}; thus $\pi_i(\phi_{F_j}(x_j)) \to \pi_i(\phi(x))$. But $\phi_{F_j}(x_j) = x_j$ for all j; thus $\pi_i(x_j) \to \pi_i(\phi(x))$. So $\pi_i(\phi(x)) = \pi_i(x)$. This is true for all $i \in I$, so $\phi(x) = x$, i.e. x is a fixed point of ϕ. 👍

XIII.21.6.4. The AC is needed at several points in this proof. But the full force of the AC is not needed for any particular I; if I is countable (Hilbert cube) only the Countable AC is needed.

XIII.21.6.5. There is a generalization, the Schauder Fixed-Point Theorem ().

XIII.22. Topological Dimension Theory

It has been a long and occasionally painful process to give a good definition of the "dimension" of a topological space. There are many reasonable definitions of dimension, and it is a remarkable and significant fact that they all give the same numerical value for "reasonable" topological spaces, including separable metrizable spaces. In this section, we describe some of the principal definitions and their properties. Although we give definitions applicable to all topological spaces, we will concentrate on their properties only in the separable metrizable case.

One obvious property any reasonable dimension theory must have is that the dimension of $\mathbb{R}^n$ should be n. This is true for all the definitions of dimension we will consider, but it is surprisingly hard to prove for most of them.

Some standard references for Dimension Theory are [Men28], [HW41], [Nag65], [Nag70], [AP73], [Pea75], [Eng95]. The first is mainly just of historical value today, but the others are worth reading for content, especially the last. Of course, newer references contain more complete results.

XIII.22.1. Essential Dimension

XIII.22.1.1. Definition. Let X be a topological space. The *essential dimension* $ed(X)$ of X is the largest n such that there is an essential map () from X to B^n. We say $ed(X) = \infty$ if there is an essential map from X to B^n for all n, and $ed(X) = 0$ if every map from X to $B^1 = [-1, 1]$ is inessential. Set $ed(\emptyset) = -1$.

The relevance of this definition is the following result, which is essentially equivalent to the Brouwer Fixed-Point Theorem:

XIII.22.1.2. Theorem. We have $ed(\mathbb{R}^n) = ed(B^n) = ed(S^n) = n$.

See XIII.22.1.5. for a more general result.

The next fact also justifies the form of the definition.

XIII.22.1.3. Proposition. Let X be a topological space. If there is an essential map from X to B^n, then there is an essential map from X to B^m for any $m < n$.

Proof: Let $f : X \to B^n$ be an essential map, with coordinate functions $(f_1, \ldots, f_n)$. Fix $m < n$, and define $\tilde{f} : X \to B^m$ by

$$\tilde{f}(x) = (f_1(x), \ldots, f_m(x)) \ .$$

Suppose $\tilde{f}$ is inessential. For $\epsilon > 0$, let $\tilde{g} : X \to B^m$ with $\|\tilde{g} - \tilde{f}\|_\infty < \epsilon$ and $\mathbf{0} \notin \tilde{g}(X)$. Let $(g_1, \ldots, g_m)$ be the coordinate functions of $\tilde{g}$. Define $g : X \to \mathbb{R}^n$ by

$$g(x) = (g_1(x), \ldots, g_m(x), f_{m+1}(x), \ldots, f_n(x)) \ .$$

Then $\|g - f\|_\infty < \epsilon$ and $\mathbf{0} \notin g(X)$. The function g does not necessarily map X into B^n, but we can retract the part that does not: define h by

$$h(x) = \begin{cases} g(x) & \text{if} \quad \|g(x)\| \leq 1 \\ \frac{g(x)}{\|g(x)\|} & \text{if} \quad \|g(x)\| > 1 \end{cases} \ .$$

Then h is a map from X to B^n, $\|h-f\|_\infty < \epsilon$, and $\mathbf{0} \notin h(X)$. Since $\epsilon > 0$ is arbitrary, this contradicts that f is essential. Thus $\tilde{f}$ is essential. 👍

XIII.22.1.4. PROPOSITION. Let X be a topological space, and $f : X \to B^n$ a map. Then f is essential if and only if there is an $\alpha > 0$ such that for every $g : X \to B^n$ sufficiently uniformly close to f, $g(X)$ contains the ball B^n_α of radius α around $\mathbf{0}$.

PROOF: It is obvious that any f with this property is essential. Conversely, suppose f is essential. Fix $0 < \epsilon < 1$ for which there is no h with $\|f-h\|_\infty < \epsilon$ and $0 \notin h(X)$. Let $g : X \to B^n$ with $\|g-f\| < \frac{\epsilon}{2}$, and suppose there is an $\boldsymbol{x}_0$ with $\|\boldsymbol{x}_0\| < \frac{\epsilon}{2}$ and $\boldsymbol{x}_0 \notin g(X)$. Define $h : B^n \to B^n$ by

$$h(\boldsymbol{x}) = \begin{cases} \boldsymbol{x} - \boldsymbol{x}_0 & \text{if} \quad \|\boldsymbol{x}-\boldsymbol{x}_0\| \leq 1 \\ \frac{\boldsymbol{x}-\boldsymbol{x}_0}{\|\boldsymbol{x}-\boldsymbol{x}_0\|} & \text{if} \quad \|\boldsymbol{x}-\boldsymbol{x}_0\| > 1 \end{cases} .$$

Then $h(\boldsymbol{x}_0) = \mathbf{0}$, and $\boldsymbol{x}_0$ is the only $\boldsymbol{x}$ with $h(\boldsymbol{x}) = 0$; and $\|h-\iota\|_\infty < \frac{\epsilon}{2}$, where ι is the identity map on B^n, so $\|h \circ g - f\|_\infty < \epsilon$ and $\mathbf{0} \notin h \circ g(X)$, a contradiction. 👍

We now prove a general result that includes XIII.22.1.2..

XIII.22.1.5. THEOREM. Let X be a smooth n-manifold with boundary. Then $ed(X) = n$.

PROOF: X contains a closed subset Y homeomorphic to B^n, and there is a retraction r from X onto Y (this requires the Tietze Extension Theorem () in general, but if $X = B^n$, $\mathbb{R}^n$, or S^n, such a Y and r can be given explicitly without resort to the Tietze Extension Theorem). Let f be a homeomorphism from Y onto B^n. Then f is an essential map by XIII.21.1.16., so $f \circ r$ is an essential map from X to B^n [if g approximates $f \circ r$, and ι is the embedding of Y into X, then $g \circ \iota$ approximates $f \circ r \circ \iota = f$] and $ed(X) \geq n$.

On the other hand, suppose $m > n$ and f is a map from X to B^m. To show that f is inessential, it suffices to show that the restriction of f to each component of X is inessential, since the components are clopen and f can be modified independently on each component. Thus we may assume X is connected, and in particular second countable. Then f can be arbitrarily closely approximated by a smooth function g, and by () the range of g has dense complement. Thus f is inessential by XIII.22.1.4. and $ed(X) < m$. 👍

XIII.22.1.6. The smooth structure was used only to prove $ed(X) \leq n$; the first half of the proof applies to any topological n-manifold with boundary (and in fact to any normal topological space X containing a copy of B_n) to give $ed(X) \geq n$. Note that the conclusion is purely topological, and the smooth structure is used only to give a simple way of approximating continuous functions by continuous functions whose range is "small." The result also holds for manifolds which do not have differentiable structures, but requires a different proof in this case (Exercise XIII.22.8.4.). An alternate proof that $ed(B^n) \leq n$ can be given using XIII.22.4.13. and Exercise XIII.22.8.3..

XIII.22.1.7. COROLLARY. Let X be a smooth n-manifold with boundary, and Y a smooth m-manifold with boundary. If $n \neq m$, then X and Y are not homeomorphic. In particular, if $n \neq m$, $\mathbb{R}^n$ is not homeomorphic to $\mathbb{R}^m$, or to any open subset of $\mathbb{R}^m$.

Extending Maps to Spheres

Essential dimension can be characterized also by an extension property for maps to spheres:

XIII.22.1.8. THEOREM. Let X be a completely regular space. Then $ed(X) \leq n$ if and only if, whenever Y is a closed subset of X and $f : Y \to S^n$ is continuous, then f extends to a continuous map $g : X \to S_n$.

PROOF: Suppose $ed(X) \leq n$, Y is a closed subspace of X, and $f : Y \to S^n$. Regard f as a map from Y to B^{n+1}. By the Tietze Extension Theorem (), f extends to a map h from X to B^{n+1}. Then h is inessential, so by an argument almost identical to the proof of (), there is a $\phi : B^{n+1} \to B^{n+1}$ which is the identity on a neighborhood of S^n and for which $\tilde{h} = \phi \circ h$ satisfies $\mathbf{0} \notin \tilde{h}(X)$. Then $h|_Y = \tilde{h}|_Y$, so $\tilde{h}$ is also an extension of f. If r is the radial retraction of $B^{n+1} \setminus \{\mathbf{0}\}$ onto S^n, then $g = r \circ \tilde{h}$ is an extension of f to $g : X \to S^n$.

Conversely, suppose X has the extension property, and suppose $f : X \to B^{n+1}$. Fix $\epsilon > 0$, and let $Y = f^{-1}(S^n_\epsilon)$ (where $S^n_\epsilon = \{\boldsymbol{x} \in B^{n+1} : \|\boldsymbol{x}\| = \epsilon\}$). Then Y is a closed subset of X, so $f|_Y$ extends to a map h from X to S^n_ϵ. Define $g : X \to B^{n+1}$ by

$$g(x) = \begin{cases} h(x) & \text{if} \quad \|f(x)\| \leq \epsilon \\ f(x) & \text{if} \quad \|f(x)\| \geq \epsilon \end{cases} .$$

Then g is continuous, $\|g - f\|_\infty \leq 2\epsilon$, and $\mathbf{0} \notin g(X)$. Thus f is inessential and $ed(X) \leq n$. 👍

A trick gives the following interesting and useful variant:

XIII.22.1.9. COROLLARY. Let X be a normal space, Y a closed subspace, and $f : Y \to S^n$ a map. If $ed(X \setminus Y) \leq n$, then f extends to a map from X to S^n.

PROOF: By (), f extends to a neighborhood U of Y in X. Since X is normal, there is a neighborhood V of Y in X with $\bar{V} \subseteq U$. Thus f extends to $\tilde{f} : \bar{V} \to S^n$. Then $Z = (X \setminus Y) \cap \bar{V}$ is a closed subspace of $X \setminus Y$, and $g = \tilde{f}|_Z$ is a map from Z to S^n. By XIII.22.1.8., g extends to $\tilde{g} : X \setminus Y \to S^n$. Then $\tilde{f}$ and $\tilde{g}$ define a map from X to S^n extending f. 👍

Essential Dimension of Subspaces

XIII.22.1.10. If X is a topological space and Y is a subspace of X, one would expect that $ed(Y) \leq ed(X)$. However, this turns out to be false in general, even if X is a compact Hausdorff space (). The problem is that a map from Y to B^n need not extend to X in general. It is true in general that $ed(Y) \leq ed(X)$ if X is separable and metrizable (). And if Y is closed, it is always true, and an immediate consequence of XIII.22.1.8.:

XIII.22.1.11. COROLLARY. Let X be a completely regular space, and Y a closed subspace of X. Then $ed(Y) \leq ed(X)$.

PROOF: Suppose $ed(X) = n$. Let Z be a closed subset of Y, and $f : Z \to S^n$ a map. Then Z is closed in X, so by XIII.22.1.8. f extends to a map $g : X \to S^n$. Then $g|_Y$ is an extension of f to Y. Thus $ed(Y) \leq n$ by the other direction of XIII.22.1.8.. 👍

XIII.22.1.12. PROPOSITION. Let X be a completely regular space. Then $ed(X) = ed(\beta X)$, where βX is the Stone-Čech compactification () of X.

PROOF: Let f be a map from X to B^n, and $\bar{f}$ its extension to βX. If $g : X \to B^n$, then $\|\bar{g} - \bar{f}\|_\infty = \|g - f\|_\infty$; thus if f is essential, $\bar{f}$ is also essential. On the other hand, if f is inessential, then for every $\epsilon > 0$ there is a $g : X \to B^n$ with $\|g - f\|_\infty < \epsilon$ and $g(X)$ bounded away from $\mathbf{0}$ (XIII.22.1.4.); thus $\mathbf{0} \notin \bar{g}(\beta X)$, so $\bar{f}$ is also inessential. 👍

Essential Dimension of Subspaces of $\mathbb{R}^n$

XIII.22.1.13. THEOREM. Let X be a subset of $\mathbb{R}^n$. Then $ed(X) \leq n$, and $ed(X) = n$ if and only if X has nonempty interior.

XIII.22.2. Invariance of Domain

We can now prove Brouwer's Invariance of Domain Theorem:

XIII.22.2.1. THEOREM. [INVARIANCE OF DOMAIN] Let U be a nonempty open subset of $\mathbb{R}^n$, and f a one-to-one continuous function from U to $\mathbb{R}^m$. Then

(i) $m \geq n$.

(ii) If $m = n$, then $f(V)$ is open in $\mathbb{R}^n$ for every open subset V of U (in particular, $f(U)$ is open in $\mathbb{R}^n$), i.e. f is an open mapping and a homeomorphism onto its range.

This will be a consequence of the following:

XIII.22.2.2. THEOREM. Let X be a subset of $\mathbb{R}^n$, and f a one-to-one continuous function from X onto a subset Y of $\mathbb{R}^n$. If $\boldsymbol{x}$ is an interior point of X, then $f(\boldsymbol{x})$ is an interior point of Y.

XIII.22.2.3. A point $\boldsymbol{x} \in X$ is an interior point of X if and only if it is an interior point of a compact subset of X: if $\boldsymbol{x}$ is an interior point of X, then $\boldsymbol{x}$ is an interior point of a sufficiently small closed ball around $\boldsymbol{x}$ which is contained in X, and the converse is obvious. Thus it suffices to prove XIII.22.2.2. for X compact. And in this case, the map f is automatically a homeomorphism.

Thus it suffices to prove that interior points of a compact subset X of $\mathbb{R}^n$ can be characterized topologically, which is the substance of the next result:

XIII.22.2.4. LEMMA. Let X be a compact subset of $\mathbb{R}^n$, and $\boldsymbol{x} \in X$. Then $\boldsymbol{x} \in \partial X$ if and only if, whenever U is a neighborhood of $\boldsymbol{x}$ in X, there is an open neighborhood V of $\boldsymbol{x}$ in X contained in U such that every map from $X \setminus V$ to S^{n-1} extends to a map from X to S^{n-1}.

PROOF: Suppose $\boldsymbol{x} \in \partial X$, and U is a neighborhood of $\boldsymbol{x}$ in X. Then there is an open ball W around $\boldsymbol{x}$ in $\mathbb{R}^n$ such that $\bar{W} \cap X \subseteq U$. Set $V = W \cap X$. Let f be a map from $X \setminus V$ to S^{n-1}. Then ∂W is homeomorphic to S^{n-1}, and $X \setminus V$ is compact and hence closed in $(X \setminus V) \cup \partial W$. The complement $Z = [(X \setminus V) \cup \partial W] \setminus (X \setminus V)$ is an open subset of ∂W, hence an $(n-1)$-manifold with a smooth structure, and thus $ed(Z) = n - 1$ by XIII.22.1.5. (only the "easy" part $ed(Z) \leq n - 1$ is needed). So by XIII.22.1.9. f extends to a map $\tilde{f}$ from $(X \setminus V) \cup \partial W$ to S^{n-1}. Since $\boldsymbol{x} \in \partial X$, there is an $\boldsymbol{x}_0 \in W$ with $\boldsymbol{x}_0 \notin X$, and there is a retraction r from $\bar{W} \setminus \{\boldsymbol{x}_0\}$ to ∂W. Define $g : X \to S^{n-1}$ by

$$g(\boldsymbol{y}) = \begin{cases} \tilde{f}(r(\boldsymbol{y})) & \text{if} \quad \boldsymbol{y} \in V \\ f(\boldsymbol{y}) & \text{if} \quad \boldsymbol{y} \in X \setminus V \end{cases} .$$

Then $g : X \to S^{n-1}$ is an extension of f to X.

Conversely, suppose $\boldsymbol{x}$ is an interior point of X, and ϵ is small enough that the closed ball B of radius ϵ around $\boldsymbol{x}$ is contained in X. Let U be the open ball of radius ϵ centered at $\boldsymbol{x}$; then $\partial U = \partial B$ is homeomorphic to S^{n-1}, and $X \setminus U$ can be radially projected onto ∂U. Call this map f. If V is any neighborhood of $\boldsymbol{x}$ contained in U, then f can be extended to a map from $X \setminus V$ to ∂W using radial projection from $\boldsymbol{x}$. But f cannot be extended to a map g from X to ∂U, since the restriction of g to B would be a retraction of B onto ∂B. 👍

XIII.22.3. Other Dimension Functions

For analysis purposes, essential dimension seems to be the most useful dimension function on spaces. However, it is not commonly used by topologists. There are three standard dimension functions used by topologists: small inductive dimension, large inductive dimension, and covering dimension.

Small Inductive Dimension

XIII.22.3.1. Small inductive dimension *ind* is the earliest and probably still the most-used dimension function on spaces. The definition is motivated by the idea that the boundary of a "nice" region of $\mathbb{R}^n$ should have dimension $n - 1$. The definition is (surprise!) an inductive one: we successively define what it means for a space to have dimension $\leq n$.

XIII.22.3.2. DEFINITION. Let X be a topological space. Then

(i) $ind(X) = -1$ if and only if $X = \emptyset$.

(ii) $ind(X) \leq n$ if and only if, whenever $x \in X$ and U is an open neighborhood of x in X, there is an open neighborhood V of x contained in U with $ind(\partial V) \leq n - 1$.

(iii) $ind(X) = n$ if $ind(X) \leq n$ and it is false that $ind(X) \leq n - 1$.

(iv) $ind(X) = \infty$ if for every n it is false that $ind(X) \leq n$.

XIII.22.3.3. Thus, for example, a topological space X satisfies $ind(X) = 0$ if and only if it is nonempty and, for each $x \in X$ and each open neighborhood U of x, there is an open neighborhood V of x contained in U with $\partial V = \emptyset$. But a subset of X has empty boundary if and only if it is clopen; so $ind(X) = 0$ if and only if for each $x \in X$ and each open neighborhood U of x, there is a clopen neighborhood V of x contained in U, i.e. if and only if X is nonempty and there is a base for the topology of X consisting of clopen sets. (Thus if X is T_0 and $ind(X) = 0$, X must be completely regular.)

XIII.22.3.4. PROPOSITION. Let X be a regular topological space. Then $ind(X) \leq n$ if and only if, whenever A is a closed set in X and $x \in X \setminus A$, there is a partition P between $\{x\}$ and A with $ind(P) \leq n-1$.

PROOF: There is an open neighborhood U of x with $\bar{U} \subseteq X \setminus A$. There is an open neighborhood V of x contained in U with $ind(\partial V) \leq n - 1$. Then $P = \partial V$ is a partition between $\{x\}$ and A. 👍

XIII.22.3.5. PROPOSITION. Let X be a topological n-manifold with boundary ($n \geq 1$). Then $1 \leq ind(X) \leq n$. In particular, if X is a topological 1-manifold with boundary, then $ind(X) = 1$. So $ind(\mathbb{R}) = ind([0, 1]) = ind(S^1) = 1$.

PROOF: By induction on n. If $n = 1$, a point of X has arbitrarily small neighborhoods whose boundary consists of either one or two points, hence of inductive dimension 0. Thus $ind(X) \leq 1$. But X is locally connected and thus $ind(X) = 0$ is false. And if X is an $(n+1)$-manifold with boundary, each point of X has arbitrarily small neighborhoods whose boundary is homeomorphic to either S^n or B^n, hence with inductive dimension $\leq n$; thus $ind(X) \leq n + 1$. Also, $ind(X) = 0$ is false for the same reason as before. 👍

XIII.22.3.6. In fact, if X is a topological n-manifold with boundary, then $ind(X) = n$, but it is much harder to prove equality, even if $X = B^n$ or $X = \mathbb{R}^n$. Actually, it is hard to prove that there is any topological space X with $ind(X) > 1$.

One simple but important property of small inductive dimension is monotonicity:

XIII.22.3.7. PROPOSITION. Let X be a topological space, and Y a subspace of X. Then $ind(Y) \leq ind(X)$.

PROOF: By induction on $n = ind(X)$. If $n = -1$ there is nothing to prove. Suppose the result is true whenever $ind(X) \leq n$. If $ind(X) \leq n + 1$ and $Y \subseteq X$, and $y \in Y$, let U be an open neighborhood of y in Y. Then there is an open neighborhood $\tilde{U}$ of y in X with $U = \tilde{U} \cap Y$. There is an open neighborhood $\tilde{V}$ of y in X with $\tilde{V} \subseteq \tilde{U}$ and $ind(\partial_X \tilde{V}) \leq n$. Set $V = \tilde{V} \cap Y$; then V is an open neighborhood of y in Y with $V \subseteq U$. We have $\partial_Y V \subseteq \partial_X \tilde{V}$ (), so by the inductive hypothesis $ind(\partial_Y V) \leq n$. Thus $ind(Y) \leq n + 1$. 👍

XIII.22.3.8. Small inductive dimension is a good notion for separable metrizable spaces, but turns out to have pathological properties outside this class and is regarded by topologists as of little use or interest for more general spaces.

Large Inductive Dimension

The small inductive dimension is strictly a local concept. There is a technical variation, called the *large inductive dimension* (Ind), which is more of a global condition.

XIII.22.3.9. DEFINITION. Let X be a topological space. Then

(i) $Ind(X) = -1$ if and only if $X = \emptyset$.

(ii) $Ind(X) \leq n$ if and only if, whenever Y is a closed subset of X and U is an open neighborhood of Y in X, there is an open neighborhood V of Y contained in U with $Ind(\partial V) \leq n - 1$.

(iii) $Ind(X) = n$ if $Ind(X) \leq n$ and it is false that $Ind(X) \leq n - 1$.

(iv) $Ind(X) = \infty$ if for every n it is false that $Ind(X) \leq n$.

XIII.22.3.10. PROPOSITION. Let X be a normal topological space. Then $Ind(X) \leq n$ if and only if, whenever A and B are disjoint closed sets in X, there is a partition P between A and B with $Ind(P) \leq n-1$.

The proof is nearly identical to the proof of XIII.22.3.4..

XIII.22.3.11. PROPOSITION. Let X be a T_1 topological space. Then $ind(X) \leq Ind(X)$.

PROOF: It is easily proved by induction that if $Ind(X) \leq n$, then $ind(X) \leq n$. 👍

XIII.22.3.12. THEOREM. If X and Y are metrizable, then

$$Ind(X \times Y) \leq Ind(X) + Ind(Y) \ .$$

Equality does not hold in general. For example, the space X of () satisfies $Ind(X) = 1$ and $Ind(X \times X) = 1$ also, since $X \times X$ is homeomorphic to X. (This space is separable metrizable, so $ind(X) = Ind(X) = dim(X)$.) There are compact metrizable spaces for which equality fails []. It is unknown whether equality can fail for ANR's; for well-behaved ANR's equality does hold [Bor36].

The inequality holds in somewhat greater generality, but does not even hold for all compact Hausdorff spaces: there are compact Hausdorff spaces X and Y with $Ind(X) = 1$, $Ind(Y) = 2$, and $Ind(X \times Y) = 4$ [Fil72]. Compare with XIII.22.3.40..

Sum and Decomposition Theorems

We now obtain some moderately complicated but crucial results about dimensions of unions, and decompositions into sets of smaller dimension.

It is clear that if X is a union of two subspaces of dimension n, then the dimension of X might be larger than n: for example, $\mathbb{R}$ can be decomposed into the rational and irrational numbers, each of which has dimension 0. There is only an inequality in general (). In fact, any sufficiently nice space of dimension n can be written as a union of $n + 1$ subspaces of dimension 0 (). But if the space is nice and the subspaces are closed, the dimension of the union does not grow, even with countably many subspaces:

XIII.22.3.13. THEOREM. [SUM THEOREM FOR IND] Let X be a separable metrizable space. If $X = \cup_{j=1}^{\infty} X_j$, with each X_j closed in X and $ind(X_j) \leq n$ for all j, then $ind(X) = Ind(X) \leq n$.

The name "Sum Theorem," which has become standard, comes from the practice in older references on set theory and topology to use *sum* as a synonym for *union*. Similarly, *product* was used to mean *intersection*.

To prove this theorem, we first prove a zero-dimensional version for Ind which is valid in much greater generality:

XIII.22.3.14. THEOREM. Let X be a normal space. If $X = \cup_{j=1}^{\infty} X_j$, with each X_j closed in X and $Ind(X_j) = 0$ for all j, then $Ind(X) = 0$.

Before proving XIII.22.3.13., we need two lemmas, the first of which includes a useful construction:

XIII.22.3.15. LEMMA. Let X be a topological space and $\mathcal{V}$ a base for the topology of X (consisting of open sets). Set

$$Y = \bigcup_{V \in \mathcal{V}} \partial V$$

and $Z = X \setminus Y$. Then $ind(Z) \leq 0$.

PROOF: Let $z \in Z$ and U an open neighborhood of z in Z. Then there is an open set W in X with $U = W \cap Z$. There is a $V \in \mathcal{V}$ with $z \in V$ and $V \subseteq U$. Since $\partial V \subseteq Y$, we have that $V \cap Z = \bar{V} \cap Z$, so $V \cap Z$ is relatively clopen in Z and $z \in V \cap Z \subseteq U$. 👍

The second lemma is a variant of XIII.21.1.23. (cf. [Eng95, 1.2.9]).

XIII.22.3.16. LEMMA. Let X be a metrizable space, or more generally a completely normal space (), and Z a subspace of Z with $Ind(Z) = 0$. If A and B are disjoint closed subsets of X, then there are open sets V_1, V_2 in X with $A \subseteq V_1$, $B \subseteq V_2$, $V_1 \cap V_2 = \emptyset$, and

$$[X \setminus (V_1 \cup V_2)] \cap Z = \emptyset \ .$$

PROOF: Let V_1' and V_2' be open neighborhoods of A and B whose closures are disjoint. Let U_1 be a relatively clopen subset of Z containing $\bar{V}_1' \cap Z$ (all closures are closures in X) and disjoint from $\bar{V}_2 \cap Z$, and $U_2 = Z \setminus U_1$. Note that V_1' is disjoint from U_2 and hence from $\bar{U}_2$ since V_1' is open; thus $A \cap \bar{U}_2 = \emptyset$. Similarly, $B \cap \bar{U}_1 = \emptyset$.

Since $\bar{U}_1 \cap Z = U_1$ and $\bar{U}_2 \cap Z = U_2$, the sets U_1 and U_2 are completely separated in X, and hence the sets $A \cup U_1$ and $B \cup U_2$ are also completely separated in X. Thus they have disjoint neighborhoods V_1 and V_2 in X. Since $Z \subseteq V_1 \cup V_2$, $[X \setminus (V_1 \cup V_2)] \cap Z = \emptyset$. 👍

We now give the proof of XIII.22.3.13..

PROOF: By induction on n. The case $n = 0$ follows immediately from XIII.22.3.14. and the fact that if Y is a T_0 Lindelöf space with $ind(Y) = 0$, then $Ind(Y) = 0$ (XIII.13.6.5. and the proof of XIII.13.6.7.((iv) $\Rightarrow$ (iii))). Now suppose $n > 0$ and the statement holds for $n-1$ (including the statement that $ind(Y) = Ind(Y)$

if $ind(Y) \leq n-1$). There is a countable base $\mathcal{V}$ for the topology of X consisting of open sets V with $ind(\partial V) \leq n-1$. Set

$$Y = \bigcup_{V \in \mathcal{V}} \partial V$$

and $Z = X \setminus Y$. Then $ind(Z) \leq 0$ by Lemma XIII.22.3.15., so $Ind(Z) \leq 0$ since Z is Lindelöf. We have that $ind(Y) \leq n-1$ by the inductive assumption. Let A and B be disjoint closed subsets of X. By Lemma XIII.22.3.16. there are disjoint neighborhoods V_1 and V_2 of A and B with $Z \subseteq V_1 \cup V_2$. Hence $P = X \setminus (V_1 \cup V_2)$ is a partition between A and B. Since $P \subseteq Y$, $ind(P) \leq n-1$, so $Ind(P) \leq n-1$ by the inductive assumption. Thus $Ind(X) \leq n$ by XIII.22.3.10.. So we have for this n that if each X_k satisfies $ind(X_k) \leq n$, then $Ind(X) \leq n$. In particular, if $ind(X) = n$, set $X_k = X$ for all k to conclude that $Ind(X) \leq n$. Since $ind(X) \leq Ind(X)$ (XIII.22.3.11.), we have $Ind(X) = n$. 👍

We now obtain some corollaries of XIII.22.3.13..

XIII.22.3.17. Corollary. Let X be a separable metrizable space. Then $ind(X) = Ind(X)$.

XIII.22.3.18. Corollary. If X is a countable regular space, then $Ind(X) = ind(X) = dim(X) = 0$.

Proof: A countable topological space is Lindelöf, so a countable regular space is normal. Each singleton set $\{x\}$ has $Ind(\{x\}) = 0$, thus $Ind(X) = 0$. We then have $ind(X) = 0$, so $dim(X) = 0$ by XIII.13.6.7.((iii) $\Rightarrow$ (iv)). 👍

The regularity assumption is necessary: there exist connected countable Hausdorff spaces ().

XIII.22.3.19. Corollary. [First Decomposition Theorem] Let X be a separable metrizable space, and $n \geq 0$. If $ind(X) = n$, then X can be written as a disjoint union $Y \cup Z$, where $ind(Y) = n-1$ and $ind(Z) = 0$. Conversely, if X can be so written, then $n-1 \leq ind(X) \leq n$.

Proof: Since $ind(X) = n$, there is an $x \in X$ and an open neighborhood U of X such that no open neighborhood V of x contained in U has $ind(\partial V) \leq n-2$. There is a countable base V for X with $ind(\partial V) \leq n-1$ for all $V \in \mathcal{V}$. Set

$$Y = \bigcup_{V \in \mathcal{V}} \partial V$$

and $Z = X \setminus Y$. Then $ind(Z) \leq 0$ by Lemma XIII.22.3.15.. We have $ind(Y) \leq n-1$ by XIII.22.3.13.. But if $V \in \mathcal{V}$ with $x \in V \subseteq U$, then $ind(\partial V) = n-1$, so $ind(Y) = n-1$ by XIII.22.3.7.. Thus $Z \neq \emptyset$, i.e. $ind(Z) = 0$.

For the converse, suppose $X = Y \cup Z$, where $ind(Y) = n-1$ and $ind(Z) = 0$. Then $ind(X) \geq n-1$ by XIII.22.3.7.. Let $x \in X$ and U an open neighborhood of x. Then by Lemma XIII.22.3.16. there are disjoint neighborhoods V_1 of x and V_2 of $X \setminus U$ such that $Z \subseteq V_1 \cup V_2$. Then $\partial V_1 \subseteq Y$, so $ind(\partial V_1) \leq n-1$; and $V_1 \subseteq U$. Thus $ind(X) \leq n$. 👍

Applying this result repeatedly, we obtain:

XIII.22.3.20. Corollary. [Second Decomposition Theorem] Let X be a separable metrizable space. Then $ind(X) \leq n$ if and only if X can be written as a disjoint union of $n+1$ subspaces $X_1, \ldots, X_{n+1}$ with $ind(X_k) \leq 0$ for all k. If $ind(X) = n$, then X can be so written with $ind(X_k) = 0$ for all k.

XIII.22.3.21. Large Inductive Dimension Ind, unlike ind, is not monotone: if Y is a subspace of X, it is not necessarily true that $Ind(Y) \leq Ind(X)$ (). But it is true if Y is closed in X: the proof is an easy induction similar to the proof of XIII.22.3.7..

Covering Dimension

This is perhaps the most generally useful notion of dimension. The idea is due to Lebesgue, but the actual definition was given by Čech.

XIII.22.3.22. The idea behind covering dimension is that $\mathbb{R}$ can be covered with small open intervals in such a way that only consecutive intervals overlap, i.e. no three of the intervals have nonempty intersection; however, since it is connected, it cannot be covered by pairwise disjoint bounded open intervals. Similarly, $\mathbb{R}^2$ can be covered by small open sets in such a way that no four of the sets have nonempty intersection (e.g. tile the plane with hexagons, or consider a brick wall, and slightly expand the tiles or bricks to open sets), but it appears that in any covering by bounded open sets at least three of the sets must have a nonempty intersection (although this turns out to be not so easy to prove). It is a little harder, but possible, to cover $\mathbb{R}^3$ with small open sets in such a way that no five of the sets have nonempty intersection (try it!), but in any open cover by bounded sets at least four of the sets must have nonempty intersection.

XIII.22.3.23. Definition. Let X be a topological space, and $\mathcal{U} = \{U_i : i \in I\}$ an open cover of X. Then $\mathcal{U}$ has *order* $\leq n$ if, whenever $i_1, \ldots, i_{n+2}$ are distinct elements of I, then

$$U_{i_1} \cap \cdots \cap U_{i_{n+2}} = \emptyset$$

(i.e. every point of X is contained in at most $n+1$ of the sets). $\mathcal{U}$ has *order* n if it has order $\leq n$ but not order $\leq n-1$, i.e. there are distinct $i_1, \ldots, i_{n+1}$ in I such that

$$U_{i_1} \cap \cdots \cap U_{i_{n+1}} \neq \emptyset \ .$$

XIII.22.3.24. Definition. Let X be a nonempty topological space. The *covering dimension* $dim(X)$ of X is $\leq n$ if every finite open cover of X has a refinement of order $\leq n$. Say $dim(X) = n$ if $dim(X) \leq n$ and it is false that $dim(X) \leq n-1$. If $dim(X) \leq n$ is false for all n, then $dim(X) = \infty$. Set $dim(\emptyset) = -1$.

XIII.22.3.25. Most, but not all, authors use $dim(X)$ to denote covering dimension, as we will; the notation $cov(X)$ is also sometimes used. But in some references such as [HW41], $dim(X)$ means something else ($ind(X)$ in [HW41]); some references on geometric measure theory such as [Edg98] use $dim(X)$ to denote the Hausdorff dimension of a metric space X (a metric, but not a topological, invariant ()).

XIII.22.3.26. An open cover $\mathcal{V}$ of a topological space X is order zero if and only if X is nonempty and $\mathcal{V}$ is a partition of X into clopen sets. Thus $dim(X) = 0$ if and only if X is nonempty and every finite open cover of X has a refinement consisting of a partition of X into clopen sets.

XIII.22.3.27. Proposition. Let X be a topological space. If $dim(X) = 0$, then $Ind(X) = 0$. If X is normal and $Ind(X) = 0$, then $dim(X) = 0$.

Proof: Suppose $dim(X) = 0$, Y is a closed subspace of X, and U is an open neighborhood of Y. The open cover $\{U, Y^c\}$ of X has a refinement $\{W_i\}$ consisting of clopen sets. Let V be the union of the W_i contained in U; then V is open. The complement V^c is a union of W_i's contained in Y^c, hence also open. So V is clopen, $V \subseteq U$, and $V^c \subseteq Y^c$, so $Y \subseteq V \subseteq U$. Thus $Ind(X) = 0$.

Conversely, suppose X is normal and $Ind(X) = 0$. Let $\mathcal{U} = \{U_1, \ldots, U_n\}$ be a finite open cover of X. By XIII.7.6.8., $\mathcal{U}$ has a closed shrinking $\{E_1, \ldots, E_n\}$. Since $Ind(X) = 0$, for each k there is a clopen set W_k with $E_k \subseteq W_k \subseteq U_k$. Thus $\{W_1, \ldots, W_n\}$ is an open cover of X. Inductively define

$$V_k = W_k \setminus \bigcup_{j<k} W_j \ .$$

Then the V_k are pairwise disjoint clopen sets with $V_k \subseteq U_k$ for each k. If $p \in X$, let k be the first index for which $p \in W_k$; then $p \in V_k$. So $\{V_1, \ldots, V_n\}$ is a partition of X into clopen sets which refines $\mathcal{U}$. Thus $dim(X) = 0$. 👍

The normality hypothesis can be eliminated using XIII.22.3.33. (XIII.22.3.34.), i.e. for any topological space X, $dim(X) = 0$ if and only if $Ind(X) = 0$. A refinement of the second half of the argument (cf. [Pea75,]) gives a partial inductive characterization of dim similar to the definition of Ind:

XIII.22.3.28. Theorem. Let X be a normal topological space, and $n \in \mathbb{N} \cup \{0\}$. Then $dim(X) \leq n$ if, whenever Y is a closed subspace of X and U an open neighborhood of Y, there is an open neighborhood V of Y contained in U with $dim(\partial V) \leq n - 1$.

An inductive argument then shows:

XIII.22.3.29. Corollary. Let X be a normal topological space. Then $dim(X) \leq Ind(X)$.

XIII.22.3.30. Note that the result XIII.22.3.28. is "if," not "if and only if" (it is "if and only if" if $n = 0$). We can have $dim(X) < Ind(X)$, even for compact Hausdorff spaces (). If X is metrizable, then $dim(X) = Ind(X)$ (). There is no relation between $dim(X)$ and $ind(X)$ for general normal spaces (), but see XIII.22.4.4.; and it follows from XIII.22.3.27. and XIII.22.3.11. that if X is T_1 and $dim(X) = 0$, then $ind(X) = 0$.

XIII.22.3.31. Covering dimension can be alternately be defined using open shrinkings (XIII.7.6.7.), a special kind of refinement. A refinement $\mathcal{V}$ of an open cover $\mathcal{U} = \{U_i : i \in I\}$ can be converted to an open shrinking using consolidation. A *consolidation* of $\mathcal{V}$ under $\mathcal{U}$ consists of, for each $V \in \mathcal{V}$, choosing an $i \in I$ with $V \subseteq U_i$ and calling this i $k(V)$. Then for each i set $W_i = \cup\{V \in \mathcal{V} : k(V) = i\}$. Then W_i is an open set contained in U_i for each i, and $\mathcal{W} = \{W_i : i \in I\}$ is an open cover of X since each $V \in \mathcal{V}$ is contained in some W_i. Thus $\mathcal{W}$ is an open shrinking of $\mathcal{U}$. If $\mathcal{U}$ is a finite open cover, then $\mathcal{W}$ is also a finite cover.

A consolidation of $\mathcal{V}$ under $\mathcal{U}$ is highly nonunique in general, and requires the AC to define.

It is easy to check that if $\mathcal{V}$ has order $\leq n$, then any consolidation of $\mathcal{V}$ under another open cover $\mathcal{U}$ also has order $\leq n$. Thus we obtain:

XIII.22.3.32. Proposition. Let X be a topological space, and $n \geq 0$. The following are equivalent:

(i) $dim(X) \leq n$.

(ii) Every finite open cover of X has a finite refinement of order $\leq n$.

(iii) Every finite open cover of X has an open shrinking of order $\leq n$.

The following sharpening is useful:

XIII.22.3.33. Theorem. Let X be a topological space, and $n \geq 0$. Then $dim(X) \leq n$ if and only if, whenever $\mathcal{U} = \{U_1, \ldots, U_{n+2}\}$ is an open cover of X with $n+2$ elements, there is an open shrinking of $\mathcal{U}$ of order $\leq n$.

Proof: If $dim(X) \leq n$, the condition trivially follows. Conversely, suppose $dim(X) > n$, and let $\mathcal{V} = \{V_1, \ldots, V_m\}$ be a finite open cover of X with no refinement of order $\leq n$ (so $m \geq n+2$). If $i_1, \ldots, i_k$ are distinct natural numbers $\leq m$, with $V_{i_1} \cap \cdots \cap V_{i_k} \neq \emptyset$, and $\mathcal{W} = \{W_1, \ldots, W_m\}$ is an open shrinking of $\mathcal{V}$ with $W_{i_1} \cap \cdots \cap W_{i_k} = \emptyset$, replace $\mathcal{V}$ by $\mathcal{W}$. Doing this process successively for each such $i_1, \ldots, i_k$ (there are only finitely many such sets), we may assume $\mathcal{V}$ has the property that whenever $i_1, \ldots, i_k \leq m$ and $V_{i_1} \cap \cdots \cap V_{i_k} \neq \emptyset$, and $\mathcal{W} = \{W_1, \ldots, W_m\}$ is an open shrinking of $\mathcal{V}$, then $W_{i_1} \cap \cdots \cap W_{i_k} \neq \emptyset$.

The reduction does not change the fact that the order of $\mathcal{V}$ is at least $n+1$, so reordering $\mathcal{V}$ if necessary we may assume $\cap_{j=1}^{n+2} V_j \neq \emptyset$. Now set $U_j = V_j$ if $1 \leq j \leq n+1$ and $U_{n+2} = \cup_{j=n+2}^{m} V_j$. If $\mathcal{W} = \{W_1, \ldots, W_{n+2}\}$ is an open shrinking of $\mathcal{U}$, then

$$\{W_1, \ldots, W_{n+1}, W_{n+2} \cap V_{n+2}, \ldots, W_{n+2} \cap V_m\}$$

is an open shrinking of $\mathcal{V}$, and hence

$$\emptyset \neq \left(\cap_{j=1}^{n+1} W_j\right) \cap (W_{n+2} \cap V_{n+2}) \subseteq \cap_{j=1}^{n+2} W_j \ .$$



An application removes the normality hypothesis from XIII.22.3.27.:

XIII.22.3.34. PROPOSITION. Let X be a topological space. If $Ind(X) = 0$, then $dim(X) = 0$. Thus $Ind(X) = 0$ if and only if $dim(X) = 0$.

PROOF: If $dim(X) = 0$, then $Ind(X) = 0$ by XIII.22.3.27.. Conversely, suppose $Ind(X) = 0$. By XIII.22.3.33., to show that $dim(X) = 0$ it suffices to show that if $\{U, V\}$ is an open cover of X, there is a refinement $\{W, W^c\}$ of $\{U, V\}$ of order 0, i.e. a clopen set W with $V^c \subseteq W \subseteq U$. But this follows immediately from the definition that $Ind(X) = 0$ since V^c is a closed set and U is an open neighborhood of V^c. 👍

There is a locally compact Hausdorff space X with $ind(X) = 0$ but $dim(X) > 0$ and $Ind(X) > 0$ (XIII.11.14.8.).

Here is another simple consequence:

XIII.22.3.35. COROLLARY. Let X be a normal space. If $card(X) < 2^{\aleph_0}$, then $dim(X) = 0$.

PROOF: Let $\{U, V\}$ be an open cover of X. Then U^c and V^c are disjoint closed sets in X, so by Urysohn's Lemma there is a continuous function $f : X \to [0, 1]$ with $f = 0$ on U^c and $f = 1$ on V^c. Then $f(X)$ is a subset of $[0, 1]$ of cardinality $< 2^{\aleph_0}$, so cannot contain any interval; thus there is an $r \in (0, 1) \setminus f(X)$. The sets $Z = \{x \in X : f(x) > r\}$ and $W = \{x \in X : f(x) < r\}$ form a partition of X into clopen sets, with $Z \subseteq U$ and $W \subseteq V$. Thus $dim(X) = 0$. 👍

XIII.22.3.36. PROPOSITION. Let X be a normal space, and $n \geq 0$. If, for every pair A, B of disjoint closed sets of X there is a partition P between A and B with $dim(P) \leq n - 1$, then $dim(X) \leq n$. 3.1.27

XIII.22.3.37. Why finite open covers, and not general open covers? Mainly, the theory just turns out to work well this way. But note that any open cover of finite order is locally finite, and there are spaces, even normal spaces, with open covers which have no locally finite (even no point-finite) refinements (XIII.11.14.12.). But for paracompact spaces, there is no difference:

XIII.22.3.38. THEOREM. [DOWKER'S THEOREM] Let X be a topological space with $dim(X) \leq n$, and $\mathcal{U}$ a locally finite open cover of X. Then $\mathcal{U}$ has a refinement of order $\leq n$.

XIII.22.3.39. COROLLARY. Let X be a paracompact space. The following are equivalent:

(i) $dim(X) \leq n$.

(ii) Every finite open cover of X has a shrinking of order $\leq n$.

(iii) Every open cover of X has a refinement of order $\leq n$.

(iv) Every open cover of X has an open shrinking of order $\leq n$.

XIII.22.3.40. THEOREM. [Hem46] Let X and Y be compact Hausdorff spaces. Then $dim(X \times Y) \leq dim(X) + dim(Y)$.

Compare with XIII.22.3.12..

XIII.22.3.41. Covering dimension is not well behaved for spaces which are not normal. There is a technical variation which works somewhat better for completely regular spaces (dimension theory is of almost no use or interest for spaces which are not at least completely regular). Recall that a set U in a topological space X is a *cozero set* if there is a continuous function $f : X \to \mathbb{R}$ with $U = \{x \in X : f(x) \neq 0\}$. A cozero set is an open F_σ; if X is normal, the converse is also true. Some references use the term *functionally open* for cozero sets.

XIII.22.3.42. DEFINITION. Let X be a topological space. Then $zdim(X) \leq n$ if every finite open cover of X by cozero sets has a finite refinement by cozero sets which is of order $\leq n$.

XIII.22.3.43. There is no obvious relation between $dim(X)$ and $zdim(X)$ for general (even completely regular) spaces X. The two can differ: the spaces of XIII.11.14.7. and XIII.11.14.8. have $zdim$ 0, but $dim > 0$ since they are not normal. It is possible that $zdim(X) \leq dim(X)$ for all completely regular X, although this is not obvious (since $zdim(X) = dim(\beta X)$ for every completely regular X, this is equivalent to the question whether $dim(\beta X) \leq dim(X)$ for every completely regular X). But every open cover of a normal space has a shrinking consisting of cozero sets (Exercise XIII.22.8.2.); hence if X is normal, $zdim(X) = dim(X)$. Some authors define the covering dimension of a space X to be $zdim(X)$, and $zdim$ is written ∂im in some references.

Nuclear Dimension

There is an interesting characterization of covering dimension for spaces, which has an important generalization to C*-algebras (cf. [**?**]). This characterization was implicit in [HW41] (it can be easily proved from XIII.22.3.20.), but was apparently not explicitly described until [KW04]. In honor of the authors who described it and the applications to the classification of C*-algebras, we call this *nuclear dimension*. This notion has close connections to the Besicovitch Covering Theorem ().

XIII.22.3.44. DEFINITION. Let X be a topological space. An open cover $\mathcal{U}$ of X is *n-decomposable* if $\mathcal{U}$ is a disjoint union $\mathcal{U} = \mathcal{U}_1 \cup \cdots \cup \mathcal{U}_{n+1}$ of $n+1$ subcollections, each of which consists of pairwise disjoint sets.

XIII.22.3.45. DEFINITION. A topological space X has *nuclear dimension* $\leq n$ if every finite open cover of X has an n-decomposable refinement. The *nuclear dimension* $dim_{nuc}(X)$ is the smallest n such that X has nuclear dimension $\leq n$ ($dim_{nuc}(X) = \infty$ if there is no such n).

XIII.22.3.46. It is obvious that an n-decomposable open cover has order $\leq n$. The converse is false: consider a cover of a circle by three arcs. (A cover is 0-decomposable if and only if it has order zero, i.e. consists of a partition into clopen subsets.) Thus $dim(X) \leq dim_{nuc}(X)$ for any X.

However, we do have:

XIII.22.3.47. PROPOSITION. Let X be a normal topological space, and $0 \leq n < \infty$. Then every open cover of X of order n has an n-decomposable refinement.

PROOF: Let $\mathcal{V}$ be an open cover of X of order n. Since $\mathcal{V}$ is locally finite, there is a partition of unity $\{f_i\}$ subordinate to $\mathcal{V}$ (), which defines a continuous function ϕ from X to $|K|$ for a simplicial complex K of dimension $\leq n$ (). Let K' be the barycentric subdivision of K. Then each vertex σ of K' is the barycenter of a unique simplex Δ_σ of K; let $d(\sigma) \in \{0, \ldots, n\}$ be the dimension of Δ_σ.

Since $|K'|$ can be identified with $|K|$, ϕ may be regarded as a map from X to $|K'|$. For each vertex σ of K', let U_σ be the preimage under ϕ of the star of σ, the points of $|K'|$ whose σ-coordinate is positive. Then $\mathcal{U} = \{U_\sigma\}$ is an open cover of X which refines $\mathcal{V}$. For each k, $0 \leq k \leq n$, let

$$\mathcal{U}_k = \{U_\sigma : d(\sigma) = k\} \ .$$

Then $\mathcal{U}$ is the disjoint union of the $\mathcal{U}_k$. If σ, τ are vertices of K' with $d(\sigma) = d(\tau) = k$, then σ and τ are the barycenters of distinct k-simplexes in K, so the stars of σ and τ are disjoint, and thus U_σ and U_τ are disjoint. So $\mathcal{U}$ is n-decomposable. If $\mathcal{V}$ is finite, K (hence K') is finite, so $\mathcal{U}$ is also finite. 👍

XIII.22.3.48. COROLLARY. Let X be a normal topological space. Then $dim(X) = dim_{nuc}(X)$.

XIII.22.3.49. Actually, $dim_{nuc}(X) = zdim(X)$ for any completely regular space ().

Partition Dimension

In analogy with the Lebesgue Covering Theorem (), we make another definition:

XIII.22.3.50. DEFINITION. Let X be a nonempty topological space. The *partition dimension* $pd(X)$ is the smallest $n \geq 0$ such that, whenever $(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})$ are $n+1$ pairs of disjoint closed sets in X, there are partitions P_k of (A_k, B_k) for $1 \leq k \leq n+1$ with $P_1 \cap \cdots \cap P_{n+1} = \emptyset$. If there is no such n, set $pd(X) = \infty$. Set $pd(\emptyset) = -1$.

XIII.22.3.51. If X is a topological space, then $pd(X) \leq n$ if and only if, whenever $(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})$ are $n+1$ pairs of disjoint closed sets in X, there are partitions P_k of (A_k, B_k) for $1 \leq k \leq n+1$ with $P_1 \cap \cdots \cap P_{n+1} = \emptyset$. If X is normal and this condition holds for one n, it holds for all larger n (and if the condition holds for some n, then X is normal). The Lebesgue Covering Theorem implies that $pd(B_n) \geq n$.

Decomposition Dimension

We define one more notion of dimension:

XIII.22.3.52. DEFINITION. Let X be a nonempty topological space. The *decomposition dimension* $dd(X)$ is the smallest n such that X can be written

$$X = X_1 \cup \cdots \cup X_{n+1}$$

where $ind(X_k) \leq 0$ for $1 \leq k \leq n+1$. If there is no such n, then $dd(X) = \infty$. Set $dd(\emptyset) = -1$.

XIII.22.3.53. The Second Decomposition Theorem (XIII.22.3.20.) says that $dd(X) = ind(X)$ for any separable metrizable X.

Local Dimensions

There have been (at least) two rather different local dimensions defined in the literature, with the second having some variations. The first was introduced by MENGER [Men28] and used in [HW41]. We use the nonstandard notation *lind* for this version to distinguish it from the second version.

XIII.22.3.54. DEFINITION. If X is a topological space and $x \in X$, define $lind_x(X)$ inductively: $lind_x(X) \leq n$ if and only if, whenever U is an open neighborhood of x in X, there is an open neighborhood V of x contained in U with $ind(\partial V) \leq n-1$. Define $lind_x(X)$ to be the smallest n such that $lind_x(X) \leq n$. If there is no such n, set $lind_x(X) = \infty$.

XIII.22.3.55. We clearly have $lind_x(X) \leq ind(X)$ for each x, and by definition $ind(X) = \sup_{x\in X} lind_x(X)$.

The second type of local dimension was defined by DOWKER [**?**] and used e.g. in [Pea75]. We give the definition using *dim*; any of the other standard dimension theories can be used in its place. We restrict to regular spaces, where the definition is more natural due to the following result:

XIII.22.3.56. PROPOSITION. Let X be a regular topological space, and $x \in X$. Then the following numbers coincide:

n_1, the smallest $n \in \{0\} \cup \mathbb{N} \cup \{\infty\}$ such that there is an open neighborhood U of x with $dim(\bar{U}) \leq n$.

n_2, the smallest $n \in \{0\} \cup \mathbb{N} \cup \{\infty\}$ such that there is a closed neighborhood U of x with $dim(U) \leq n$.

n_3, the smallest $n \in \{0\} \cup \mathbb{N} \cup \{\infty\}$ such that there is a neighborhood U of x with $dim(U) \leq n$.

The same result holds if *dim* is replaced by *ind* or *Ind*.

PROOF: Trivially $n_3 \leq n_2 \leq n_1$. Let U be a neighborhood of x with $dim(U) = n_3$. Then U contains an open neighborhood V of x, and V contains an open neighborhood W of x with $\bar{W} \subseteq V$ (). Since $\bar{W}$ is a closed subset of U, by () we have

$$n_1 \leq dim(\bar{W}) \leq dim(U) = n_3 \ .$$

XIII.22.3.57. DEFINITION. Let X be a regular topological space, and $x \in X$. The local dimension $dim_x(X)$ of X at x is the common value of the three numbers in XIII.22.3.56.. (If X is not regular, it is customary to define $dim_x(X)$ to be n_1.) Similarly define ind_x and $Ind_x(X)$. Define

$$loc\, dim(X) = \sup_{x\in X} dim_x(X)$$
$$loc\, ind(X) = \sup_{x\in X} ind_x(X)$$
$$loc\, Ind(X) = \sup_{x\in X} Ind_x(X) \ .$$

XIII.22.3.58. We clearly have $dim_x(X) \leq dim(X)$, $ind_x(X) \leq ind(X)$, and $Ind_x(X) \leq Ind(X)$ for any X and x; thus $loc\,dim(X) \leq dim(X)$, $loc\,ind(X) \leq ind(X)$, $loc\,Ind(X) \leq Ind(X)$ for any X.

If X is T_1, then $ind(Y) \leq Ind(Y)$ for any subspace Y of X, so $ind_x(X) \leq Ind_x(X)$ for any $x \in X$ and hence $loc\,ind(X) \leq loc\,Ind(X)$. Similarly, if X is normal, then any closed subset Y of X is normal, so $dim(Y) \leq Ind(Y)$; thus $dim_x(X) \leq Ind_x(X)$ for any $x \in X$ and hence $loc\,dim(X) \leq loc\,Ind(X)$.

XIII.22.3.59. PROPOSITION. Let X be a topological space. Then $loc\,ind(X) = ind(X)$.

PROOF: We have $loc\,ind(X) \leq ind(X)$. For the reverse inequality, set $loc\,ind(X) = n$. Let $x \in X$ and let U be an open neighborhood of X with $ind(\bar{U}) = n$. Let V be an open neighborhood of x. Then $U \cap V$ is an open neighborhood of X contained in U. Since $ind(\bar{U}) = n$, there is a relatively open neighborhood W of x contained in $U \cap V$ with $ind(\partial_{\bar{U}} W) \leq n - 1$. But since $U \cap V$ is open in X, so is W, and since $W \subseteq U$, $\partial_{\bar{U}} W = \partial_X W$. Since $W \subseteq V$ and $ind(\partial_X W) \leq n - 1$, and $x \in X$ is arbitrary, $ind(X) \leq n$. 👍

XIII.22.3.60. Thus $loc\,ind$ is of little or no interest. However, $loc\,dim$ can differ from dim, and $loc\,Ind$ from Ind, so they are of some interest. The most dramatic example is that, under some set-theoretic assumptions such as the Continuum Hypothesis or the Diamond Principle, for any $n \geq 4$ a normal locally Euclidean space X can be constructed in which each point has an open neighborhood homeomorphic to $\mathbb{R}^n$ (X is a "nonmetrizable n-manifold"), so $ind(X) = loc\,dim(X) = loc\,Ind(X) = n$, but for which $dim(X)$ and $Ind(X)$ are greater than n. See [Fed98] (which contains a detailed survey of the history of dimension theory). Other examples show that the failure of equality for general compact Hausdorff spaces does not depend on additional set-theoretic assumptions.

XIII.22.3.61. COROLLARY. If X is T_1, then $ind(X) \leq loc\,Ind(X)$.

To compare and contrast the approaches to local dimension of MENGER and DOWKER:

XIII.22.3.62. PROPOSITION. Let X be a regular topological space, and $x \in X$. Then $lind_x(X) \leq ind_x(X)$.

PROOF: Set $ind_x(X) = n$. Let U be an open neighborhood of X with $ind(\bar{U}) = in$. Then there is an open neighborhood V of x with $\bar{V} \subseteq U$. Since $ind(\bar{U}) = n$, there is a relatively open neighborhood W of x contained in V with $ind(\partial_{\bar{U}} W) \leq n-1$. But since V is open in X, so is W, and since $\bar{V} \subseteq U$, $\partial_{\bar{U}} W = \partial_X W$. Thus $lind_x(X) \leq n$. 👍

For regular spaces X, this gives another proof that $loc\,ind(X) = ind(X)$:

$$ind(X) = \sup_{x \in X} lind_x(X) \leq \sup_{x \in X} ind_x(X) = loc\,ind(X) \leq ind(X) \ .$$

This also implies that if X is regular and $ind(X) < \infty$, then $lind_x(X) = ind_x(X) = ind(X)$ for at least one $x \in X$. But not necessarily for all x:

XIII.22.3.63. EXAMPLE. The local dimensions $lind_x$ and ind_x can differ dramatically at some points, even in nice spaces. Let X be the one-point compactification of the separated union of $[0,1]^n$ for all n. Then X is a compact metrizable space. If p is the point at infinity, then $lind_p(X) = 0$ since there is a local base at p of clopen sets; but $ind(U) = \infty$ for any neighborhood U of p, so $ind_p(X) = \infty$.

The set of x for which $lind_x(X) = ind(X)$ is interesting to consider, and is the primary interest of MENGER'S approach; cf. Exercise XIII.22.8.5..

XIII.22.3.64. REMARK. If X is a topological space, we can also define

n_4, the smallest $n \in \{0\} \cup \mathbb{N} \cup \{\infty\}$ such that there is an open neighborhood U of x with $dim(U) \leq n$.

If X is regular, $n_4 \geq n_3 = dim_x(X)$, but they are not obviously equal since dim is not monotone in general. Similarly, if Ind is used in place of dim, we obtain that $n_4 \geq Ind_x(X)$. If X is totally normal (e.g. if X is metrizable), then $n_4 = dim_x(X)$ or $n_4 = Ind_x(X)$, as the case may be, for any $x \in X$.

XIII.22.4. Comparing the Dimensions

All the dimensions defined so far agree for separable metric spaces:

XIII.22.4.1. THEOREM. Let X be a topological space which is separable and metrizable (equivalently, second countable and regular). Then $ind(X) = Ind(X) = dim(X) = ed(X) = pd(X) = dd(X)$.

This result not only confirms that for "nice" spaces there is an unambiguous notion of topological dimension, but is also useful since some properties of dimension are proved easily for some of the dimension functions and not so easily for others. Proofs of the theorem can be found in [HW41] and [Eng95].

For general spaces, even for metrizable spaces or compact Hausdorff spaces, the dimension functions do not always agree. We state without proof some of the principal relations between them; see e.g. [HW41], [Pea75], [Eng95].

XIII.22.4.2. THEOREM. ([Kat52], [Mor54b]) If X is a metrizable space (not necessarily separable), then $Ind(X) = dim(X)$.

XIII.22.4.3. There is a completely metrizable space X with $ind(X) = 0$ and $Ind(X) = dim(X) = 1$ [Roy68]. There are first countable compact Hausdorff spaces X with $ind(X) < Ind(X)$ ([Fil69], [Pas70]). More examples can be found in [Nag70] and [Pea75].

XIII.22.4.4. THEOREM. [?] If X is a compact Hausdorff space, then $dim(X) \leq ind(X)$.

The result extends to strongly paracompact spaces, and in particular to Lindelöf spaces, but not even to completely metrizable spaces as ROY'S example shows. There is a compact Hausdorff space X with $dim(X) < ind(X)$ ([Lun49]; cf. [Eng95, 3.1.31]).

XIII.22.4.5. THEOREM. [Ale47] If X is completely regular, then $ed(X) = zdim(X)$. If X is normal, then $ed(X) = dim(X)$.

The next result is a corollary of XIII.22.3.7. and XIII.22.4.1.:

XIII.22.4.6. COROLLARY. Let X be a separable metrizable space and Y a subset of X. Then $dim(Y) \leq dim(X)$.

XIII.22.4.7. There is an example of a compact Hausdorff space X and a normal subspace Y such that $dim(X) = Ind(X) = 0$ and $dim(Y) = Ind(Y) = 1$ ([Dow55]; cf. [Eng95, 2.2.11], [GJ76, 16M]).

For separable metrizable spaces, we obtain a strengthening of XIII.22.3.29.:

XIII.22.4.8. PROPOSITION. Let X be a separable metrizable space (or, more generally, a completely normal hereditarily Lindelöf space). Then $dim(X) \leq dd(X) (= ind(X))$.

PROOF: Suppose $dd(X) \leq n$, i.e. $X = X_1 \cup \cdots \cup X_{n+1}$ with $ind(X_k) \leq 0$ for all k. Let $\mathcal{U} = \{U_1, \ldots, U_m\}$ be a finite open cover of X. For each k, we have $dim(X_k) \leq 0$ (), so the cover $\{U_1 \cap X_k, \ldots, U_m \cap X_k\}$ of X_k has a shrinking $\{F_1, \ldots, F_m\}$ consisting of relatively clopen subsets of X_k. Since X is completely normal, there are disjoint open neighborhoods $\{V_{1,k}, \ldots, V_{m,k}\}$ of $F_1, \ldots, F_m$ in X. Consider the open cover $\mathcal{W}$ of X consisting of all sets of the form $U_j \cap V_{j,k}$. Then $\mathcal{W}$ refines $\mathcal{U}$, and has order $\leq n$ since for any $n+2$ distinct sets $U_{j_i} \cap V_{j_i,k_i}$ in $\mathcal{W}$ there must be some k with $k_i = k$ for two distinct i. Thus $dim(X) \leq n$. 👍

This result also follows from XIII.22.3.29. and the fact that $dd(X) = ind(X) = Ind(X)$ for separable metrizable spaces (XIII.22.3.17., XIII.22.3.53.), but this requires the Sum Theorem for ind; the inequality $dim(X) \leq dd(X)$ for separable metrizable spaces does not require the Sum Theorem.

XIII.22.4.9. PROPOSITION. Let X be a normal topological space. Then $pd(X) \leq dim(X)$.

PROOF: Suppose $dim(X) \leq n$, and let $\{(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})\}$ be a set of $n+1$ pairs of disjoint closed subsets of X. Then the open cover

$$\left\{ X \setminus A_1,\ \ldots\ ,\ X \setminus A_{n+1},\ X \setminus \left(\bigcup_{k=1}^{n+1} B_k \right) \right\}$$

has a shrinking $\mathcal{V} = \{V_1, \ldots, V_{n+2}\}$ of order $\leq n$. Let $\{F_1, \ldots, F_{n+2}\}$ be a closed shrinking of $\mathcal{V}$. For $1 \leq k \leq n+1$ define

$$\tilde{A}_k = A_k \cup (F_{n+2} \setminus V_k), \quad \tilde{B}_k = B_k \cup F_k\ .$$

Since $A_k \cap F_k = \emptyset$ and $B_k \cap F_{n+2} = \emptyset$ for each k, we have $\tilde{A}_k \cap \tilde{B}_k = \emptyset$. Also,

$$\bigcup_{k=1}^{n+1} F_k \subseteq \bigcup_{k=1}^{n+1} \tilde{B}_k\ .$$

We have that

$$F_{n+2} \cap \left(\bigcap_{k=1}^{n+1} V_k \right) \subseteq V_{n+2} \cap \left(\bigcap_{k=1}^{n+1} V_k \right) = \emptyset$$

since $\mathcal{V}$ has order $\leq n$. Thus $F_{n+2} \subseteq \cup_{k=1}^{n+1} \tilde{A}_k$. Since $\cup_{k=1}^{n+2} F_k = X$, we have

$$\bigcup_{k=1}^{n+1} (\tilde{A}_k \cup \tilde{B}_k) = X \ .$$

Let P_k be a partition between $\tilde{A}_k$ and $\tilde{B}_k$ for each k. Then P_k is a partition between A_k and B_k, and $\cap_{k=1}^{n+1} P_k = \emptyset$ since it is disjoint from $\cup_{k=1}^{n+1} (\tilde{A}_k \cup \tilde{B}_k) = X$. 👍

XIII.22.4.10. PROPOSITION. Let X be a normal space. Then $dim(X) = pd(X)$.

PROOF: Because of XIII.22.4.9., we need to show that $dim(X) \leq pd(X)$. Assume $pd(X) \leq n$, and let $\mathcal{U} = \{U_1, \ldots, U_{n+2}\}$ be an open cover of X. There is a closed shrinking $\{A_1, \ldots, A_{n+2}\}$ of $\mathcal{U}$. Set $B_k = X \backslash U_k$. Then $\{(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})\}$ is a family of pairs of disjoint closed sets in X, so there are partitions P_k between A_k and B_k for $1 \leq k \leq n+1$ with $\cap_{k=1}^{n+1} P_k = \emptyset$, i.e. for each $k \leq n+1$ there are disjoint open sets V_k, W_k containing A_k and B_k respectively with

$$\bigcap_{k=1}^{n+1} X \setminus (V_k \cup W_k) = \emptyset \ .$$

Set

$$V_{n+2} = U_{n+2} \cap \left(\bigcup_{k=1}^{n+1} W_k \right) \ .$$

Then $V_k \subseteq U_k$ for all k, and

$$\begin{aligned}
\bigcup_{k=1}^{n+2} V_k &= \left(\bigcup_{k=1}^{n+1} V_k \right) \cup \left(U_{n+2} \cap \left[\bigcup_{k=1}^{n+1} W_k \right] \right) \\
&= \left(\bigcup_{k=1}^{n+1} V_k \cup U_{n+2} \right) \cap \left(\left[\bigcup_{k=1}^{n+1} V_k \right] \cup \left[\bigcup_{k=1}^{n+1} W_k \right] \right) \\
&= \left(\bigcup_{k=1}^{n+1} V_k \cup U_{n+2} \right) \cap X = \bigcup_{k=1}^{n+1} V_k \cup U_{n+2} \supseteq \bigcup_{k=1}^{n+2} A_k = X
\end{aligned}$$

so $\mathcal{V} = \{V_1, \ldots, V_{n+2}\}$ is a shrinking of $\mathcal{U}$. We also have

$$\begin{aligned}
\bigcap_{k=1}^{n+2} V_k &= \left(\bigcap_{k=1}^{n+1} V_k \right) \cap \left(U_{n+2} \cap \left[\bigcup_{k=1}^{n+1} W_k \right] \right) \\
&\subseteq \left(\bigcap_{k=1}^{n+1} V_k \right) \cap \left(\bigcup_{k=1}^{n+1} W_k \right) = \emptyset
\end{aligned}$$

so $\mathcal{V}$ has order $\leq n$. By XIII.22.3.33., $dim(X) \leq n$. 👍

XIII.22.4.11. PROPOSITION. Let X be a completely regular space. Then $pd(X) \leq ed(X)$.

PROOF: We will use XIII.22.1.8.. Suppose $ed(X) \leq n$, and let $(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})$ be $n+1$ pairs of disjoint closed sets in X. Let Y be the union of all the A_k and B_k. Then Y is a closed subset of X. For each k, $1 \leq k \leq n+1$, set $f_k(x) = 0$ if $x \in A_k$ and $f_k(x) = 1$ if $x \in B_k$. If $f = (f_1, \ldots, f_{n+1})$, then f is a map from Y to $\partial I^{n+1} \cong S^n$. By XIII.22.1.8., f extends to a map g from X to ∂I^{n+1}. Let g_k $(1 \leq k \leq n+1)$ be the k'th coordinate function of g; then g_k extends f_k to X. Set $P_k = g_k^{-1}(1/2)$. Then P_k is a partition between A_k and B_k, and $P_1 \cap \cdots P_{k+1} = \emptyset$ since $\left(\frac{1}{2}, \ldots, \frac{1}{2}\right) \notin \partial I^{n+1}$. Thus $pd(X) \leq n$. 👍

XIII.22.4.12. PROPOSITION. Let X be a normal space. Then $ed(X) = pd(X)$.

PROOF: We need only show $ed(X) \leq pd(X)$. We again use XIII.22.1.8.. Suppose $pd(X) \leq n$, Y is a closed subset of X, and f is a map from Y to ∂I^{n+1}. For $1 \leq k \leq n+1$, let A_k [resp. B_k] be the set of $x \in X$ such that the k'th coordinate of $f(x)$ is 0 [resp. 1]. Then there are partitions P_k between A_k and B_k for each k, with $\cap_{k=1}^{n+1} P_k = \emptyset$ since $pd(X) \leq n$, i.e. for each $k \leq n+1$ there are disjoint open sets V_k, W_k containing A_k and B_k respectively with

$$\bigcap_{k=1}^{n+1} X \setminus (V_k \cup W_k) = \emptyset \ .$$

Let $\{U_1, \ldots, U_{n+1}\}$ be a swelling of $\{P_1, \ldots, P_{n+1}\}$ (XIII.7.6.11.), so $P_k \subseteq U_k$ and $\cap_{k=1}^{n+1} U_k = \emptyset$. Then $V_k \cup P_k$ and $W_k \cup P_k$ are closed in X, hence normal, so by Urysohn's Lemma there is a map $\phi_k : V_k \cup P_k \to \left[0, \frac{1}{2}\right]$ with $\phi_k(V_k \setminus U_k) = \{0\}$ and $\phi_k(P_k) = \left\{\frac{1}{2}\right\}$, and a map $\psi_k : W_k \cup P_k \to \left[\frac{1}{2}, 1\right]$ with $\psi_k(P_k) = \left\{\frac{1}{2}\right\}$ and $\psi_k(W_k \setminus U_k) = \{1\}$. Then ϕ_k and ψ_k define a map f_k from X to $[0,1]$ since they agree on the closed set $P_k = (V_k \cup P_k) \cap (W_k \cup P_k)$. We have $g_k(A_k) = \{0\}$ and $g_k(B_k) = \{1\}$. The g_k define a map $g : X \to I^n$. Since $g_k^{-1}\left(\left\{\frac{1}{2}\right\}\right) \subseteq U_k$ for each k and $\cap_{k=1}^{n+1} U_k = \emptyset$, g never takes the value $p = \left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$. If r is a retraction from $I^n \setminus \{p\}$ to ∂I^n, then $r \circ g$ is an extension of f to X. 👍

Combining XIII.22.4.12. with XIII.22.4.10., we obtain:

XIII.22.4.13. COROLLARY. Let X be a normal space. Then $dim(X) = pd(X) = ed(X)$.

XIII.22.4.14. In fact, by an extension of the same methods it can be shown that $ed(X) = zdim(X)$ for any completely regular X ().

The hardest of the inequalities in XIII.22.4.1. is that $ind(X) \leq dim(X)$. We now work on this.

XIII.22.4.15. We have $Ind(X) \leq dim(X)$ (and hence $Ind(X) = dim(X)$ by XIII.22.3.29.) if X is metrizable.

Summary

We summarize the inequalities for the various dimension functions we have considered (except dd). To simplify to the only cases of substantial interest, we will only consider *normal* spaces.

XIII.22.4.16. If X is a normal topological space, then

$$ind(X) \leq loc\, Ind(X) \leq Ind(X)$$

$$loc\, dim(X) \leq dim(X) \leq Ind(X)$$

$$ed(X) = pd(X) = dim(X) \ .$$

Examples show that there are no further inequalities among $ind(X)$, $Ind(X)$, $dim(X)$, $loc\, Ind(X)$, $loc\, dim(X)$, $ed(X)$, and $pd(X)$ for general normal X.

If X is a compact Hausdorff space, then we have

$$loc\, dim(X) = dim(X) \leq ind(X) \leq loc\, Ind(X) = Ind(X)$$

$$ed(X) = pd(X) = dim(X) \ .$$

Examples show that there are no further inequalities among $ind(X)$, $Ind(X)$, $dim(X)$, $loc\, Ind(X)$, $loc\, dim(X)$, $ed(X)$, and $pd(X)$ for general compact Hausdorff X.

If X is a metrizable topological space, then we have

$$ind(X) \leq loc\, Ind(X) = loc\, dim(X) = dim(X) = Ind(X) = ed(X) = pd(X) \ .$$

If X is separable and metrizable (equivalently, second countable and regular), then $ind(X)$ is equal to the others. Examples show that $ind(X)$ can be less than the others for general metrizable X.

XIII.22.5. More Dimension Theories

Analytic Dimension

XIII.22.5.1. Let X be a completely regular topological space. We will give a characterization due to Katětov of $zdim(X)$ in terms of the ring (algebra) $BC_{\mathbb{R}}(X)$ of bounded real-valued continuous functions on X. (This ring is denoted $C^*(X)$ in many references such as [GJ76], which is a thorough study of topology via these rings. We have adopted different notation since $C^*(X)$ is too easily confused with C*-algebra notation.) $BC_{\mathbb{R}}(X)$ has a norm

$$\|f\| = \sup_{x \in X} |f(x)|$$

under which it is a real Banach space. The norm is determined by the algebraic structure since it can be defined using the order structure, and a function is nonnegative if and only if it is a square. Note that every subring of $BC_{\mathbb{R}}(X)$ containing the constant functions is a subalgebra (all subrings we consider will contain the constant functions).

XIII.22.5.2. If X is compact, or more generally countably compact, then $BC_{\mathbb{R}}(X) = C_{\mathbb{R}}(X)$, the set of all real-valued continuous functions on X [if $f : X \to \mathbb{R}$ is continuous, then $\{f^{-1}((-n, n)) : n \in \mathbb{N}\}$ is a countable open cover of X which has a finite subcover]. A space X with $BC_{\mathbb{R}}(X) = C_{\mathbb{R}}(X)$ is called *pseudocompact*; there exist pseudocompact spaces which are not countably compact, but every normal pseudocompact space is countably compact [GJ76, 3D]. In general, we have a natural identification $BC_{\mathbb{R}}(X) \cong C_{\mathbb{R}}(\beta X)$, where βX is the Stone-Čech compactification of X ().

XIII.22.5.3. The idea behind KATĚTOV'S characterization is that if X is a compact subset of $\mathbb{R}^n$, then polynomials in the coordinate functions are dense in $C_{\mathbb{R}}(X)$, i.e. $C_{\mathbb{R}}(X)$ has a dense unital subalgebra generated by n elements (the coordinate functions). This turns out to be the minimum number of generators of a dense unital subalgebra if $dim(X) = n$, e.g. if X has nonempty interior.

XIII.22.5.4. Simply taking the minimum number of generators of a dense unital subalgebra does not work as a satisfactory notion of dimension for a general X, however. For one thing, $BC_{\mathbb{R}}(X)$ can easily be nonseparable, and will be if βX is not metrizable, e.g. if X is not compact; then even no countably generated subalgebra can be dense. And the wrong answer is obtained even for a space as simple as a circle S^1: the minimum number of generators of a dense unital subalgebra of $C_{\mathbb{R}}(S^1)$ is two, since no real-valued continuous function on S^1 separates points.

The key trick is to consider analytic subrings:

XIII.22.5.5. DEFINITION. Let X be a completely regular space. An *analytic subring* of $BC_{\mathbb{R}}(X)$ is a closed unital subalgebra A with the property that if $f^2 \in A$, then $f \in A$. The smallest analytic subring containing a set $\mathcal{F}$ of functions is called the *analytic subring generated by* $\mathcal{F}$.

Note that any intersection of analytic subrings is an analytic subring, so there is always a smallest analytic subring containing any given set of functions.

XIII.22.5.6. EXAMPLES. (i) The example of the unit circle $S^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ nicely illustrates the idea. Let $f(x, y) = x$. Then the unital subalgebra generated by f is not dense in $C_{\mathbb{R}}(S^1)$. However, if A is the analytic subring of $C_{\mathbb{R}}(S^1)$ generated by f, then A contains $g = 1 - f^2$. We have $g(x, y) = 1 - x^2 = y^2$. Thus A also contains $h(x, y) = y$ since $g = h^2$, and it follows from the Stone-Weierstrass Theorem () that A is all of $C_{\mathbb{R}}(S^1)$. Thus $C_{\mathbb{R}}(S^1)$ is generated as an analytic subring by one function, i.e. "S^1 is one-dimensional."

(ii) Let X be compact. Consider the analytic subring A generated by the empty set, i.e. the smallest analytic subring of $C_{\mathbb{R}}(X)$. Then A contains the constant function 1 by definition, so it also contains every function taking exactly the values ± 1, and thus the indicator function of every clopen subset of X. So all continuous functions with finite range are in A. A is just the closure of this set, i.e. A is the closed subalgebra of $C_{\mathbb{R}}(X)$ generated by the indicator functions of clopen sets in X. So A is all of $C_{\mathbb{R}}(X)$ if and only if X is totally disconnected, i.e. zero-dimensional ().

XIII.22.5.7. If $\mathcal{F}$ is a subset of $BC_{\mathbb{R}}(X)$, a subset Y of X is a *stationary set* for $\mathcal{F}$ if every $f \in \mathcal{F}$ is constant on Y. If Y is a stationary set for $\mathcal{F}$, so is $\bar{Y}$, so we need consider only closed stationary sets. Of course, each singleton is a stationary set, and the converse is true if and only if $\mathcal{F}$ separates points of X. The Stone-Weierstrass Theorem can be rephrased:

XIII.22.5.8. THEOREM. Let X be a compact Hausdorff space, and $\mathcal{F}$ a subset of $C_{\mathbb{R}}(X)$. Then the closed unital subalgebra of $C_{\mathbb{R}}(X)$ generated by $\mathcal{F}$ consists of all functions in $C_{\mathbb{R}}(X)$ which are constant on each stationary set of $\mathcal{F}$.

A variation of this, due to KATĚTOV, is the key result:

XIII.22.5.9. THEOREM. Let X be a compact Hausdorff space, and $\mathcal{F}$ a subset of $C_{\mathbb{R}}(X)$. Then the analytic subring of $C_{\mathbb{R}}(X)$ generated by $\mathcal{F}$ consists of all functions in $C_{\mathbb{R}}(X)$ which are constant on each connected stationary set of $\mathcal{F}$.

One can restrict attention to maximal connected stationary sets: every connected stationary set is contained in a maximal connected stationary set (XIII.13.3.2.), which is automatically closed (XIII.13.1.9.).

XIII.22.5.10. We can provisionally define the analytic dimension of X to be the smallest n such that $BC_{\mathbb{R}}(X)$ is generated as an analytic subring by n elements. This is a good definition if $BC_{\mathbb{R}}(X)$ is separable, which occurs if and only if X is compact and metrizable (). In general, we make the following definition:

XIII.22.5.11. DEFINITION. Let X be a completely regular space. The *analytic dimension* $ad(X)$ is the smallest n, $0 \leq n \leq \aleph_0$, such that every countably generated (i.e. separable) closed subalgebra of $BC_{\mathbb{R}}(X)$ is contained in an analytic subring generated by n elements.

The analytic dimension $ad(X)$ is the smallest n such that every finite subset of $BC_{\mathbb{R}}(X)$ is contained in an analytic subring generated by n elements. If X is compact and metrizable, then $ad(X)$ is the smallest n such that $BC_{\mathbb{R}}(X)$ is generated as an analytic subring by n elements. Note that the analytic subring generated by even finitely many elements need not be separable: if X is the product of uncountably many two-point spaces, then X is compact and zero-dimensional, and the analytic subring of $C_{\mathbb{R}}(X)$ generated by $\emptyset$ is all of $C_{\mathbb{R}}(X)$, which is nonseparable since X is not metrizable.

The main theorem of KATĚTOV is:

XIII.22.5.12. THEOREM. Let X be a completely regular space. Then $ad(X) = zdim(X)$. Thus, if X is normal, $ad(X) = dim(X)$.

See [GJ76] for a complete discussion and proof, which is fairly involved. Since $ad(X)$ is transparently equal to $ad(\beta X)$ for any X, we obtain:

XIII.22.5.13. COROLLARY. Let X be a completely regular space. Then $zdim(X) = zdim(\beta X) = dim(\beta X)$. If X is normal, then $dim(X) = dim(\beta X)$.

The corollary can be proved directly without introducing analytic dimension (cf. XIII.22.1.12., XIII.22.4.5.; [GJ76, 16.11]).

Affine Dimension

Homological Dimension

XIII.22.5.14. There are important metric notions of dimension which are not topological invariants, such as Hausdorff dimension, packing dimension, and Assouad dimension ().

XIII.22.6. Embedding in Euclidean Space

In this subsection, we prove two important theorems about embedding of topological spaces in Euclidean space.

If X is a topological space of dimension n, we want to embed X in $\mathbb{R}^p$ for some p which is a function of n. Of course, for X to embed in any Euclidean space it is necessary that X be separable and metrizable, so we make an assumption that all spaces in this section are separable and metrizable. Thus all standard dimension functions agree on such an X, so "dimension n" is unambiguous.

XIII.22.6.1. THEOREM. Let X be a subset of $\mathbb{R}^n$. Then X has dimension n if and only if the interior of X is nonempty, i.e. if and only if X contains a nonempty open set in $\mathbb{R}^n$.

The proof is deferred to XIII.22.6.14..

XIII.22.6.2. If a (separable metrizable) space X has dimension n, it cannot be embedded in $\mathbb{R}^n$ in general (consider a circle). It cannot even generally be embedded in $\mathbb{R}^{n+1}$: a graph cannot always be embedded in $\mathbb{R}^2$, and the projective plane and the Klein bottle are closed 2-manifolds which cannot be embedded in $\mathbb{R}^3$. The n-skeleton of a $2n+2$-simplex cannot be embedded in $\mathbb{R}^{2n}$ []. The best which can be done in general is the following:

XIII.22.6.3. THEOREM. Let X be a separable metrizable space with $dim(X) \leq n$. Then X embeds in $\mathbb{I}^{2n+1}$. In fact, every continuous function from X to $\mathbb{I}^{2n+1}$ is a uniform limit of embeddings.

To prove XIII.22.6.3. (and in fact to prove a slightly better result), we introduce notation for some useful subsets of $\mathbb{R}^n$:

XIII.22.6.4. DEFINITION. Let $0 \leq n \leq m$. Then

$\mathfrak{P}_n^m$ is the set of points in $\mathbb{R}^m$ with precisely n rational coordinates.

$\mathfrak{M}_n^m$ is the set of points in $\mathbb{R}^m$ with at most n rational coordinates.

$\mathfrak{L}_n^m$ is the set of points in $\mathbb{R}^m$ with at least n rational coordinates.

$\mathfrak{N}_n = \mathfrak{M}_n^{2n+1}$ is the *Nöbeling universal space of dimension n*.

Note that $\mathfrak{M}_n^m \cup \mathfrak{L}_n^m = \mathbb{R}^m$ and $\mathfrak{M}_n^m \cap \mathfrak{L}_n^m = \mathfrak{P}_n^m$ for any n and m.

XIII.22.6.5. PROPOSITION. For any $0 \leq n \leq m$, $dim(\mathfrak{P}_n^m) = 0$, $dim(\mathfrak{M}_n^m) = n$, $dim(\mathfrak{L}_n^m) = m - n$, $dim(\mathfrak{N}_n) = n$.

PROOF: $\mathfrak{P}_n^m$ is a countable union of (relatively) closed subsets where n specified coordinates are specified fixed rational numbers and the other $m-n$ coordinates are irrational. Such a set is homeomorphic to $\mathbb{J}^{m-n}$, which has dimension 0. Thus $dim(\mathfrak{P}_n^m) = 0$ by the Sum Theorem (the zero-dimensional version XIII.22.3.14. suffices).

The observations

$$\mathfrak{M}_n^m = \mathfrak{P}_0^m \cup \mathfrak{P}_1^m \cup \cdots \cup \mathfrak{P}_n^m$$
$$\mathfrak{L}_n^m = \mathfrak{P}_n^m \cup \mathfrak{P}_{n+1}^m \cup \cdots \cup \mathfrak{P}_m^m$$

and the Sum Theorem imply that $dim(\mathfrak{M}_n^m) \leq n$ and $dim(\mathfrak{L}_n^m) \leq m-n$. But $\mathfrak{M}_n^m$ contains a subspace homeomorphic to $\mathbb{R}^n$ (say the set of all points with the first $m-n$ coordinates π), and $\mathfrak{L}_n^m$ similarly contains a copy of $\mathbb{R}^{m-n}$, so $dim(\mathfrak{M}_n^m) \geq n$ and $dim(\mathfrak{L}_n^m) \geq m-n$ by XIII.22.3.7.. 👍

The theorem we will prove is:

XIII.22.6.6. THEOREM. Let X be a separable metrizable space of dimension $\leq n$. Then there is an embedding $f : X \to \mathfrak{N}_n$ such that the closure of $f(X)$ in $\mathfrak{N}_n$ is compact. In fact, any bounded continuous function from X to $\mathbb{R}^{2n+1}$ can be uniformly approximated arbitrarily closely by such a function.

This result justifies the name "Nöbeling universal space." Theorem XIII.22.6.3. is an immediate corollary. Another important corollary is:

XIII.22.6.7. COROLLARY. Let X be a separable metrizable space of dimension n. Then X has a metrizable compactification of dimension n. In fact, there is a compact subset K_n of $\mathbb{R}^{2n+1}$ with $dim(K_n) = n$ such that every separable metrizable space of dimension $\leq n$ embeds in K_n.

PROOF: There is an embedding $f : \mathfrak{N}_n \to \mathfrak{N}_n$ such that $K_n = \overline{f(\mathfrak{N}_n)}$ is a compact subset of $\mathfrak{N}_n$. 👍

Actually, the Menger Universal Space M_n^{2n+1} is such a space (XIII.19.5.15.).

XIII.22.6.8. The strategy of the proof of XIII.22.6.6. will be to show that the set E of embeddings of X into $\mathfrak{N}_n$ such that the closure of $f(X)$ in $\mathfrak{N}_n$ is compact is a countable intersection of dense open sets in $BC(X, \mathbb{R}^{2n+1})$. The theorem will then follow from the Baire Category Theorem ().

We fix a totally bounded metric ρ on X. There is such a metric since X embeds in the Hilbert cube (). There is thus for each m a finite open cover $\mathcal{U}_m$ of X by open balls of diameter $\leq 1/m$; $\cup_m \mathcal{U}_m$ is a base for the topology of X, and in fact any finite open cover of X is refined by one of the U_m (). By passing to a subsequence, we may assume $\mathcal{U}_{m+1}$ refines U_m for each m.

Let $\mathcal{F}$ be the set of all tuples of the form

$$\mathcal{F} = \{(i_1, \ldots, i_{n+1}; q_1, \ldots, q_{n+1}) : 1 \leq i_1 < i_2 < \cdots < i_{n+1} \leq 2n+1,\ q_k \in \mathbb{Q}\}\ .$$

Then $\mathcal{F}$ is countable. For each $F \in \mathcal{F}$, define

$$L_F = \{(x_1, \ldots, x_{2n+1}) \in \mathbb{R}^{2n+1} : x_{i_k} = q_k \text{ for } 1 \leq k \leq n+1\}\ .$$

Then L_F is a hyperplane in $\mathbb{R}^{2n+1}$ of dimension n, $L_F \cap \mathfrak{N}_n = \emptyset$, and

$$\bigcup_{F \in \mathcal{F}} L_F = \mathfrak{L}_{n+1}^{2n+1} = \mathbb{R}^{2n+1} \setminus \mathfrak{N}_n\ .$$

For each $F \in \mathcal{F}$ and $m \in \mathbb{N}$, let $W_{F,m}$ be the set of all $f \in BC(X, \mathbb{R}^{2n+1})$ such that $\overline{f(X)}$ is disjoint from L_F and f is an $\mathcal{U}_m$-map, i.e. for every $y \in \overline{f(X)}$ there is an open neighborhood V of y such that $f^{-1}(V) \subseteq U$ for some $U \in \mathcal{U}_m$. We will show that $W_{F,m}$ is a dense open set in $BC(X, \mathbb{R}^{2n+1})$. This will prove the theorem since

$$E = \bigcap_{F,m} W_{F,m} \ .$$

XIII.22.6.9. To show that $W_{F,m}$ is open in $BC(X, \mathbb{R}^{2n+1})$, fix $f \in W_{F,m}$. Since $\overline{f(X)}$ is compact and disjoint from L_F, the distance from $\overline{f(X)}$ to L_F is a positive number η. For each $y \in \overline{f(X)}$, there is an open neighborhood V_y in $\mathbb{R}^{2n+1}$ such that $f^{-1}(V_y)$ is contained in some $U \in \mathcal{U}_m$. Let γ be the Lebesgue number of the cover $\mathcal{V} = \{V_y : y \in \overline{f(X)}\}$ of $\overline{f(X)}$, i.e. if B is a ball of radius $\leq \gamma$ in $\mathbb{R}^n$ intersecting $\overline{f(X)}$, then $B \subseteq V_y$ for some $y \in \overline{f(X)}$, and thus $f^{-1}(B)$ is contained in some $U \in \mathcal{U}_m$.

Set $\delta = \min(\gamma/3, \eta/2)$. If $g \in BC(X, \mathbb{R}^{2n+1})$ with $\|f - g\| < \delta$, then for any $x \in X$ and $a \in L_F$ we have

$$\|g(x) - a\| \geq \|f(x) - a\| - \|f(x) - g(x)\| > \frac{\eta}{2}$$

so $\overline{g(X)} \cap L_f = \emptyset$.

Let $y \in \overline{g(X)}$. There is a $z \in X$ with $\|y - g(z)\| < \delta$. If B is the open ball of radius δ around y, and $x \in g^{-1}(B)$, then

$$\|g(x) - f(z)\| \leq \|g(x) - g(z)\| + \|g(z) - f(z)\| < 3\delta \leq \gamma$$

since $g(x)$ and $g(z)$ are both in B and $\|f - g\| < \delta$. Thus $g(x)$ is in the open ball C around $f(z)$ of radius γ, i.e. $g^{-1}(B) \subseteq f^{-1}(C)$, which is contained in some $U \in \mathcal{U}_m$. Thus $g \in W_{F,m}$. So $W_{F,m}$ is open.

Showing that $W_{F,m}$ is dense in $BC(X, \mathbb{R}^{2n+1})$ takes more work. We need to approximate using vectors in general position.

XIII.22.6.10. Proposition. If S is a finite union of affine subspaces of $\mathbb{R}^n$ of dimension $\leq n-1$, then the complement of S is dense (and open) in $\mathbb{R}^n$.

Proof: The complement of an affine subspace of $\mathbb{R}^n$ of dimension $\leq n-1$ is dense and open, and in any topological space a finite intersection of dense open sets is dense and open. 👍

In fact, by the Baire Category Theorem, a countable union of affine subspaces of $\mathbb{R}^n$ of dimension $\leq n-1$ has dense (but usually not open) complement. This also follows from the Sum Theorem XIII.22.3.13..

XIII.22.6.11. Corollary. If $G = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$ is a subset of $\mathbb{R}^n$ in general position (), $\boldsymbol{y} \in \mathbb{R}^n$, and $\epsilon > 0$, then there is an $\boldsymbol{x}_{p+1} \in \mathbb{R}^n$ with $\|\boldsymbol{y} - \boldsymbol{x}_{p+1}\| < \epsilon$ and $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{x}_{p+1}\}$ in general position.

Proof: If $\{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_r}\}$ is a subset of G with $r \leq n$, and $\boldsymbol{x} \in \mathbb{R}^n$, then $\{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_r}, \boldsymbol{x}\}$ span an r-dimensional affine subspace if and only if $\boldsymbol{x}$ is not in the $(r-1)$-dimensional affine subspace spanned by $\{\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_r}\}$. So we need only choose $\boldsymbol{x}_{p+1}$ to not be in any of these finitely many affine subspaces. 👍

Applying this result repeatedly, we obtain:

XIII.22.6.12. COROLLARY. If $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$ is a subset of $\mathbb{R}^n$ in general position, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q \in \mathbb{R}^n$, and $\epsilon > 0$, then there are $\boldsymbol{x}_{p+1}, \ldots, \boldsymbol{x}_{p+q} \in \mathbb{R}^n$ such that $\|\boldsymbol{x}_{p+k} - \boldsymbol{y}_k\| < \epsilon$ for $1 \leq k \leq q$ and $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{p+q}\}$ is in general position.

In fact, if $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$ is in general position, $(\boldsymbol{y}_k)$ is any sequence in $\mathbb{R}^n$, and (ϵ_k) is any sequence of positive numbers, there is a sequence $(\boldsymbol{x}_{p+k})$ in $\mathbb{R}^n$ such that $\|\boldsymbol{x}_{p+k} - \boldsymbol{y}_k\| < \epsilon_k$ for every k and the infinite set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{x}_{p+1}, \ldots\}$ is in general position.

XIII.22.6.13. Now fix F and m, and let $f \in BC(X, \mathbb{R}^{2n+1})$ and $\epsilon > 0$. Let $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n\}$ be a set of points in general position spanning L_F as an affine subspace. Cover $\overline{f(X)}$ by finitely many open balls of diameter $\leq \frac{\epsilon}{2}$; by increasing m if necessary, we may assume $\mathcal{U}_m$ refines the cover of X consisting of inverse images of these balls, i.e. that the diameter of $f(U)$ is less than $\frac{\epsilon}{2}$ for all $U \in \mathcal{U}_m$. Let

$$\mathcal{V} = \{V_1, \ldots, V_q\}$$

be a shrinking of $\mathcal{U}_m$ of order $\leq n$. For each k, $1 \leq k \leq q$, fix $z_k \in V_k$. Choose $\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q} \in \mathbb{R}^{2n+1}$ such that $\|\boldsymbol{x}_{n+k} - f(z_k)\| < \frac{\epsilon}{2}$ for each k and $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{n+q}\}$ is in general position.

Let $\{h_k : 1 \leq k \leq q\}$ be a partition of unity on X subordinate to $\mathcal{V}$, e.g. the standard partition of unity from ρ (). For $x \in X$, set

$$g(x) = \sum_{k=1}^{q} h_k(x)\boldsymbol{x}_{n+k} \ .$$

We claim $g \in W_{F,m}$ and $\|f - g\| \leq \epsilon$. If $x \in X$, then for each k for which $h_k(x) > 0$, we have $x \in V_k$, so $\|f(x) - f(z_k)\| < \frac{\epsilon}{2}$ and thus $\|f(x) - \boldsymbol{x}_{n+k}\| < \epsilon$. We thus have

$$\|f(x) - g(x)\| = \left\| \sum_{k=1}^{q} h_k(x)f(x) - \sum_{k=1}^{q} h_k(x)\boldsymbol{x}_{n+k} \right\|$$

$$= \left\| \sum_{k=1}^{q} h_k(x)[f(x) - \boldsymbol{x}_{n+k}] \right\| \leq \sum_{k=1}^{q} h_k(x)\|f(x) - \boldsymbol{x}_{n+k}\| < \sum_{k=1}^{q} \epsilon h_k(x) = \epsilon$$

so $\|f - g\| < \epsilon$.

To show $g \in W_{F,m}$, we first show $\overline{g(X)} \cap L_F = \emptyset$. If $x \in X$, then since $\mathcal{V}$ has order $\leq n$, $g(x)$ is a convex combination of at most $n+1$ of the $\boldsymbol{x}_{n+k}$, say $\boldsymbol{x}_{n+i_0}, \ldots, \boldsymbol{x}_{n+i_n}$. Since

$$\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+i_0}, \ldots, \boldsymbol{x}_{n+i_n}\}$$

is a set of at most $2n+2$ vectors in $\mathbb{R}^{2n+1}$ which is in general position, it is affinely independent; thus the affine subspaces generated by $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n\}$ (which is L_F) and $\{\boldsymbol{x}_{n+i_0}, \ldots, \boldsymbol{x}_{i_n}\}$ do not intersect (). In particular, the convex hull of $\{\boldsymbol{x}_{n+i_0}, \ldots, \boldsymbol{x}_{i_n}\}$ (which is a simplex) does not intersect L_F. There are finitely many such simplexes as x ranges over X; the union of these simplexes is a compact subset (polyhedron) P of $\mathbb{R}^{2n+1}$ containing $g(X)$ which does not meet L_F. Thus $\overline{g(X)}$ does not meet L_F.

If $\boldsymbol{y} \in P$, let $\phi_k(\boldsymbol{y})$ be the barycentric coordinate of $\boldsymbol{y}$ with respect to $\boldsymbol{x}_{n+k}$ (). Then ϕ_k is a continuous function from P to $[0,1]$, and $h_k = \phi_k \circ g$. If $\boldsymbol{y} \in \overline{g(X)}$, then since the $\phi_k(\boldsymbol{y})$ sum to 1, there is a k with $\phi_k(\boldsymbol{y}) > 0$. Fix such a k, and let $Z = \{\boldsymbol{z} \in P : \phi_k(\boldsymbol{z}) > 0\}$. Then Z is a neighborhood of $\boldsymbol{y}$, and $g^{-1}(Z) \subseteq V_k$. Thus $g \in W_{F,m}$.

This completes the proof of Theorem XIII.22.6.6.. 

XIII.22.6.14. We now turn to the proof of Theorem XIII.22.6.1.. If X contains an open subset of $\mathbb{R}^n$, it obviously has dimension n. The proof of the converse involves a somewhat delicate construction, and uses the following functions. If $\boldsymbol{a} \in \mathbb{R}^n$ and $\epsilon > 0$, let $h_{\boldsymbol{a},\epsilon}$ be the function from $\mathbb{R}^n \setminus \{\boldsymbol{a}\}$ to $\mathbb{R}^n$ which moves each point of $\mathbb{R}^n \setminus \{\boldsymbol{a}\}$ radially away from $\boldsymbol{a}$ a distance ϵ. This function has the following properties for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n \setminus \{\boldsymbol{a}\}$, as is easily verified:

(i) $h_{\boldsymbol{a},\epsilon}$ is a homeomorphism from $\mathbb{R}^n \setminus \{\boldsymbol{a}\}$ onto $\mathbb{R}^n \setminus \bar{B}_\epsilon(\boldsymbol{a})$.

(ii) $\|h_{\boldsymbol{a},\epsilon}(\boldsymbol{x}) - \boldsymbol{x}\| = \epsilon$.

(iii) $\|\boldsymbol{x} - \boldsymbol{y}\| \leq \|h_{\boldsymbol{a},\epsilon}(\boldsymbol{x}) - h_{\boldsymbol{a},\epsilon}(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\| + 2\epsilon$.

We will actually prove the following stronger result:

XIII.22.6.15. THEOREM. Let X be a subset of $\mathbb{R}^n$ with empty interior. Then X is homeomorphic to a subset Y of $\mathbb{R}^n$ with $\bar{Y} \subseteq \mathfrak{M}^n_{n-1}$. In particular, Y is nowhere dense in $\mathbb{R}^n$ and $dim(Y) = dim(X) \leq n - 1$.

PROOF: By XIII.22.2.2., any subset of $\mathbb{R}^n$ homeomorphic to X has empty interior. The complement of $\mathfrak{M}^n_{n-1}$ in $\mathbb{R}^n$ is the set of points with all rational coordinates, hence is countable. Let $\{\boldsymbol{a}_k : k \in \mathbb{N}\}$ be an enumeration of this complement. Set $Y_1 = X$, and inductively define a sequence Y_k of subsets of $\mathbb{R}^n$ and homeomorphisms $h_k : Y_k \to Y_{k+1}$ as follows. Y_k has empty interior, so there is a $\boldsymbol{b}_k \in \mathbb{R}^n \setminus Y_k$ with

$$\|\boldsymbol{b}_k - \boldsymbol{a}_k\| < \frac{3^{-k}}{4} .$$

Define h_k to be the restriction of $h_{\boldsymbol{b}_k,3^{-k}}$ to Y_k, and $Y_{k+1} = h_k(Y_k)$. We have

$$\|h_k(\boldsymbol{y}) - \boldsymbol{a}_k\| \geq \|h_k(\boldsymbol{y}) - \boldsymbol{b}_k\| - \|\boldsymbol{b}_k - \boldsymbol{a}_k\| > 3^{-k} - \frac{3^{-k}}{4} = \frac{3^{-k+1}}{4}$$

for all $\boldsymbol{y} \in Y_k$.

For each k set $f_k = h_k \circ \cdots \circ h_1$. Then f_k is a homeomorphism from X onto Y_{k+1}. By XIII.22.6.14.(ii) and the triangle inequality we have, for $j < k$ and $\boldsymbol{x} \in X$,

$$\|f_k(\boldsymbol{x}) - f_j(\boldsymbol{x})\| \leq \sum_{m=j+1}^{k} 3^{-m} < \sum_{m=j+1}^{\infty} 3^{-m} = \frac{3^{-j}}{2} .$$

Thus the f_k form a uniform Cauchy sequence and converge uniformly to a continuous function $f : X \to \mathbb{R}^n$. Set $Y = f(X)$. For $\boldsymbol{x}, \boldsymbol{z} \in X$, we have $\|f(\boldsymbol{x}) - f(\boldsymbol{z})\| \geq \|\boldsymbol{x} - \boldsymbol{z}\|$ by XIII.22.6.14.(iii), so f is injective and $f^{-1}|_Y$ is a contraction, hence continuous; thus f is a homeomorphism from X onto Y.

For any k we have, for any $\boldsymbol{x} \in X$,

$$\|f(\boldsymbol{x}) - f_k(\boldsymbol{x})\| \leq \frac{3^{-k}}{2}$$

and hence

$$\|f(\boldsymbol{x}) - \boldsymbol{a}_k\| \geq \|f_k(\boldsymbol{x}) - \boldsymbol{a}_k\| - \|f(\boldsymbol{x}) - f_k(\boldsymbol{x})\| > \frac{3^{-k+1}}{4} - \frac{3^{-k}}{2} = \frac{3^{-k}}{4} .$$

Thus $\boldsymbol{a}_k$ is not in $\bar{Y}$, i.e. $\bar{Y} \subseteq \mathfrak{M}^n_{n-1}$. 👍

XIII.22.7. Infinite-Dimensional Spaces

XIII.22.7.1. A topological space can be infinite-dimensional in any of the above senses. There are in fact examples of (normal) spaces which are infinite-dimensional in some senses but not others. For separable metrizable spaces the notion of infinite-dimensionality is unambiguous.

XIII.22.7.2. There are degrees of infinite-dimensionality, however. One might expect, for example, that $[0,1]^\kappa$ should be "κ-dimensional." A more important distinction, however, at least for metrizable spaces, is between "countable-dimensional" and "uncountable-dimensional" spaces:

XIII.22.7.3. DEFINITION. Let X be a topological space. Then X is *countable-dimensional* if it is a countable union of finite-dimensional subspaces. X is *countably infinite-dimensional* if it is countable-dimensional but not finite-dimensional.

This definition can be made for any of the dimension functions; we will be imprecise as to the meaning of "finite-dimensional" but note this is well-defined for separable metrizable spaces.

XIII.22.7.4. By (), a normal space X is countable-dimensional (in the sense of Ind) if and only if it can be written as a countable union of subspaces X_n with $Ind(X_n) = 0$ for all n. X is countably infinite-dimensional if and only if it can be written as a countable union of zero-dimensional subspaces, but not as a finite union of zero-dimensional subspaces.

XIII.22.7.5. It is easy to give examples of countably infinite-dimensional separable metrizable spaces: the disjoint union $\sqcup_n [0,1]^n$ works. For a compact metrizable example, take the one-point compactification of the disjoint union, or the infinite join if a connected compact metrizable example is desired. It is only slightly harder to string together the cubes to get a compact AR which is countably infinite-dimensional.

Another distinction comes from considering a strengthening of the partition dimension concept:

XIII.22.7.6. DEFINITION. Let X be a normal topological space. Then X is *strongly infinite-dimensional* if there is a sequence of pairs (A_k, B_k) of disjoint closed sets in X such that, if P_k is any partition between A_k and B_k for each k, then $\cap_{k=1}^\infty P_k \neq \emptyset$.

X is *weakly infinite-dimensional* if $pd(X) = \infty$ but X is not strongly infinite-dimensional, i.e. for every sequence of pairs (A_k, B_k) of disjoint closed sets in X, there are partitions P_k such that $\cap_{k=1}^\infty P_k = \emptyset$, but for each n there are pairs $(A_1, B_1), \ldots, (A_n, B_n)$ such that, if P_k is any partition between A_k and B_k for each k, then $\cap_{k=1}^n P_k \neq \emptyset$.

XIII.22.7.7. PROPOSITION. Every countably infinite-dimensional completely normal space is weakly infinite-dimensional.

PROOF: This is a simple adaptation of (). Write $X = \cup_{n=1}^\infty X_n$, where X_n is zero-dimensional. If $\{(A_n, B_n) : n \in \mathbb{N}\}$ is a sequence of pairs of disjoint closed sets in X, then by XIII.22.3.16. there is for each n a partition P_n between A_n and B_n with $P_n \cap X_n = \emptyset$. Then $\cap_{n=1}^\infty P_n = \emptyset$. 👍

On the contrary, we have:

XIII.22.7.8. THEOREM. The Hilbert cube $\mathbb{H}$ is strongly infinite-dimensional.

PROOF: (cf. [HW41, p. 49]) In $\mathbb{H} = [0,1]^{\mathbb{N}}$, let A_k be the set of all points with k'th coordinate 0, and B_k the set of points with k'th coordinate 1. Then A_k and B_k are disjoint closed sets in X. Suppose P_k is a partition between A_k and B_k for each k. For each n let $f_n : [0,1]^n \to \mathbb{H}$ be defined by

$$f_n(x_1, \dots, x_n) = (x_1, \dots, x_n, 0, 0, \dots) \ .$$

Fix n. If $\tilde{A}_k = f_n^{-1}(A_k)$, $\tilde{B}_k = f_n^{-1}(B_k)$, $\tilde{P}_k = f_n^{-1}(P_k)$ for $1 \leq k \leq n$, then $\tilde{A}_k$ and $\tilde{B}_k$ are just the opposite k-faces of $[0,1]^n$, and $\tilde{P}_k$ is a partition in $[0,1]^n$ between $\tilde{A}_k$ and $\tilde{B}_k$, so by the Face Partition Theorem (XIII.21.1.19.), $\cap_{k=1}^n \tilde{P}_k \neq \emptyset$; thus $\cap_{k=1}^n P_k \neq \emptyset$ in X. This is true for all n, so the P_k have the Finite Intersection Property, and thus $\cap_{k=1}^\infty P_k \neq \emptyset$ since the P_k are compact. 👍

XIII.22.7.9. COROLLARY. The Hilbert cube is not countable-dimensional.

Since the Hilbert cube is separable and metrizable, it is unambiguous what dimension function we use.

XIII.22.7.10. It is thus highly questionable whether we should say that $dim\left([0,1]^{\aleph_0}\right) = \aleph_0$.

There is a theory of transfinite dimension for (some) countable-dimensional spaces (Exercise XIII.22.8.9.).

XIII.22.7.11. There are many pathologies in infinite-dimensional spaces. For example, if X is n-dimensional for finite n, then X has closed subspaces of dimension k for all $0 \leq k \leq n$ (cf. the proof of XIII.22.3.19.). But there are examples of an infinite-dimensional compact metrizable space X such that every subspace of X is either zero-dimensional or infinite-dimensional (such a space is called *hereditarily infinite-dimensional*). Using the CH, an X can be constructed with the property that every uncountable subspace is infinite-dimensional (the CH is necessary; existence of such a space cannot be proved in ZFC, cf. XIII.22.3.35.). Note that every (regular) space has zero-dimensional subspaces, e.g. any countable subspace (XIII.22.3.18.).

Examples of the following types of spaces (and many more) can be found in [vM89]:

A weakly infinite-dimensional compact space which is not countable-dimensional.

A strongly infinite-dimensional compact space which is hereditarily infinite-dimensional.

A strongly infinite-dimensional totally disconnected space (XIII.22.8.7.; there can be no such compact example by XIII.13.6.7.).

All these examples are separable and metrizable.

XIII.22.7.12. Until quite recently, the structure theory of infinite-dimensional spaces lagged far behind that of finite-dimensional spaces. But there has been a lot of recent progress. See [vM01] and [Eng95] for more details.

XIII.22.8. Exercises

XIII.22.8.1. (a) Show that if X is T_1 and $dim(X) = 0$, then $ind(X) = 0$. [If U is an open neighborhood of a point p, consider the open cover $\{U, X \setminus \{p\}\}$ of X.]

(b) If X is the Sierpiński space, show that $dim(X)$ and $Ind(X)$ are vacuously 0, but $ind(X) = 1$ (XIII.13.9.1.).

XIII.22.8.2. Let X be a normal space.

(a) Use XIII.7.6.8. and Urysohn's Lemma to show that any open cover $\mathcal{U}$ of X has an open shrinking consisting of cozero sets.

(b) Show that $zdim(X) = dim(X)$.

XIII.22.8.3. [vM89] This problem gives a direct proof that $pd(B^n) \leq n$ (cf. XIII.22.1.5., XIII.22.4.13.); we actually show that if X is any compact subset of $\mathbb{R}^n$, then $pd(X) \leq n$.

(a) Show that if D is a dense subset of $\mathbb{R}$, and

$$D^{(n)} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_k \in D \text{ for at least one } k\}$$

and if A and B are disjoint compact subsets of $\mathbb{R}^n$, then there is a partition between A and B in $\mathbb{R}^n$ contained in $D^{(n)}$. [Cover A with finitely many open rectangular solids which are products of intervals with endpoints in D, whose closures are disjoint from B.]

(b) Show that if $(A_1, B_1), \ldots, (A_{n+1}, B_{n+1})$ are $n+1$ pairs of disjoint closed sets in X, then there are partitions P_k between A_k and B_k in X with $\cap_{k=1}^{n+1} P_k = \emptyset$. [Let $D_1, \ldots, D_{n+1}$ be $n+1$ disjoint dense subsets of $\mathbb{R}$, and choose $P_k \subseteq D_k^{(n)}$.]

XIII.22.8.4. Show that every topological n-manifold X with boundary has $dim(X) = n$. [Write each component of X as a countable union of subsets homeomorphic to B_n, and apply **??**.]

XIII.22.8.5. (a) [Men28] Let X be a separable metrizable space of dimension n. The *dimensional kernel* of X is

$$DK(X) = \{x \in X : lind_x(X) = n\} .$$

Show that $DK(X)$ is an F_σ and $ind(DK(X)) \geq n-1$. [Write the complement as a union of a set of dimension $\leq n-2$ and a set of dimension 0. Use the Sum Theorem and the argument of XIII.22.3.15..] Of course, $ind(DK(X)) \leq n$ by XIII.22.3.7.. For "most" X, $ind(DK(X)) = n$.

(b) X is *weakly n-dimensional* if $ind(DK(X)) = n-1$. Show that the space of () is weakly 1-dimensional.

For every n there exist weakly n-dimensional spaces. A compact metrizable space cannot be weakly n-dimensional; in fact, a weakly n-dimensional space cannot contain a compact subset of dimension n [HW41, p. 95].

XIII.22.8.6. Let X be a topological space and Y a (nonempty) subspace of X.

(a) Show that $ind(Y) = 0$ if and only if there is a base $\mathcal{B}$ for the topology of X such that $Y \cap \partial U = \emptyset$ for all $U \in \mathcal{B}$.

(b) Apply XIII.1.2.10. to show that if X is second countable and $ind(Y) = 0$, there is a countable base $\mathcal{B}$ for the topology of X such that $Y \cap \partial U = \emptyset$ for all $U \in \mathcal{B}$.

(c) Show that if X is second countable and $ind(Y) = 0$, then there is a G_δ subset Y_δ of X with $Y \subseteq Y_\delta$ and $ind(Y_\delta) = 0$. [Consider $X \setminus \cup_{U \in \mathcal{B}} \partial U$, where $\mathcal{B}$ is a countable base as in (b); cf. XIII.22.3.15..]

(d) Using the Second Decomposition Theorem XIII.22.3.20., prove the *Extension Theorem*: if X is second countable, $Y \subseteq X$, and $ind(Y) = n$, then there is a G_δ subset Y_δ of X with $Y \subseteq Y_\delta$ and $ind(Y_\delta) = n$.

XIII.22.8.7. ([?]; cf. [Eng95, 1.4.F]) Let K be the Cantor set, and $n \in \mathbb{N} \cup \{\infty\}$. Set $X = [0,1]^n$. Let $\mathcal{G}$ be the set of G_δ's in $K \times X$. Then $card(\mathcal{G}) = 2^{\aleph_0}$. Let ϕ be a bijection from K to $\mathcal{G}$. Fix a point $x_0 \in X$, and define a function $f : K \to X$ by choosing $f(t)$ to be a point such that $(t, f(t)) \notin \phi(t)$ if there is such a point and $f(t) = x_0$ otherwise. Let $Y = \{(t, f(t)) : t \in K\} \subseteq K \times X$ be the graph of f; give Y the subspace topology.

(a) Show that projection onto the first coordinate is an injective continuous function from Y onto K. Thus Y is totally disconnected. (So Y may be thought of as K with a topology stronger than the usual topology.)

(b) If G is a G_δ in $K \times X$ containing Y, then $G = \phi(t)$ for some $t \in K$, and hence G contains $\{t\} \times X$. Conclude that $ind(G) = n$.

(c) Apply the Extension Theorem (XIII.22.8.6.(d)) to conclude that $ind(Y) = n$.

Note that Y is separable and metrizable, so all dimension functions coincide on Y. This argument uses the AC.

XIII.22.8.8. Increasing and Decreasing Dimensions. Prove the following theorem:

THEOREM. (AC) Let $d \in \mathbb{N}$, and let X be a separable metrizable space of dimension d. Then for any $n \in \mathbb{N} \cup \{\infty\}$ there is a separable metrizable space Y_n of dimension n and a continuous bijection $\phi_n : Y \to X$. Thus for any such n there is a stronger separable metrizable topology on X of dimension n. The same holds if $dim(X) = 0$ and $card(X) = 2^{\aleph_0}$. Thus continuous bijections can raise or lower dimension arbitrarily, even for separable metrizable spaces.

(a) A separable metrizable space of cardinality less than $2^{\aleph_0}$ has dimension 0 (XIII.22.3.35.). Thus, if $d > 0$, X has cardinality $2^{\aleph_0}$ (XIII.7.10.6.(c)).

(b) If $0 \leq n < d$, write X as a disjoint union $X_1 \cup \cdots \cup X_{d+1}$, where each X_k has dimension 0 (XIII.22.3.20.). Expand the topology on X by making $X_1, \ldots, X_{d-n}$ clopen. This can be done concretely by taking the following subspace of $X \times \{0, 1, \ldots, d-n\}$:

$$Y_n = \{(x,0) : x \in \cup_{k>d-n} X_k\} \cup \bigcup_{1 \leq k \leq d-n} \{(x,k) : x \in X_k\} .$$

Use XIII.22.3.20. to show that $ind(Y_n) = n$.

If $X = \mathbb{R}$ and $n = 0$, we can take $[(\mathbb{R} \setminus \mathbb{Q}) \times \{0\}] \cup [\mathbb{Q} \times \{1\}] \subseteq \mathbb{R}^2$, the graph of $\chi_{\mathbb{Q}}$. One can also obtain such a topology by expanding the usual topology by making each rational number an isolated point (cf. XIII.11.14.8., which is not quite the same). This space can be realized as the following subspace of $\mathbb{R}^2$:

$$\{(x,0) : x \in \mathbb{R} \setminus \mathbb{Q}\} \cup \left\{ \left(r, \frac{1}{q} \right) : r = \frac{p}{q} \in \mathbb{Q} \right\}$$

where $\frac{p}{q}$ is in lowest terms and $q > 0$ (writing $0 = \frac{0}{1}$). This is the graph of the "Thomae function" which is continuous at the irrationals and discontinuous at the rationals. Actually the graph of any function with a dense set of jump discontinuities would do.

(c) For the case $n > d$, note that in XIII.22.8.7. K and X can be replaced in the construction by any separable metrizable spaces of cardinality $2^{\aleph_0}$. If we use $[0,1]^n$, the space Y will satisfy $n \leq ind(Y) \leq d + n$. If $ind(Y) > n$, reduce it to n using (b).

In particular, for any $n \in \mathbb{N} \cup \{\infty\}$ there is a separable metrizable topology $\mathcal{T}_n$ on $\mathbb{R}$, or $[0,1]$, stronger than the usual topology, with $dim(\mathbb{R}, \mathcal{T}_n) = n$. [In this case the Y has dimension exactly n: note that Y cannot contain any nonempty open set in $\mathbb{R} \times \mathbb{R}^n \cong \mathbb{R}^{n+1}$ by Invariance of Domain, and apply XIII.22.6.1..]

XIII.22.8.9. [] Extend the definitions of ind and Ind transfinitely: if α is an ordinal, then $ind(X) \leq \alpha$ if and only if, whenever $x \in X$ and U is an open neighborhood of x in X, there is an open neighborhood V of x contained in U with $ind(\partial V) < \alpha$, and similarly for Ind.

(a) If X is a T_1 space, and $Ind(X)$ is defined, then $ind(X)$ is defined and $ind(X) \leq Ind(X)$.

(b) If X is a separable metrizable space, and $ind(X)$ is defined, then X is countable-dimensional. Thus $ind(X)$ is not defined even for every compact metrizable space; in particular, $ind(\mathbb{H})$ is not defined. If X is compact, metrizable, and countable-dimensional, then $ind(X)$ and $Ind(X)$ are defined.

(c) If X is a separable metrizable space, and $ind(X)$ is defined, then $ind(X)$ is a countable ordinal.

(d) For every countable ordinal α, there are compact metrizable spaces X_α and Y_α with $ind(X_\alpha) = \alpha$ and $Ind(Y_\alpha) = \alpha$. There are compact metrizable spaces X with $\omega < ind(X) < Ind(X)$.

(e) If $ind_x(X)$ is finite for every $x \in X$, then $ind(X) \leq \omega$. The converse is false, even for compact metrizable X. Define a transfinite version of local dimension ind_x, and show that $ind(X) = \sup_{x \in X} ind_x(X)$ for any X.

XIII.22.8.10. Let X be a topological space.

DEFINITION. A *block decomposition* of X is a locally finite cover $\mathcal{F} = \{F_i\}$ of X by blocks (), where $F_i = \bar{U}_i$ for a pairwise disjoint set $\{U_i\}$ of open sets in X.

In some references, a block decomposition is called a *canonical covering* (perhaps since blocks are sometimes called canonically closed sets), but this is poor terminology since there is nothing canonical about a block decomposition.

(a) Show that if $\mathcal{F} = \{F_i\}$ is a block decomposition of X, then for $i \neq j$ the interiors of F_i and F_j are disjoint, i.e. F_i and F_j intersect only along their boundaries. Thus in the definition we may take the U_i to be the interiors of the F_i.

(b) If $\mathcal{F} = \{F_i\}$ and $\mathcal{G} = \{G_j\}$ are block decompositions of X, and whenever $F_i^o \cap G_j^o \neq \emptyset$ we set H_{ij} to be the closure of $F_i^o \cap G_j^o$, show that the H_{ij} form a block decomposition of X refining both $\mathcal{F}$ and $\mathcal{G}$.

(c) If $\mathcal{F} = \{F_i\}$ and $\mathcal{G} = \{G_j\}$ are block decompositions of X and $\mathcal{G}$ refines $\mathcal{F}$, then each G_j is contained in a unique F_i and each F_i is the union of the G_j contained in it.

XIII.22.8.11. Approximation Dimension. [Pon63]; cf. [Pea75] Here is an approach to dimension closer to the earlier intuitive ones. Let X be a (Hausdorff) topological space.

DEFINITION. If $X = \emptyset$, set $\delta(X) = -1$. If X is nonempty, $\delta(X)$ is the smallest n such that X has a directed regular () family $\mathcal{F}_\lambda$ of block decompositions, each of order $\leq n$, and $\delta(X) = \infty$ if there is no such n.

If $X = \emptyset$, set $\Delta(X) = -1$. If X is nonempty, $\Delta(X)$ is the smallest n such that X has a directed fine () family $\mathcal{F}_\lambda$ of block decompositions, each of order $\leq n$, and $\Delta(X) = \infty$ if there is no such n.

The functions $\delta(X)$ and $\Delta(X)$ are called the *approximation dimension* and *cofinal approximation dimension* of X respectively. We have $\delta(X) \leq \Delta(X)$ by (); $\delta(X) = \infty$ if X is not regular, and $\Delta(X) = \infty$ if X is not paracompact ().

(a) Show that for any (Hausdorff) space X, $\delta(X) = 0$ if and only if $ind(X) = 0$, and that the following are equivalent:

(i) $\Delta(X) = 0$.

(ii) X is paracompact and $dim(X) = 0$.

(iii) Every open cover of X has a refinement consisting of a partition into clopen sets (X is *perfectly zero-dimensional*).

(b) Show that if X is a topological space and $n \in \mathbb{N} \cup \{0\}$, then $\delta(X) \leq n$ if and only if there is a space Z with $ind(Z) = 0$ and a continuous closed surjective map $f : Z \to X$ such that $|f^{-1}(\{x\})| \leq n + 1$ for all $x \in X$.

(c) Show that if X is a regular space and $n \in \mathbb{N} \cup \{0\}$, then $\Delta(X) \leq n$ if and only if there is a perfectly zero-dimensional space Z and a continuous closed surjective map $f : Z \to X$ such that $|f^{-1}(\{x\})| \leq n + 1$ for all $x \in X$.

(d) Show that $ind(X) \leq \delta(X)$ for any topological space X, and $Ind(X) \leq \Delta(X)$ for any regular space X.

(e) Show that if X is metrizable, then $\Delta(X) = dim(X) = Ind(X)$. If X is separable and metrizable, then $\delta(X) = ind(X) = dim(X) = Ind(X) = \Delta(X)$, i.e. δ and Δ coincide with all other dimension functions we have considered.

XIII.22.8.12. (a) Let X be a normal space, and A a closed subspace. Use XIII.22.1.8. to prove that

$$dim(X) \leq \max(dim(A), dim(X \setminus A)) \ .$$

(b) Conclude that if X is separable metric and A is a closed subspace, then

$$dim(X) = \max(dim(A), dim(X \setminus A)) \ .$$

[Use that $ind = dim$ for separable metric spaces, and conclude the opposite inequality from XIII.22.3.7..]

(c) Conclude that if X and Y are separable metric, A is a closed subspace of X, and f a quotient mapping from A to a subset of Y, then the adjunction space $X \cup_f Y$ () satisfies

$$dim(X \cup_f Y) = \max(dim(Y), dim(X \setminus A)) \leq \max(dim(X), dim(Y)) \ .$$

In particular, if X is separable metric and CX is the cone over X, then

$$dim(CX) = dim(X \times [0, 1)) \leq dim(X) + 1 \ .$$

(d) Show that if X is normal and A and B are arbitrary subspaces with $X = A \cup B$, then

$$dim(X) \leq dim(A) + dim(B) + 1 \ .$$

Equality can hold (e.g. $\mathbb{R}$ is a union of two zero-dimensional subspaces).

See [Pea75, 5.8] for (a) and [Pea75, 5.11] for (d).

XIII.22.9. Fubini's Differentiation Theorem

Although even uniformly convergent infinite series of functions cannot be differentiated term-by-term in general, there are a few types of series for which term-by-term differentiation is valid. The most notable example is convergent power series (). Another example is convergent series of nondecreasing functions, the topic of this subsection.

The main theorem is due to Fubini. The name "Fubini's Theorem" is, however, normally reserved for another of Fubini's results on interchanging the order of multiple integration (). This result is sometimes called "Fubini's Differentiation Theorem," a name we will use. This theorem can be stated for sequences (as can any result about infinite series, since the theories of sequences and series are formally equivalent), but the hypotheses are clumsy to state in the sequence setting (cf. Exercise XIII.22.10.1.).

XIII.22.9.1. Theorem. [Fubini's Differentiation Theorem] Let $\sum_{k=1}^{\infty} f_k(x)$ be an infinite series of nondecreasing functions on an interval $[a,b]$, converging pointwise to a (nondecreasing) function f. Then

(i) The series converges uniformly to f on $[a,b]$.

(ii) The series $\sum_{k=1}^{\infty} f_k'(x)$ converges pointwise a.e. on $[a,b]$ to $f'(x)$.

In particular, $f_k'(x) \to 0$ for almost all $x \in [a,b]$.

(Conclusion (ii) is the main part of the theorem; (i) is an easy observation.)

Proof: Begin with (i). For any $x \in [a,b]$, we have

$$f(x) = [f(x) - f(a)] + f(a) = \sum_{k=1}^{\infty} [f_k(x) - f_k(a)] + \sum_{k=1}^{\infty} f_k(a)$$

with both series convergent. If $\epsilon > 0$, choose N such that for all $n \geq N$,

$$\left| \sum_{k=n+1}^{\infty} [f_k(b) - f_k(a)] \right| < \frac{\epsilon}{2} \quad \text{and} \quad \left| \sum_{k=n+1}^{\infty} f_k(a) \right| < \frac{\epsilon}{2} .$$

Then, for any $n \geq N$ and any $x \in [a,b]$, we have

$$\left| f(x) - \sum_{k=1}^{n} f_k(x) \right| = \left| \sum_{k=n+1}^{\infty} [f_k(x) - f_k(a)] + \sum_{k=n+1}^{\infty} f_k(a) \right|$$

$$\leq \left| \sum_{k=n+1}^{\infty} [f_k(x) - f_k(a)] \right| + \left| \sum_{k=n+1}^{\infty} f_k(a) \right| \leq \left| \sum_{k=n+1}^{\infty} [f_k(b) - f_k(a)] \right| + \left| \sum_{k=n+1}^{\infty} f_k(a) \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

(the middle inequality on the last line holds because f_k is nondecreasing) and thus the series converges uniformly to f on $[a,b]$.

We now turn to (ii). First assume all the f_k are nonnegative. There is a subset A of $[a,b]$ whose complement has measure 0 and such that f and all the f_k are differentiable at all $x \in A$ (). Let $s_n(x) = \sum_{k=1}^{n} f_k(x)$ be the n'th partial sum and $r_n(x) = f(x) - s_n(x) = \sum_{k=n+1}^{\infty} f_k(x)$ the remainder. Then s_n and r_n are nonnegative, nondecreasing, and differentiable on A, with $s_n'(x) \geq 0$ and $r_n'(x) \geq 0$ for all $x \in A$. Since $s_n'(x) \leq s_n'(x) + r_n'(x) = f'(x)$ for all $x \in A$, we have that $\sum_{k=1}^{\infty} f_k'(x)$ converges for all $x \in A$ since it is a nonnegative series with bounded partial sums, and

$$\sum_{k=1}^{\infty} f_k'(x) \leq f'(x)$$

for all $x \in A$.

To show that equality holds a.e. on $[a,b]$, it suffices to find a subsequence (s_{n_j}') which converges to f' a.e. on $[a,b]$. Since $s_n(b) \to f(b)$, for each j there is an $n_j > n_{j-1}$ such that $r_{n_j}(b) = f(b) - s_{n_j}(b) < 2^{-j}$ and hence, for each $x \in [a,b]$, we have

$$0 \leq \sum_{j=1}^{\infty} r_{n_j}(x) \leq \sum_{j=1}^{\infty} r_{n_j}(b) = \sum_{j=1}^{\infty} [f(b) - s_{n_j}(b)] < \infty$$

and thus the series $\sum_{j=1}^{\infty} r_{n_j}(x)$ satisfies the hypotheses in the preceding paragraph, so there is a subset B of $[a,b]$ such that $[a,b] \setminus B$ has measure 0 and

$$0 \leq \sum_{j=1}^{\infty} r_{n_j}'(x) = \sum_{j=1}^{\infty} [f'(x) - s_{n_j}'(x)] < \infty$$

for all $x \in B$. Thus, for every $x \in B$, we have $\lim_{j\to\infty} [f'(x) - s_{n_j}'(x)] = 0$, so

$$\lim_{j\to\infty} s_{n_j}'(x) = f'(x)$$

for almost all $x \in [a,b]$.

Now drop the assumption that the f_k are nonnegative. (The only place this assumption was used was to obtain that the r_n were nonnegative.) Set $g(x) = f(x) - f(a)$ and $g_k(x) = f_k(x) - f_k(a)$ for each k and each $x \in [a,b]$. Then the g_k are nonnegative and nondecreasing, $g' = f'$, $g_k' = f_k'$ for all k, and $g(x) = \sum_{k=1}^{\infty} g_k(x)$ for all k. Thus the previous argument applies to g and the g_k, and shows that

$$f'(x) = g'(x) = \sum_{k=1}^{\infty} g_k'(x) = \sum_{k=1}^{\infty} f_k'(x) \text{ a.e.}$$

XIII.22.10. Exercises

XIII.22.10.1. Give a correct restatement of the Fubini Differentiation Theorem for sequences. Show that the following is *not* a correct statement:

"Let (f_n) be a nondecreasing sequence of nondecreasing functions on an interval $[a,b]$ converging uniformly on $[a,b]$ to a (nondecreasing) function f. Then $f_n' \to f'$ pointwise a.e. on $[a,b]$."

[Consider a strictly increasing sequence (f_n) of nondecreasing step functions converging uniformly to $f(x) = x$ on $[0,1]$. Modify this example to make the f_n continuous by connecting the steps by line segments of large slope. In these examples, there is no $x \in [0,1]$ for which $f_n'(x) \to f'(x)$.]

XIII.22.10.2. Let g be the Cantor function (). Regard g as a function on $\mathbb{R}$ by setting $g(x) = 0$ for $x < 0$ and $g(x) = 1$ for $x > 1$. Then g is nonnegative, continuous, nondecreasing, $g(0) = 0$, $g(1) = 1$, and $g' = 0$ a.e. Let $\{[a_k, b_k] : k \in \mathbb{N}\}$ be an enumeration of the closed bounded intervals with rational endpoints $a_k < b_k$.

(a) For each k let f_k be a scaled copy of g transferred to $[a_k, b_k]$:

$$f_k(x) = 2^{-k} f\left(\frac{x - a_k}{b_k - a_k}\right) .$$

Show that f_k is nonnegative, continuous, nondecreasing, $f_k(x) = 0$ for $x \le a_k$, $f_k(x) = 2^{-k}$ for $x \ge b_k$, and $f_k' = 0$ a.e.

(b) Set $f(x) = \sum_{k=1}^{\infty} f_k(x)$. Show that the series converges uniformly on all of $\mathbb{R}$, and that f is continuous and strictly increasing on $\mathbb{R}$.

(c) Apply XIII.22.9.1. to conclude that $f' = 0$ a.e.

XIII.22.10.3. [Move to section on series of functions]

(a) Let $\{a_k : k \in \mathbb{N}\}$ be an enumeration of $\mathbb{Q} \cap [0,1]$. Show that the infinite series $\sum_{k=1}^{\infty} \frac{3}{k!} \sqrt[3]{x - a_k}$ converges uniformly on $[0,1]$ to a function g which is continuous and strictly increasing.

(b) Let $[a,b]$ be the range of g, and $f = g^{-1} : [a,b] \to [0,1]$. Show that f is continuous, strictly increasing, and that $f' = 0$ on the dense set $g(\mathbb{Q})$.

(c) Show that g is differentiable at every irrational number x in $[0,1]$ and that $g'(x) > 1$. Thus f is differentiable at every point of $[a,b]$ and $0 \le f'(x) < 1$ for all $x \in [a,b]$.

This example is due to D. Pompeiu ([**?**]; cf. [vRS82, 13.3] or [Str81, p. 215-216]). For this f, the set $\{x \in [a,b] : f'(x) = 0\}$, while dense in $[a,b]$, is countable. There are more dramatic examples, e.g. there is a strictly increasing continuous function f on $[a,b]$ such that $f' = 0$ a.e. (Exercise XIII.22.10.2.).

XIII.22.10.4. Let $K \subseteq [0,1]$ be a Cantor set of positive measure, and let g be a nonnegative real-valued continuous function on $[0,1]$ which is zero precisely on K, e.g. $g(x)$ is the distance from x to K. Set

$$f(x) = \int_0^x g(t)\, dt .$$

Then f is strictly increasing on $[0,1]$, differentiable everywhere with $f' = g$ (), i.e. $\mathcal{C}^1$, and $f' = 0$ on a set of positive measure.

XIII.22.10.5. Does there exist a strictly increasing differentiable function f on an interval $[a, b]$ such that the set

$$S = \{x \in [a, b] : f'(x) = 0\}$$

satisfies that $S \cap [c, d]$ has positive measure for all $a \leq c < d \leq b$? (Compare XIII.22.10.3., XIII.22.10.4..)

XIII.22.10.6. Let $f_k(x) = \frac{1}{k} \tan^{-1} kx - \frac{1}{k+1} \tan^{-1}(k+1)x$.

(a) Show that f_k is strictly increasing on $[-1, 1]$.

(b) Show that $\sum_{k=1}^{\infty} f_k(x)$ converges (uniformly) on $[-1, 1]$ to a differentiable function f.

(c) Although f and each f_k are differentiable everywhere, it is not true that $f'(x) = \sum_{k=1}^{\infty} f_k'(x)$ everywhere (we do not have $f'(0) = \sum_{k=1}^{\infty} f_k'(0)$).

XIII.23. Sard's Theorem

Sard's Theorem in its various forms is one of the most remarkable properties of Euclidean space with its differentiable structure, and is perhaps *the* fundamental result of the subject of differential topology.

Consider a smooth map f from $\mathbb{R}^n$ to $\mathbb{R}^m$, or, more generally, from a smooth n-manifold N to a smooth m-manifold M. Then, for any $\boldsymbol{x}_0 \in \mathbb{R}^n$, $Df(\boldsymbol{x}_0)$ is a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. The point $\boldsymbol{x}_0$ is *singular* if $Df(\boldsymbol{x}_0)$ has rank strictly smaller than m; otherwise, $\boldsymbol{x}_0$ is *regular*. The image in M of the set of singular points of f is called the set of *singular values* of f; a $\boldsymbol{y}_0$ in the range of f which is not a singular value is called a *regular value* of f. (The image of a regular point can be a singular value if it is also the image of a singular point.) The names reflect the fact that f is well-behaved in a neighborhood of a regular point but can be quite pathological near a singular point (cf. ()). Points of M not in the range of f are also (rather perversely) usually called regular values of f. Every critical value of f is a singular value; the converse is not necessarily true if $m > 1$.

Although the set of singular points of f can be a large subset of $\mathbb{R}^n$ (f could even be a constant function!), Sard's Theorem states that the set of singular *values* of f is a "negligibly small" subset of $\mathbb{R}^m$ (or M), in a certain precise sense (measure zero and meager). Thus, in many situations, it can be concluded that "almost all" values of f are regular values; even the existence of one regular value for f can sometimes have important consequences (cf. XIII.21.5.6.).

Sard's Theorem is intuitively quite plausible, and relatively easy to prove, if $n \leq m$, and only quite mild differentiability assumptions on f are needed. In the case $n < m$, where every point of N is a singular point (since the rank of $Df(\boldsymbol{x}_0)$ is always $\leq n$ for any $\boldsymbol{x}_0$), it simply says that the range of f is a "small" subset of M (intuitively it has dimension $\leq n$); although the existence of "space-filling curves" () can shake one's confidence about this, Sard's Theorem in the case $n < m$ says that there can be no smooth space-filling curve or lower-dimensional surface. And in the case $n = m = 1$, the statement says that for a continuously differentiable function, "most" horizontal lines are not tangent to the graph.

In the case $n > m$, or $n = m > 1$, the result is also intuitively plausible: if $\boldsymbol{x}_0$ is a singular point, then near $\boldsymbol{x}_0$ the function f is approximately a linear function whose range is a proper subspace of $\mathbb{R}^m$, and such a subspace is negligibly small in a measure or category sense, and thus a small neighborhood of $\boldsymbol{x}_0$ should have an image of "much smaller" size in $\mathbb{R}^m$. However, this intuitive idea turns out not to be directly translatable into a proof in this case, and is even false in general unless some delicate smoothness conditions are assumed (the underlying intuitive reason for the smoothness conditions is closely related to V.10.7.11.). The fact that the conclusion holds in the case $n > m$ is rather subtle and quite remarkable, and has all sorts of important consequences, e.g. it easily implies the Brouwer Fixed-Point Theorem, which is the most fundamental fact about the topology of Euclidean space, as well as underlying transversality, the fundamental technical tool of differential topology.

Sard's Theorem is often called the *Morse-Sard Theorem* since the case $m = 1$ was proved by A. P. Morse in 1939 [Mor39]; the general theorem was proved by A. Sard in 1942 [Sar42]. Sard, a student of Marston Morse, obtained some of the results independently of A. P. Morse in his Ph.D. dissertation of 1936. Note that A. P. Morse was not the same person as Marston Morse (they were apparently unrelated). Sard states in [Sar42] that Marston Morse previously proved the theorem of A. P. Morse in the cases $n \leq 6$, in part jointly with Sard (unpublished). A closely related result was obtained in [Bro35]. There have been various generalizations and refinements, cf. [Fed69], [Yom83], [Bat93].

XIII.23.1. Sets of Measure Zero

The notion of a set of measure zero is really a concept from measure theory (). But little if any actual measure theory is needed to develop the definition and basic properties needed in this section.

XIII.23.1.1. DEFINITION. A subset $A \subseteq \mathbb{R}^n$ *has measure zero in* $\mathbb{R}^n$ if, for every $\epsilon > 0$, there is a sequence (B_k) of open balls in $\mathbb{R}^n$ such that $A \subseteq \cup_k B_k$ and

$$\sum_{k=1}^{\infty} \rho_k^n < \epsilon$$

where ρ_k is the radius of B_k. We allow empty balls, i.e. open balls of radius 0; thus the collection $\{B_k\}$ may be finite or countably infinite.

Note that up to a constant factor, the sum $\sum_{k=1}^{\infty} \rho_k^n$ is just the sum of the n-dimensional volumes $\sum_{k=1}^{\infty} V_n(B_k)$ of the B_k. By adjusting ϵ, the sum of the volumes can be used instead in the definition.

XIII.23.1.2. We should properly say "A has n-dimensional Lebesgue measure zero" instead of "A has measure zero in $\mathbb{R}^n$." Note that the "*in* $\mathbb{R}^n$" is a crucial part of the terminology: it does not make sense to simply say that a subset of Euclidean space has measure zero without specifying which Euclidean space it is with respect to. For example, the interval $[0, 1]$ does not have measure zero in $\mathbb{R}^1$, but if it is regarded as a segment on the x-axis it does have measure zero in $\mathbb{R}^2$. The term *null set* is often used for a set of measure zero.

XIII.23.1.3. A ball of radius r in $\mathbb{R}^n$ is contained in a cube of side length $2r$ with sides parallel to the coordinate axes; conversely, a cube of side length $2s$ is contained in the ball of radius $s\sqrt{n}$ around its center. So the definition can be equivalently rephrased replacing "open ball" with "open cube" and "radius" with "side length." The cubes can be restricted to ones with sides parallel to the axes, or not.

An open ball intersecting A is contained in an open ball of twice the radius centered at a point of A. Thus all the balls in the definition may be taken with centers in A. The same consideration can be made with cubes.

There is much more flexibility. If B is a ball, there is a ball with rational radius with the same center whose radius is less than twice the radius of B; thus we may restrict to using open balls with rational radius. Similarly, open cubes with rational edges can be used. In a similar manner, we may restrict to using only open balls or cubes whose centers have rational coordinates (and rational radii or sides).

Another especially useful variation is to use open rectangular solids. For if R is an open rectangular solid, R can be covered by a finite number of open cubes whose total volume is less than $2V_n(R)$. Thus we obtain:

XIII.23.1.4. PROPOSITION. Let $A \subset \mathbb{R}^n$. Then A has measure zero in $\mathbb{R}^n$ if and only if, for every $\epsilon > 0$, there is a sequence (R_k) of open rectangular solids in $\mathbb{R}^n$ with $A \subseteq \cup_k R_k$ and

$$\sum_{k=1}^{\infty} V_n(R_k) < \epsilon \ .$$

As before, some of the R_k may be empty, i.e. there may be either a finite or countably infinite number of R_k.

XIII.23.1.5. PROPOSITION. (i) Any subset of a subset of $\mathbb{R}^n$ of measure zero in $\mathbb{R}^n$ has measure zero in $\mathbb{R}^n$.

(ii) A countable union of subsets of $\mathbb{R}^n$ of measure zero in $\mathbb{R}^n$ has measure zero in $\mathbb{R}^n$.

PROOF: Part (i) is obvious. For (ii), let (A_j) be a sequence of subsets of $\mathbb{R}^n$ of measure zero in $\mathbb{R}^n$, and $A = \cup_j A_j$. Let $\epsilon > 0$. For each j let (B_{jk}) be a sequence of open balls in $\mathbb{R}^n$ covering A_j for which the sum of the volumes is less than $2^{-j}\epsilon$. Then the B_{jk} cover A and the sum of the volumes is less than ϵ. 👍

XIII.23.1.6. COROLLARY. Let $A \subseteq \mathbb{R}^n$. Then A has measure zero in $\mathbb{R}^n$ if and only if every $\boldsymbol{x}_0 \in A$ has a neighborhood U such that $A \cap U$ has measure zero in $\mathbb{R}^n$.

PROOF: This is really the Lindelöf property of $\mathbb{R}^n$. If $\boldsymbol{x}_0 \in A$ and U is any neighborhood of $\boldsymbol{x}_0$, then there is a ball V of rational radius, whose center has rational coordinates, with $\boldsymbol{x}_0 \in V \subseteq U$. If $A \cap U$ has measure zero in $\mathbb{R}^n$, so does $A \cap V$. There are only countably many distinct such V, so A is a countable union of sets of measure zero in $\mathbb{R}^n$. The converse is trivial. 👍

XIII.23.1.7. EXAMPLES. (i) Let S be the $(n-1)$-dimensional subspace

$$\{(x_1, \ldots, x_{n-1}, 0) : x_1, \ldots, x_{n-1} \in \mathbb{R}\}$$

of $\mathbb{R}^n$. Then S is a countable union of bounded subsets. Each bounded subset can be contained in a single open rectangular solid whose thickness is so small that its volume is arbitrarily small. Thus S has measure zero in $\mathbb{R}^n$.

(ii) No (nonempty) open rectangle in $\mathbb{R}^n$ can have measure zero in $\mathbb{R}^n$. This fact, while intuitively quite plausible, is somewhat tricky to prove; see (). Since any nonempty open set in $\mathbb{R}^n$ contains an open rectangle, no nonempty open set can have measure zero, i.e. every set of measure zero in $\mathbb{R}^n$ has empty interior.

XIII.23.1.8. PROPOSITION. Let U be a subset of $\mathbb{R}^n$, $\boldsymbol{f} : U \to \mathbb{R}^n$ a Lipschitz function (). If $A \subseteq U$ has measure zero in $\mathbb{R}^n$, so does $\boldsymbol{f}(A)$. In particular, isometries and affine functions preserve sets of measure zero.

PROOF: Suppose $A \subseteq U$ has measure zero in $\mathbb{R}^n$. Let M be the Lipschitz constant of $\boldsymbol{f}$. Let $\epsilon > 0$, and let (B_k) be a sequence of open balls with centers $\boldsymbol{x}_k \in A$ covering A with total volume less than $\frac{\epsilon}{M^n}$. If ρ_k is the radius of B_k, then $\boldsymbol{f}(B_k \cap U)$ is contained in the open ball of radius $M\rho_k$ around $\boldsymbol{f}(\boldsymbol{x}_k)$; thus $\boldsymbol{f}(A)$ is contained in the union of these open balls, which have total volume $< \epsilon$. 👍

XIII.23.1.9. COROLLARY. Let U and V be open subsets of $\mathbb{R}^n$, and $\boldsymbol{f}$ a $(\mathcal{C}^1)$ diffeomorphism from U to V. If $A \subseteq U$, then A has measure zero in $\mathbb{R}^n$ if and only if $\boldsymbol{f}(A)$ does.

PROOF: Every point of U has a compact neighborhood V contained in U, on which $\boldsymbol{f}$ is Lipschitz (); thus if $A \cap V$ has measure zero in $\mathbb{R}^n$, so does $\boldsymbol{f}(A \cap V)$. 👍

XIII.23.1.10. Sets of measure zero are not preserved under homeomorphisms in general (). Thus having measure zero is a metric concept but not a topological concept.

The next result is a special case of Tonelli's Theorem (), needed for the proof of Sard's Theorem. This proof appears in [Ste83, p. 51] and is attributed to H. FURSTENBERG. An F_σ is a countable union of closed sets. Since every closed set in $\mathbb{R}^n$ is a countable union of compact sets, a subset of $\mathbb{R}^n$ is an F_σ if and only if it is σ-compact. Of course, every closed set in $\mathbb{R}^n$ is an F_σ. See ().

XIII.23.1.11. THEOREM. [BABY TONELLI THEOREM] Let K be an F_σ in $\mathbb{R}^n$. Suppose, for each $t \in \mathbb{R}$, we have that

$$K_t := \{(x_1, \ldots, x_{n-1}) : (x_1, \ldots, x_{n-1}, t) \in K\} \subseteq \mathbb{R}^{n-1}$$

has measure zero in $\mathbb{R}^{n-1}$. Then K has measure zero in $\mathbb{R}^n$.

Informally, if each horizontal cross-section of K has $(n-1)$-dimensional measure zero, then K has n-dimensional measure zero.

We need two lemmas for the proof. The first is a variation of () (note that this result is not quite correctly stated in [Ste83]):

XIII.23.1.12. LEMMA. Let $[a, b]$ be a closed bounded interval in $\mathbb{R}$ $c > 0$, and $\mathcal{U}$ a cover of $[a, b]$ by open intervals of length $\leq c$. Then there is a finite subcover $\{I_1, \ldots, I_n\}$ of $\mathcal{U}$ for which

$$\sum_{k=1}^{n} \ell(I_k) \leq 2(b - a + c) \ .$$

PROOF: Since $[a, b]$ is compact there is a finite subcover. Any finite subcover contains a minimal subcover $\mathcal{V}$ (i.e. $\mathcal{V}$ is a cover but no proper subset of $\mathcal{V}$ is a cover). Number the intervals in $\mathcal{V}$ in increasing order of left endpoint (no two can have the same left endpoint by minimality), i.e. $\mathcal{V} = \{I_1, \ldots, I_n\}$ where $I_k = (a_k, b_k)$ with $a_k < a_{k+1}$ for $1 \leq k < n$. These endpoints necessarily satisfy

$$a_1 < a \leq a_2 < b_1 \leq a_3 < b_2 \leq \cdots \leq a_n < b_{n-1} \leq b < b_n$$

(if one of these inequalities is violated, the collection of intervals is not minimal). We then have

$$\sum_{k=1}^{n} \ell(I_k) = \sum_{k=1}^{n} (b_k - a_k) = (a - a_1) + (b - a) + (b_n - b) + \sum_{k=1}^{n-1} (b_k - a_{k+1}) \ .$$

The last sum is the total length of the overlaps of the intervals; since the overlaps are disjoint and contained in $[a, b]$, this total length is $\leq b - a$. Also, $a - a_1 \leq c$ and $b_n - b \leq c$. Thus we have

$$\sum_{k=1}^{n} \ell(I_k) \leq c + (b - a) + c + (b - a) = 2(b - a + c) \ .$$



XIII.23.1.13. LEMMA. Let K be a compact subset of $\mathbb{R}^n \cong \mathbb{R}^{n-1} \times \mathbb{R}$, and $t \in \mathbb{R}$. Set

$$K_t := \{(x_1, \ldots, x_{n-1}) : (x_1, \ldots, x_{n-1}, t) \in K\} \subseteq \mathbb{R}^{n-1} \ .$$

If U is an open set in $\mathbb{R}^{n-1}$ containing K_t, then there is a $\delta > 0$ such that

$$K \cap \left[\mathbb{R}^{n-1} \times (t-\delta, t+\delta)\right] \subseteq U \times (t-\delta, t+\delta) \ .$$

PROOF: Suppose there is no such δ. Then there is a sequence $(\boldsymbol{x}_k, t_k)$ of points of K with $t_k \to t$ and $\boldsymbol{x}_k \notin U$ for all k. Passing to a subsequence, we may assume $(\boldsymbol{x}_k, t_k) \to (\boldsymbol{x}, t) \in K$ for some $\boldsymbol{x}$ by compactness of K. But then $\boldsymbol{x} \in K_t$, and $\boldsymbol{x}_k \to \boldsymbol{x}$ in $\mathbb{R}^{n-1}$, contradicting that U is an open neighborhood of $\boldsymbol{x}$. 👍

XIII.23.1.14. We now prove Theorem XIII.23.1.11.. It suffices to assume that K is compact. Then there is an interval $[a, b]$ such that $K \subseteq \mathbb{R}^{n-1} \times [a, b]$. Fix $\epsilon > 0$. For each $t \in [a, b]$ there is a countable (even finite) collection $\mathcal{R}_t$ of open rectangular solids in $\mathbb{R}^{n-1}$ covering K_t whose total $(n-1)$-dimensional volume is

$$v_t \leq \frac{\epsilon}{2(b-a+1)} \ .$$

Let U_t be the union of the rectangular solids in $\mathcal{R}_t$. By XIII.23.1.13. there is a $\delta_t > 0$ such that

$$K \cap \left[\mathbb{R}^{n-1} \times (t-\delta_t, t+\delta_t)\right] \subseteq U_t \times (t-\delta_t, t+\delta_t) \ .$$

We may assume $\delta_t \leq 1/2$ for all t. The intervals $\{(t-\delta_t, t+\delta_t) : t \in [a, b]\}$ form an open cover of $[a, b]$, so by Lemma XIII.23.1.12. there is a subcover

$$\{(t_1 - \delta_{t_1}, t_1 + \delta_{t_1}), \ldots, (t_m - \delta_{t_m}, t_m + \delta_{t_m})\}$$

whose total length $\sum_{k=1}^m 2\delta_{t_k}$ is $\leq 2(b-a+1)$. If $\mathcal{R}$ is the collection of open rectangular solids of the form

$$R \times (t_k - \delta_{t_k}, t_k + \delta_{t_k})$$

for $R \in \mathcal{R}_{t_k}$, $1 \leq k \leq m$, then the collection $\mathcal{R}$ covers K and the total n-dimensional volume is

$$\sum_{k=1}^m 2\delta_{t_k} v_{t_k} \leq \frac{\epsilon}{2(b-a+1)} \sum_{k=1}^m 2\delta_{t_k} \leq \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, K has measure zero in $\mathbb{R}^n$.
This completes the proof of Theorem XIII.23.1.11.. 👍

XIII.23.1.15. This theorem extends (XVIII.3.1.12.) from F_σ's to a much larger class of subsets of $\mathbb{R}^n$ (Lebesgue measurable subsets), but not to arbitrary subsets. See XVIII.3.3.1..

Sets of Measure Zero in a Manifold

We can extend the notion of subset of measure zero to smooth manifolds:

XIII.23.1.16. Definition. Let M be a $\mathcal{C}^r$ n-manifold ($r \geq 1$). A subset A of M *has measure zero in* M if, for every $x \in A$, there is a coordinate chart (U, ϕ) for M with $x \in U$ such that $\phi(A \cap U)$ has measure zero in $\mathbb{R}^n$.

XIII.23.1.17. It follows easily from XIII.23.1.9. that if $A \subseteq M$ has measure zero in M, and (U, ϕ) is any coordinate chart for M, then $\phi(A \cap U)$ has measure zero in $\mathbb{R}^n$. This could have been taken as the definition of a subset of measure zero, but the condition in XIII.23.1.16. is easier to check.

If M is not second countable, what we have technically defined is a locally null subset of M. (This technicality may be safely ignored as irrelevant in our applications.)

There is no analogous reasonable definition of a subset of measure zero in a topological manifold ().

The next result is a straightforward corollary of XIII.23.1.8.:

XIII.23.1.18. Proposition. Let N and M be $\mathcal{C}^r$ n-manifolds ($r \geq 1$), with N separable, and $f : N \to M$ a $\mathcal{C}^r$ function. If A is a subset of N which has measure zero in N, then $f(A)$ has measure zero in M.

XIII.23.1.19. This result holds also if N is n-dimensional and M is m-dimensional with $n < m$ (), but not necessarily if $m < n$.

XIII.23.2. Statement of Sard's Theorem

XIII.23.2.1. Theorem. [Sard's Theorem] Let N be a separable $\mathcal{C}^r$ n-manifold, and M a $\mathcal{C}^r$ m-manifold, with $r \geq \max(n - m + 1, 1)$, and let $f : N \to M$ be a $\mathcal{C}^r$ function. Then the set of singular values of f is a subset of M of measure zero in M.

XIII.23.2.2. We will prove special cases of the theorem separately before turning to the most general version. The special cases are written separately because the proof simplifies in these cases and sometimes a more general result holds. The cases we discuss are:

(i) $n < m$.

(ii) $n = m = 1$.

(iii) $n = m > 1$.

(iv) $n > m$, r large.

(v) $n > m$, sharp r bound.

Our proof is adapted from [Ste83] and [Hir94].

XIII.23.2.3. The first reduction in all cases is to note that N can be covered by countably many coordinate charts, each of which is contained in the preimage of a coordinate chart in M, and since the union of countably many null sets is a null set and null sets are preserved under diffeomorphisms, it suffices to prove the theorem for $N = \mathbb{R}^n$ and $M = \mathbb{R}^m$. In fact, since N is σ-compact, it suffices to prove the theorem for the set of singular values in the image of an arbitrary compact subset of $\mathbb{R}^n$.

XIII.23.2.4. The restriction that N be separable, i.e. that it has only countably many components, is necessary in general: if N has $2^{\aleph_0}$ components, a function from N to M can be surjective even if it is constant on each component of N; such a function is $\mathcal{C}^\infty$ and has identically zero derivative, and hence every point of M is a singular value.

XIII.23.3. The Case $n < m$

XIII.23.4. The Case $n = m = 1$

XIII.23.4.1. The simplest case is $n = m = 1$. Suppose f is a $\mathcal{C}^1$ function from $\mathbb{R}$ (or an open interval in $\mathbb{R}$) to $\mathbb{R}$. The singular points of f are just the critical numbers (the x_0 for which $f'(x_0) = 0$). It suffices to show that the image of the critical numbers in any closed bounded subinterval of the domain has measure zero in $\mathbb{R}$.

So suppose $[a, b]$ is such an interval. Since f' is continuous on $[a, b]$, it is uniformly continuous there (). Fix $\epsilon > 0$. Then there is a $\delta > 0$ such that $|f'(x) - f'(y)| < \epsilon$ whenever $x, y \in [a, b]$, $|x - y| < \delta$. Partition $[a, b]$ into subintervals $I_k = [x_{k-1}, x_k]$ with $\ell(I_k) = x_k - x_{k-1} < \delta$ for all k. For every I_k containing a singular point of f, we have $|f'(x)| < \epsilon$ for all $x \in I_k$. Thus by the Mean Value Theorem we have $|f(x) - f(y)| < \epsilon\ell(I_k)$ for all $x, y \in I_k$, for any such k. Thus $f(I_k)$ is contained in an interval J_k of length $< \epsilon\ell(I_k)$ for any such k (actually $f(I_k)$ is itself such an interval). Thus the set of singular values of f in $[a, b]$ is contained in a finite number of intervals of total length

$$\sum \ell(J_k) < \sum_{k=1}^{n} \epsilon\ell(I_k) = \epsilon(b - a)$$

and since $\epsilon > 0$ is arbitrary, this set of singular values has measure zero (in fact zero content) in $\mathbb{R}$.

This completes the proof in the case $n = m = 1$. 👍

The $\mathcal{C}^1$ assumption on f can be weakened and even eliminated in this case. The following result is due to Luzin (XIX.5.1.14.); we give an alternate proof due to S. Ivanov (http://mathoverflow.net/questions/113991/counterexample-to-sard-theorem-for-a-not-c1-map/114000#114000):

XIII.23.4.2. Theorem. Let f be a function from $\mathbb{R}$ (or an interval in $\mathbb{R}$) to $\mathbb{R}$. Set

$$\mathcal{C} = \{x \in \mathbb{R} : f'(x) \text{ exists and equals } 0\}$$

and let $\mathcal{S} = f(\mathcal{C})$. Then $\mathcal{S}$ has measure zero in $\mathbb{R}$.

Proof: We will use the Vitali Covering Theorem (). It suffices to show that the part of $\mathcal{C}$ in (a, b) has image of measure zero, for any bounded open interval (a, b) in the domain of f, so we may assume the domain of f is (a, b) and hence that $\mathcal{C}$ has finite outer measure. Fix $\epsilon > 0$. For every $x_0 \in \mathcal{C}$, there is a δ_0 such that $(x_0 - \delta_0, x_0 + \delta_0) \subseteq (a, b)$ and

$$\left|\frac{f(x) - f(x_0)}{x - x_0}\right| < \frac{\epsilon}{5}$$

for all x with $|x - x_0| < \delta_0$. Set

$$\mathcal{I}_{x_0} = \left\{\left(x - \frac{\delta}{5}, x + \frac{\delta}{5}\right) : 0 < \delta \leq \delta_0\right\}$$

and let $\mathcal{I} = \cup_{x_0 \in \mathcal{C}} \mathcal{I}_{x_0}$. Then $\mathcal{I}$ is a Vitali covering of $\mathcal{C}$.

For each $I = \left(x_0 - \frac{\delta}{5}, x_0 + \frac{\delta}{5}\right) \in \mathcal{I}$, let I' be the interval $(x_0 - \delta, x_0 + \delta)$ with the same midpoint as I of five times the length. By () there is a sequence (I_n) of disjoint sets in $\mathcal{I}$ such that the I'_n cover $\mathcal{C}$. If $x \in I'_n$, and x_n is the midpoint of I_n, then

$$|f(x) - f(x_n)| < \frac{\epsilon}{5}|x - x_n| < \frac{\epsilon}{5}\frac{\ell(I'_n)}{2} = \epsilon\frac{\ell(I_n)}{2}$$

and thus, if $x, y \in I'_n$ we have

$$|f(x) - f(y)| \leq |f(x) - f(x_n)| + |f(x_n) - f(y)| < \epsilon\ell(I_n)$$

and so $f(I'_n)$ is contained in an interval J_n of length $\epsilon\ell(I_n)$. The J_n cover $\mathcal{S}$.

Since the I_n are disjoint, we have $\sum_n \ell(I_n) \leq b - a$, so

$$\sum_{n=1}^{\infty} \ell(J_n) = \epsilon \sum_{n=1}^{\infty} \ell(I_n) \leq \epsilon(b - a)$$

and since $\epsilon > 0$ is arbitrary, we have that $\mathcal{S}$ has measure zero in $\mathbb{R}$. 👍

XIII.23.5. The Case $n = m > 1$

XIII.23.6. The Case $n > m$, r Large

We now show that the theorem holds for general n and m if r is sufficiently large. Most importantly, this result applies for all n and m if f is $\mathcal{C}^\infty$. The estimate on the minimum r from this argument is not sharp in general (it is quadratic in $n - m$, not linear). This proof is adapted from [Hir94].

XIII.23.6.1. For $n, m \in \mathbb{N}$, define $s(n, m) \in \mathbb{N}$ as follows:

$$s(n,m) = \begin{cases} 1 & \text{if} \quad n \leq m \\ \frac{(n-m+1)(n-m+2)}{2} & \text{if} \quad n > m \end{cases}.$$

The following properties needed for the proof of the theorem are easily verified:

XIII.23.6.2. Proposition. The numbers $s(n, m)$ have the following properties:

(i) $s(n, m) > \frac{n}{m}$ if $n > m$.

(ii) $s(n, m) - k \geq s(n - 1, m)$ if $1 \leq k \leq \frac{n}{m}$.

(iii) $s(n, m) = s(n - 1, m - 1)$ if $m > 1$.

XIII.23.6.3. Theorem. Let $n, m \in \mathbb{N}$. Then if $r \geq s(n, m)$, N is a separable $\mathcal{C}^r$ manifold, M a $\mathcal{C}^r$ m-manifold, and $f : N \to M$ is a $\mathcal{C}^r$ function, then the set of singular values of f is a subset of M of measure zero in M.

XIII.23.6.4. The proof will be by double induction: induction on m, and for fixed m induction on n. The cases $n \leq m$ have already been proved (), so fix n and m and assume that $n > m$ and that the theorem is proved for $(n-1, m)$, and for $(n-1, m-1)$ if $m > 1$. If Σ is the set of singular points of N, it suffices to show that every $\boldsymbol{x}_0 \in \Sigma$ has a neighborhood U such that $f(\Sigma \cap U)$ has measure 0 in M. Write

$$\Sigma = \Sigma_1 \cup \Sigma_2 \cup \Sigma_3$$

where

Σ_1 is the set of $\boldsymbol{x}_0 \in \Sigma$ for which all partial derivatives of f of orders $\leq \frac{n}{m}$ vanish at $\boldsymbol{x}_0$.

Σ_2 is the set of $\boldsymbol{x}_0 \in \Sigma$ such that the first-order partials of f vanish at $\boldsymbol{x}_0$ but some partial derivative of f of order k, $2 \leq k \leq \frac{n}{m}$, is nonzero at $\boldsymbol{x}_0$.

Σ_3 is the set of $\boldsymbol{x}_0 \in \Sigma$ such that some first-order partial of f is nonzero at $\boldsymbol{x}_0$.

XIII.23.6.5. First suppose $\boldsymbol{x} \in \Sigma_1$.

XIII.23.6.6. Now suppose $\boldsymbol{x}_0 \in \Sigma_2$, $\boldsymbol{y}_0 = f(\boldsymbol{x}_0)$. For each multiindex α of order k, $2 \leq k < \frac{n}{m}$, set

$$g_{\alpha,i} = \frac{\partial^\alpha f_i}{\partial x_\alpha}$$

where f_i is the i'th coordinate function of f. Then $g_{\alpha,i}$ is $\mathcal{C}^q$, where

$$q = r - k \geq s(n,m) - \frac{n}{m} \geq s(n-1,m) \ .$$

For each multiindex α of order k, $2 \leq k < \frac{n}{m}$, and each i and j, let $X_{\alpha,i,j}$ be the set of points $\boldsymbol{x} \in \Sigma_2$ such that $g_{\alpha,i}(\boldsymbol{x}) = 0$ and $\frac{\partial g_{\alpha,i}}{\partial x_j}(\boldsymbol{x}) \neq 0$. Then $\boldsymbol{x}_0 \in X_{\alpha,i,j}$ for some α, i, j.

Fix α, i, j with $\boldsymbol{x}_0 \in X_{\alpha,i,j}$. Then, since $g_{\alpha,i}(\boldsymbol{x}_0) = 0$ and $\frac{\partial g_{\alpha,i}}{\partial x_j}(\boldsymbol{x}_0) \neq 0$, $\boldsymbol{x}_0$ is a regular point of $g_{\alpha,i}$. Thus there is a neighborhood U of $\boldsymbol{x}_0$ such that 0 is a regular value of $h_{\alpha,i} = g_{\alpha,i}|_U$ and therefore $Z_{\alpha,i} = h_{\alpha,i}^{-1}(\{0\})$ is a $\mathcal{C}^q$-submanifold of N of dimension $n-1$ containing $\boldsymbol{x}_0$. Since $q \geq s(n-1,m)$ and $f|_{Z_{\alpha,i}}$ is $\mathcal{C}^q$, the inductive hypothesis applies. Any point of Σ_2 which is in $Z_{\alpha,i}$ is a singular point of $f|_{Z_{\alpha,i}}$ since its differential is zero, so we conclude that the measure of $f(\Sigma_2 \cap Z_{\alpha,i})$ is zero. Since $X_{\alpha,i,j} \cap U \subseteq Z_{\alpha,i}$, we have that the measure of $X_{\alpha,i,j} \cap U$ is zero. There is a neighborhood of $\boldsymbol{x}_0$ in Σ_2 contained in a finite union of such sets for various α, i, j, so we conclude that the measure of $f(\Sigma_2)$ is zero.

XIII.23.6.7. If $m = 1$, $\Sigma_3 = \emptyset$ so the proof is finished. If $m > 1$, suppose $\boldsymbol{x}_0 \in \Sigma_3$, $\boldsymbol{y}_0 = f(\boldsymbol{x}_0)$. By (), there are neighborhoods U of $\boldsymbol{x}_0$ and V of $\boldsymbol{y}_0$, and diffeomorphisms $g : \mathbb{R}^{n-1} \times \mathbb{R} \to U$ and $h : V \to \mathbb{R}^{m-1} \times \mathbb{R}$ with $g(\boldsymbol{0}, 0) = \boldsymbol{x}_0$, $h(\boldsymbol{y}_0) = (\boldsymbol{0}, 0)$, such that $\phi = h \circ f \circ g$ is of the form

$$\phi(\boldsymbol{z}, t) = (\psi(\boldsymbol{z}, t), t)$$

for some $\mathcal{C}^r$-function $\psi : \mathbb{R}^{n-1} \times \mathbb{R} \to \mathbb{R}^{m-1}$. The singular values of f in V are precisely the images under h^{-1} of the singular values of ϕ, so it suffices to show that the set $\mathcal{S}$ of singular values of ϕ has measure 0. Fix t, and let $\omega_t : \mathbb{R}^{n-1} \to \mathbb{R}^{m-1}$ be defined by $\omega_t(\boldsymbol{z}) = \psi(\boldsymbol{z}, t)$. Then ω is $\mathcal{C}^r$, so by the inductive hypothesis

the set S_t of singular values of ω_t has measure 0 in $\mathbb{R}^{m-1}$. We have that $\boldsymbol{z}$ is a singular point of ω_t if and only if (z,t) is a singular point of ϕ, so

$$\mathcal{S} = \{(\boldsymbol{z},t) : t \in \mathbb{R}, \boldsymbol{z} \in S_t\} \ .$$

Thus by the Baby Tonelli Theorem (XIII.23.1.11.) $\mathcal{S}$ has measure zero in $\mathbb{R}^{m-1} \times \mathbb{R}$.

This completes the proof of Theorem XIII.23.6.3.. 👍

XIII.23.7. The Case $n > m$, Sharp r Bound

We now turn to the final and most delicate part of the theorem, establishing the sharp bound on r. It is questionable whether it is worth proving this version, since the softer result XIII.23.6.3. suffices for almost all applications. But the conceptual reason behind the sharp bound on r is important and instructive.

The reason for the sharp bound $r \geq n - m + 1$ is a higher-dimensional version of the following one-dimensional fact (cf. ()):

XIII.23.7.1. PROPOSITION. Let $f : \mathbb{R} \to \mathbb{R}$ be a $\mathcal{C}^r$ function ($r \geq 0$). If f has a nonisolated zero at x_0, then $f(x_0) = f'(x_0) = \cdots = f^{(r)}(x_0) = 0$. It follows that f vanishes at x_0 at least to order r^+, i.e.

$$\lim_{x \to x_0} \frac{f(x)}{(x - x_0)^r} = 0 \ .$$

PROOF: If $r = 0$ there is nothing to prove. If $r \geq 1$, let (x_n) be a sequence of zeroes of f converging to x_0. By Rolle's Theorem, for each n there is a y_n between x_0 and x_n with $f'(y_n) = 0$. We have $y_n \to x_0$ by the Squeeze Theorem, so $f'(x_0) = 0$ since f' is continuous. If $r > 1$, repeat the argument using f' in place of f and the y_n in place of the x_n to conclude that $f''(x_0) = 0$, and continue inductively. The last statement follows from V.10.1.3.. 👍

XIII.23.7.2. This result is sharp. If $f(x) = \frac{|x|}{1+(\log|x|)^2} \sin(\log|x|)$ for $x \neq 0$ and $f(0) = 0$, then f is $\mathcal{C}^1$ and has a nonisolated zero at 0, but vanishes only to order 1^+ at 0. There are similar counterexamples for larger r.

XIII.23.8. The Category Version

The following result is a simple corollary of the measure-theoretic version of Sard's Theorem (cf. [Bro35]).

XIII.23.8.1. THEOREM. Let N be a separable $\mathcal{C}^r$ n-manifold, M a $\mathcal{C}^r$ m-manifold, and $f : N \to M$ a $\mathcal{C}^r$ function, where $r \geq \max(n - m + 1, 1)$. Then the set of singular values of f is a meager set in M.

PROOF: First note that the set of singular points of f is closed in N: it is the set of $\boldsymbol{x}_0 \in N$ such that $Det(Df(\boldsymbol{x}_0)Df(\boldsymbol{x}_0)^t) = 0$ (). Thus, although the set of singular values of f need not be closed in M, it is an F_σ (XIII.11.7.11.), and has measure zero by XIII.23.2.1.; thus it has empty interior (), i.e. its complement in M is a dense G_δ. 👍

XIII.23.8.2. Corollary. Let N be a separable n-manifold, M an m-manifold, and $f : N \to M$ a $\mathcal{C}^1$ function. If $n < m$, then the range of f is a meager subset of M.

XIII.23.8.3. These results apply to the case where N and M are Euclidean spaces. They imply that manifolds of different dimensions cannot be diffeomorphic; in particular, $\mathbb{R}^n$ and $\mathbb{R}^m$ are not diffeomorphic if $n \neq m$ (actually more is true: they cannot even be homeomorphic, cf. ()).

XIII.23.9. Counterexamples

The differentiability restrictions in Sard's theorem are sharp (as long as only $\mathcal{C}^r$ functions are considered for $r \in \mathbb{N}$; some relaxation is possible if more delicate degrees of differentiability are allowed, cf. [Bat93]). Even before any of the publications of A. P. Morse or Sard, H. Whitney [Whi35] described counterexamples to stronger statements.

XIII.23.9.1. Whitney described a (highly nondifferentiable) parametrized curve $\gamma : [0,1] \to \mathbb{R}^2$, with γ injective, such that

$$\lim_{t \to t_0} \frac{|t - t_0|}{\|\gamma(t) - \gamma(t_0)\|} = 0$$

for every $t_0 \in [0,1]$. Then the function $\gamma^{-1} : C = \gamma([0,1]) \to [0,1]$ has "zero derivative" at each point $\boldsymbol{a}$ of C in the sense that

$$\lim_{\boldsymbol{x} \to \boldsymbol{a}} \frac{|\gamma^{-1}(\boldsymbol{x}) - \gamma^{-1}(\boldsymbol{a})|}{\|\boldsymbol{x} - \boldsymbol{a}\|} = 0$$

where the limit is taken within C. By Whitney's Extension Theorem (), γ^{-1} extends to a $\mathcal{C}^1$ function f from $\mathbb{R}^2$ to $\mathbb{R}$; every point of C is a critical point of f, so every number in $[0,1]$ is a critical value.

As described in [Hir94], "This [example] leads to the following paradox: the graph of f is a surface $S \subset \mathbb{R}^3$ on which there is an arc A, at every point of which the surface has a horizontal tangent plane, yet A is not at a constant height. To make this more vivid, imagine that S is a hill and A a path on the hill. The hill is level at every point of the path, yet the path goes up and down." In fact, the path can be made to go monotonically up from bottom to top. We should point out that such a path would not be a very good route to follow to ascend the hill, since it has infinite arc length; in fact, its projection into the xy-plane has positive Lebesgue measure (the curve has Hausdorff dimension 2 and positive 2-dimensional Hausdorff measure). There cannot be a smooth path with this property because of the Chain Rule.

Whitney briefly described a higher-dimensional version, giving a $\mathcal{C}^{n-1}$ function f_n ($n \geq 2$) from $\mathbb{R}^n$ to $\mathbb{R}$ for which the set of critical values contains an interval.

If f_n is such a function, then for any m the function $g_{n,m} : \mathbb{R}^{n+m-1} \to \mathbb{R}^m$ given by

$$g_{n,m}(x_1, \dots, x_n, x_{n+1}, \dots, x_{n+m-1}) = (f_n(x_1, \dots, x_n), x_{n+1}, \dots, x_{n+m-1})$$

is a $\mathcal{C}^{n-1}$ function whose singular values contain an entire rectangle in $\mathbb{R}^m$.

XIII.23.9.2. Here is a simpler example from [Gri85] for the case $n = 2$, $m = 1$.

We need a $\mathcal{C}^1$ function f from $\mathbb{R}$ to $\mathbb{R}$ for which every point of the Cantor set K is a singular value. Here is one construction. Fix γ, $2 < \gamma < 3$, and let $\alpha = (\alpha_1, \alpha_2, \dots)$ with $\alpha_n = \gamma^{-n}$. Construct the Cantor set K_α as in XIII.18.2.1., i.e. subintervals of length γ^{-n} are removed at each step. Let (I_k) be a list of the

removed intervals. Let ϕ be a mesa function () on $[0,1]$ and $\beta = \int_0^1 \phi(t)\,dt$; then $0 < \beta < 1$. If the length of I_k is γ^{-n}, let h_k be a translated and scaled version of ϕ on I_k whose integral is 3^{-n}, i.e. scaled horizontally by γ^{-n} and vertically by $\frac{1}{\beta}\left(\frac{\gamma}{3}\right)^n$. Since

$$\lim_{n\to\infty}\left(\frac{\gamma}{3}\right)^n = 0$$

we have that $\sum_{k=1}^{\infty} h_k$ converges uniformly to a continuous function h. Then $h(t) \geq 0$ for all $t \in \mathbb{R}$ and $h(t) = 0$ for all $t \in K_\alpha$, and for $t < 0$, $t > 1$.

Set $g(x) = \int_{-\infty}^{x} h(t)\,dt$. Then g is a nondecreasing $\mathcal{C}^1$ function on $\mathbb{R}$, $g(x) = 0$ for $x \leq 0$ and $g(x) = 1$ for $x \geq 1$, and $g(x) \in K$ for $x \in K_\alpha$. For each such x, $g'(x) = h(x) = 0$. Thus every point of K is a critical value of g (the converse holds too).

Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x,y) = g(x) + g(y)$. Then f is $\mathcal{C}^1$, and if x and y are in K_α, then $Df(x,y) = 0$. If $z \in [0,2]$, then there are $c, d \in K$ such that $z = c + d$ (Exercise XIII.18.3.1.). There are $x, y \in K_\alpha$ with $g(x) = c$ and $g(y) = d$; so $f(x,y) = z$ and z is a critical value, hence singular value, of f. Thus the set of singular values of f does not have measure zero in $\mathbb{R}$. (Actually $[0,2]$ is the entire range of f, and consists of singular values.)

Note that h is not differentiable everywhere, hence g, and therefore f, are not $\mathcal{C}^2$; of course a $\mathcal{C}^2$ function with this property cannot exist by Sard's Theorem.

XIII.23.10. Applications

XIII.23.10.1. THEOREM. Let X be a compact subset of $\mathbb{R}^n$. Then for any $\epsilon > 0$ there is an open neighborhood U of X in $\mathbb{R}^n$ with $\bar{U}$ compact, such that ∂U is a smooth submanifold of $\mathbb{R}^n$ (i.e. $\bar{U}$ is a compact smooth submanifold with boundary), such that $\rho(\mathbf{y}, X) < \epsilon$ for all $\mathbf{y} \in U$, U contains all $\mathbf{y} \in \mathbb{R}^n$ with $\rho(\mathbf{y}, X) < \frac{\epsilon}{4}$, and, for every $\mathbf{y} \in \partial U$, $\frac{\epsilon}{4} < \rho(\mathbf{y}, X) < \epsilon$.

PROOF: Let $V = \{\mathbf{y} \in \mathbb{R}^n : \rho(\mathbf{y}, X) < \epsilon\}$. Then V is a neighborhood of X in $\mathbb{R}^n$ with compact closure. Define $g : \mathbb{R}^n \to \mathbb{R}$ by $g(\mathbf{y}) = \rho(\mathbf{y}, X)$; then g is continuous. Let f be a C^∞ function from $\mathbb{R}^n$ to $\mathbb{R}$ (e.g. a polynomial in the coordinates) uniformly approximating g within $\frac{\epsilon}{4}$ on $\bar{V}$; there is such an f by the Stone-Weierstrass Theorem (). There is a regular value α for f with $\frac{\epsilon}{2} < \alpha < \frac{3\epsilon}{4}$ by Sard's Theorem. Set

$$U = \{\mathbf{y} \in V : f(\mathbf{y}) < \alpha\}\ .$$

Then U is open. Since for $\mathbf{y} \in \bar{V}$ we have $f(\mathbf{y}) - \frac{\epsilon}{4} < \rho(\mathbf{y}, X) < f(\mathbf{y}) + \frac{\epsilon}{4}$, we have that $\bar{U} \subseteq V$, so

$$\partial U = \{\mathbf{y} \in V : f(y) = \alpha\}$$

is a smooth closed submanifold of $\mathbb{R}^n$, and the desired inequalities follow. 👍

XIII.23.10.2. In fact the $\frac{\epsilon}{4}$ can be replaced by any $\eta < \epsilon$ by a suitable modification of the proof.

XIII.23.10.3. In the case $n = 2$, if X is a compact subset of the plane there is for any $\epsilon > 0$ an ϵ-neighborhood U of X in $\mathbb{R}^2$ whose boundary consists of finitely many disjoint smooth closed curves. (Such a neighborhood is not necessarily homeomorphic to a disjoint union of closed disks, however, e.g. if X is a circle.) See XIII.23.12.3. for a constructive proof not using Sard's Theorem.

XIII.23.11. Refined Versions

Even sharper versions of Sard's Theorem are known, such as:

XIII.23.11.1. THEOREM. [Fed69, 3.4.3] Let $n > k \geq 0$ and $r \geq 1$ be integers, and let f be a $\mathcal{C}^r$ function from an open subset U of $\mathbb{R}^n$ into a normed vector space Y. Set

$$B_k = f(\{\boldsymbol{x} \in U : rank\ Df(\boldsymbol{x}) \leq k\}) \subseteq Y\ .$$

Then

$$\mathscr{H}^{k+\frac{n-k}{r}}(B_k) = 0$$

(where $\mathscr{H}^s$ denotes s-dimensional Hausdorff measure on Y (XVII.6.3.3.).)

Note that the usual form (XIII.23.2.1.) of Sard's Theorem follows from the case $Y = \mathbb{R}^m$, $k = m-1$, $r \geq \max(n-m+1, 1)$ (the set of singular values of f is B_{m-1}, or actually B_k, $k = \min(n, m-1)$).

The result can be rephrased in terms of Hausdorff dimension (XVII.6.3.16.), although the statement is slightly weaker:

XIII.23.11.2. COROLLARY. Let $n > k \geq 0$ and $r \geq 1$ be integers, and let f be a $\mathcal{C}^r$ function from an open subset U of $\mathbb{R}^n$ into a normed vector space Y. Set

$$B_k = f(\{\boldsymbol{x} \in U : rank\ Df(\boldsymbol{x}) \leq k\}) \subseteq Y\ .$$

Then

$$dim_H(B_k) \leq k + \frac{n-k}{r}\ .$$

XIII.23.11.3. COROLLARY. Let f be a $\mathcal{C}^\infty$ function from an open set in $\mathbb{R}^n$ to $\mathbb{R}^m$. If $S \subseteq \mathbb{R}^m$ is the set of singular values of f, then

$$dim_H(S) \leq \min(n, m-1)\ .$$

The case $k = 0$ (the set of critical values) is also interesting:

XIII.23.11.4. COROLLARY. Let f be a $\mathcal{C}^r$ function from an open set in $\mathbb{R}^n$ to a normed vector space Y. If $C \subseteq Y$ is the set of critical values of f, then

$$dim_H(C) \leq \frac{n}{r}\ .$$

If f is $\mathcal{C}^\infty$, then $dim_H(C) = 0$.

XIII.23.11.5. These results extend immediately to the case where X is a separable smooth n-manifold and f a $\mathcal{C}^r$ function from X to a normed vector space Y. Extending to the case where Y is a separable smooth manifold is more difficult because of subtleties in defining Hausdorff measure and Hausdorff dimension for subsets of a manifold (cf. ()).

XIII.23.11.6. There is another fairly recent generalization of XIII.23.2.1.. A function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is $\mathcal{C}^{r,1}$ if it is r-times differentiable on U and $D^r f$ is Lipschitz (), and *locally* $\mathcal{C}^{r,1}$ if it is $\mathcal{C}^{r,1}$ in a neighborhood of each point of U. A locally $\mathcal{C}^{r,1}$ function is $(r+1)$-times differentiable a.e. on U, and a $\mathcal{C}^{r+1}$ function is locally $\mathcal{C}^{r,1}$.

XIII.23.11.7. THEOREM. [Bat93] Let N be a separable $\mathcal{C}^r$ n-manifold, and let $f : N \to \mathbb{R}^m$ be a locally $\mathcal{C}^{r-1,1}$ function, with $r \geq \max(n-m+1, 1)$. Then the set of singular values of f is a subset of $\mathbb{R}^m$ of measure zero in $\mathbb{R}^m$.

There is again a slight problem in extending to the case where the target is a manifold, since the definition of locally $\mathcal{C}^{r-1,1}$ must be extended to this case.

XIII.23.12. Exercises

XIII.23.12.1. Carefully justify all the assertions in XIII.23.1.3., i.e. show that adding each of the various restrictions or modifications to the definition does not change the class of null sets.

XIII.23.12.2. Show that if C is any Cantor set in $\mathbb{R}$ of Lebesgue measure 0, there is a $\mathcal{C}^1$ function $f : \mathbb{R} \to \mathbb{R}$ whose set of critical values is C. [Write $C = C_\lambda$ as in XIII.18.2.9., and mimic the argument of XIII.23.9.2., using IV.2.16.2..]

XIII.23.12.3. [?] Give a constructive proof of the case $n = 2$ of XIII.23.10.1., as follows.

(a) Show that the diameter of a regular hexagon of side s is $2s$. [Write the hexagon as a union of six equilateral triangles.]

(b) If X is a compact subset of $\mathbb{R}^2$ and V is an open neighborhood of X, there is a bounded open neighborhood W of X contained in V. Let ϵ be the distance between X and ∂W. Show that $\epsilon > 0$.

(c) Cover the plane with a hexagonal honeycomb with sides of length $s < \frac{\epsilon}{4}$, and let A be the union of all the closed hexagons which are contained in W. Show that X is contained in the interior U of A.

(d) Show that $\partial A = \partial U$ consists of finitely many disjoint simple closed polygonal curves.

(e) If desired, the corners of these boundary curves can be rounded off to give smooth boundary curves.

(f) Try to do the same argument using squares (try staggering the squares so no four edges meet at a point), and try generalizing to $\mathbb{R}^n$.

XIII.23.12.4. [Edg98] Here is another example of how $\mathcal{C}^1$ differs from $\mathcal{C}^2$ and can be pathological. Let f be a strictly increasing singular function from $\mathbb{R}$ to $\mathbb{R}$, i.e. a continuous function with $f' = 0$ a.e. (XIX.2.3.5.(i)), and let g be an antiderivative for f. Then g is $\mathcal{C}^1$. Let γ be the graph of g, a parametrized curve in $\mathbb{R}^2$.

(a) Show that g is a strictly convex function (), and hence γ is a strictly convex $\mathcal{C}^1$ curve.

(b) Show that the curvature of γ () is 0 a.e.

(c) If Pompeiu's example (XIX.2.3.5.(iii)) is used for f, then γ has everywhere defined curvature, which is 0 on a dense set. (But note that γ is not $\mathcal{C}^2$ since γ'' is not continuous.)

XIII.24. Miscellaneous Aspects of Topology

XIII.24.1. The Countable Chain Condition

Various forms of the AC are needed at points of this discussion, so for simplicity we will assume the AC.

XIII.24.1.1. Definition. Let X be a topological space. X has the *Countable Chain Condition (CCC)* if every collection of pairwise disjoint nonempty open sets in X is countable (finite or countably infinite).

XIII.24.1.2. The CCC is a nice condition for a space to have; but the Countable Chain Condition turns out to be logically subtle, and it has been studied by set theorists. It is related to Martin's Axiom (), a weakening of the Continuum Hypothesis. There are spaces, even compact Hausdorff spaces, which do not have the CCC: the simplest example is the one-point compactification of an uncountable discrete space.

XIII.24.1.3. As an example of the logical subtlety, it cannot be proved or disproved in ZFC that a product of two spaces with the CCC has the CCC. This is connected with the question of the existence of a Suslin line ().

There are some things which can be said, such as:

XIII.24.1.4. Proposition. A separable topological space has the CCC.

This is obvious, since every nonempty open set contains points of a dense subset.

XIII.24.1.5. Proposition. Let X and Y be topological spaces. If X has the CCC and there is a surjective map from X to Y, then Y also has the CCC.

Proof: Let $f : X \to Y$ be a surjective map. If $\{V_i : i \in I\}$ is a collection of pairwise disjoint nonempty open sets in Y, then $\{f^{-1}(V_i) : i \in I\}$ is a collection of pairwise disjoint nonempty open sets in X, so I is countable. 👍

XIII.24.1.6. Proposition. Let X be a space with the CCC, and U an open subset. Then U has the CCC in the relative topology.

Proof: Any relatively open set in U is open in X. 👍

What about products of separable spaces? "Small" products are separable, hence have the CCC, but large products are not separable in general. However, an arbitrary product of separable spaces turns out to have the CCC (XIII.24.1.13.). This is rather tricky to prove; we give a proof using an intermediate property from measure theory.

Measurable Separability

XIII.24.1.7. DEFINITION. Let X be a topological space. X is *measurably separable (MS)* if there is a probability measure μ defined on a σ-algebra $\mathcal{A}$ of subsets of X, such that every nonempty open set in X contains a set in $\mathcal{A}$ of strictly positive measure. Such a measure is called a *separating measure* for X.

Note that the open sets in X are not assumed be μ-measurable (i.e. in $\mathcal{A}$).

XIII.24.1.8. PROPOSITION. Every separable topological space is MS.

PROOF: Let X be a separable topological space and (x_n) a dense sequence. Then the purely atomic measure μ supported on the x_n with $\mu(\{x_n\}) = 2^{-n}$ is a separating measure. 👍

XIII.24.1.9. PROPOSITION. Every MS space has the CCC.

PROOF: Let X be an MS space, μ a separating measure on $(X, \mathcal{A})$, and $\{U_i : i \in I\}$ a collection of pairwise disjoint nonempty open sets in X. For each i there is an $A_i \in \mathcal{A}$ with $A_i \subseteq U_i$ and $\mu(A_i) > 0$. A finite measure space cannot have an uncountable family of pairwise disjoint measurable sets of strictly positive measure, so I is countable. 👍

XIII.24.1.10. PROPOSITION. Let X and Y be topological spaces. If X is MS and there is a surjective map from X to Y, then Y is also MS.

PROOF: Let $f : X \to Y$ be a surjective map. If μ is a separating measure on A, defined on $(X, \mathcal{A})$, let $\mathcal{B} = f^{\to}(\mathcal{A})$ (XVI.1.6.6.), and $f_*(\mu)$ the push-forward (XVII.1.6.14.) of μ to $(Y, \mathcal{B})$. Then $f_*(\mu)$ is a separating measure for Y. 👍

XIII.24.1.11. PROPOSITION. Let X be an MS space, and U an open subset. Then U is MS in the relative topology.

PROOF: Let $(X, \mathcal{A}, \mu)$ be a separating measure. Form the restriction $\mu|_U$ of μ to U. Since U might not be μ-measurable, the restriction must be defined as in XVII.1.6.7.. Then $\mu|_U$ is a nonzero finite measure on $(U, \mathcal{A}_U)$. The renormalization to a probability measure is a separating measure for U. 👍

The concept of measurable separability was designed to prove the following theorem:

XIII.24.1.12. THEOREM. An arbitrary product of MS spaces is MS (in the product topology).

PROOF: Let $\{X_i : i \in I\}$ be a set of MS spaces, and for each i let μ_i be a separating measure for X_i, defined on a σ-algebra $\mathcal{A}_i$. Let $X = \prod_i X_i$ and $\mu = \times_i \mu_i$ on the σ-algebra $\mathcal{A} = \otimes_i \mathcal{A}_i$ (). We claim that μ is a separating measure for X.

Let U be a nonempty open set in X. Then there is a finite set $F \subseteq I$ and nonempty open sets V_i in X_i $(i \in F)$ such that

$$\left[\prod_{i\in F} V_i\right] \times \left[\prod_{i\in I\setminus F} X_i\right] \subseteq U \ .$$

For each $i \in F$ there is an $A_i \in \mathcal{A}_i$ with $A_i \subseteq V_i$ and $\mu_i(A_i) > 0$. Set

$$A = \left[\prod_{i\in F} A_i\right] \times \left[\prod_{i\in I\setminus F} X_i\right] \in \mathcal{A} \ .$$

Then $A \subseteq U$ and $\mu(A) = \prod_{i\in F} \mu_i(A_i) > 0$. 👍

Note that if I is large, open sets in X are not in $\mathcal{A}$ in general.

Putting together XIII.24.1.8., XIII.24.1.12., and XIII.24.1.9., we obtain:

XIII.24.1.13. COROLLARY. An arbitrary product of separable topological spaces has the CCC.

XIII.25. Piecewise-Linear Topology

The idea behind piecewise-linear (PL) topology is to reduce the study of the topology of "reasonable" subsets of Euclidean space, and eventually all "reasonable" (e.g. compact metrizable) topological spaces, to the affine structure of $\mathbb{R}^n$. This is clearly possible in principle since the affine structure on $\mathbb{R}^n$ determines its topological structure, and it turns out in practice to be a slick and clever way to express many topological notions in essentially combinatorial language.

PL topology has somewhat passed out of fashion. However, the basic setup (which is all we will do here, what topologists would consider to be "preliminaries") consisting of polyhedra and simplicial complexes and their maps, is material every mathematician should be familiar with.

The standard reference for PL topology is [RS72], on which much of our exposition is based.

XIII.25.1. Polyhedra, Cells, and Simplexes

The term "polyhedron" traditionally refers to a three-dimensional figure made up of a finite number of vertices, edges, and faces; it is the three-dimensional analog of a polygon. We will use the term in a somewhat different sense, applying to objects of arbitrary dimension; thus (most) polyhedra in the traditional sense will be polyhedra in our sense, as will be polygons in the plane, closed intervals, points, and analogous higher-dimensional figures.

There is an ambiguity in the usual use of terms like "polyhedron," "polygon," "triangle," "square," etc.: do we include the "interior" or just the "surface"? Our definition of polyhedron will include both cases, although when a specific polyhedron is considered we must carefully specify which we mean. For example, a solid triangle (homeomorphic to a disk) and a triangle not including the interior (homeomorphic to a circle) will both be polyhedra; the first is a 2-dimensional polyhedron and the second a 1-dimensional polyhedron.

Cells

XIII.25.1.1. Definition. A *cell* in $\mathbb{R}^n$ is a compact convex set with only finitely many extreme points. An extreme point of a cell is called a *vertex* of the cell. (The plural of *vertex* is *vertices*.) The *dimension* of a cell is the dimension of the affine subspace it generates.

XIII.25.1.2. Since a compact convex subset of $\mathbb{R}^n$ is the convex hull of its extreme points (), a cell is precisely the convex hull of a finite set S of points in $\mathbb{R}^n$. (Note, however, that not all the points of S are necessarily vertices of the cell defined by S, although each vertex of the cell is in S.)

XIII.25.1.3. Example. A rectangle (product of intervals) is a cell. A rectangle whose sides all have the same length is called a *cube*.

It is convenient in PL-topology to use the ρ_∞ metric on $\mathbb{R}^n$. The closed balls in this metric are cubes, hence cells. A closed ball in the usual ρ_2 metric, or the ρ_p metric for $1 < p < \infty$, is not a cell except in $\mathbb{R}^1$, since it has infinitely many extreme points.

XIII.25.1.4. If a cell C in $\mathbb{R}^m$ has $n+1$ vertices $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, then the affine subspace it generates is parallel to the subspace spanned by $\{\boldsymbol{v}_1 - \boldsymbol{v}_0, \ldots, \boldsymbol{v}_n - \boldsymbol{v}_0\}$, hence the dimension of C is the dimension of this subspace, which is $\leq n$. It can be strictly less than n (e.g. for a convex polygon in $\mathbb{R}^2$).

XIII.25.1.5. Recall () that a *face* of a convex set C is a subset F such that, whenever an interior point of a line segment in C is in F, the endpoints of the segment are in F. We have easily from () (or with an elementary proof):

XIII.25.1.6. PROPOSITION. Let C be a cell in $\mathbb{R}^n$, and F a face of C. Then F is also a cell, and the vertices of F are a subset of the vertices of C. The dimension of F is strictly smaller than the dimension of C if F is a proper face of C.

XIII.25.1.7. Not every set of vertices of a cell C determines a face: for example, if C is a (solid) square in $\mathbb{R}^2$, the faces are just the vertices, the edges, and the whole square. Two opposite vertices do not determine a face.

XIII.25.1.8. PROPOSITION. Let C be a cell in $\mathbb{R}^n$ and $\boldsymbol{y} \in C$. Let F_y be the smallest face of C containing $\boldsymbol{y}$. Then F_y is either $\{\boldsymbol{y}\}$ or the union F of all line segments in C containing $\boldsymbol{y}$ in the interior.

PROOF: If F is empty, then $\boldsymbol{y}$ is an extreme point of C and $F_y = \{\boldsymbol{y}\}$. Suppose $F \neq \emptyset$. By definition, any face containing y contains F, so we need only show that F is a face of C. To show F is convex, suppose $\boldsymbol{a}$ and $\boldsymbol{b}$ are in F, and that $\boldsymbol{c}$ is a convex combination of $\boldsymbol{a}$ and $\boldsymbol{b}$. There are line segments $\boldsymbol{ad}$ and $\boldsymbol{be}$ in C containing $\boldsymbol{y}$ in their interior. Let Q be the quadrilateral in the plane generated by $\boldsymbol{bd}$ and $\boldsymbol{be}$ with vertices $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{d}$, and $\boldsymbol{e}$; then $Q \subseteq C$ by convexity. Let L be the line through $\boldsymbol{c}$ and $\boldsymbol{y}$. Then $L \cap C$ contains $L \cap Q$, which is a line segment with $\boldsymbol{y}$ in its interior. Thus $\boldsymbol{c} \in F$ and F is convex.

Now suppose $\boldsymbol{a}, \boldsymbol{b} \in C$, and $\boldsymbol{c}$ is a (nontrivial) linear combination of $\boldsymbol{a}$ and $\boldsymbol{b}$ which is in F. Then there is a line segment $\boldsymbol{cd}$ in C containing $\boldsymbol{y}$ in its interior. Let T be the triangle (in the plane generated by $\boldsymbol{ab}$ and $\boldsymbol{cd}$) with vertices $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{d}$; then $T \subseteq C$ by convexity. If L is the line through $\boldsymbol{a}$ and $\boldsymbol{y}$, then $L \cap C$ contains $L \cap T$, which is a line segment containing $\boldsymbol{y}$ in its interior. Thus $\boldsymbol{a} \in F$. Similarly $\boldsymbol{b} \in F$. So F is a face of C. 👍

XIII.25.1.9. If C is a cell in $\mathbb{R}^n$, let $\mathring{C}$ be the set of all points of C not contained in any proper face of C, and $\dot{C}$ the union of the proper faces of C. Then $\mathring{C}$ and $\dot{C}$ are called the *interior* and *boundary* of C respectively (they are the topological interior and topological boundary of C in the affine subspace it generates with the relative topology). The interior $\mathring{C}$ is nonempty: if $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are the vertices of C, then $\mathring{C}$ contains all points of the form $\sum_{k=1}^m \lambda_k \boldsymbol{v}_k$ where all $\lambda_k > 0$ (and $\sum_{k=1}^m \lambda_k = 1$). In fact, $\mathring{C}$ is precisely this set. $\mathring{C}$ is a relatively open subset of C, and $\dot{C} = C \setminus \mathring{C}$ is compact.

XIII.25.1.10. PROPOSITION. Let C be a cell in $\mathbb{R}^n$, and $f : \mathbb{R}^n \to \mathbb{R}^m$ an affine function. Then $f(C)$ is a cell in $\mathbb{R}^m$.

PROOF: If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ are the vertices of C, then $f(C)$ is exactly the convex hull of $\{f(\boldsymbol{v}_1), \ldots, f(\boldsymbol{v}_m)\}$, hence a cell. (But $f(\boldsymbol{v}_k)$ is not necessarily a vertex of $f(C)$ for every k.) 👍

XIII.25.1.11. PROPOSITION. Let C be a convex subset of $\mathbb{R}^n$. Then C is a cell if and only if C is a finite intersection of (closed) half-spaces.

One direction of this is proved in XIII.25.1.31., and the converse in Exercise XIII.25.5.1..

Simplexes

XIII.25.1.12. DEFINITION. An *n-simplex* is an n-dimensional cell with exactly $n+1$ extreme points. A *simplex* is an n-simplex for some n.

The plural of *simplex* should properly be *simplices*, but (unlike with *vertex*) the anglicized plural *simplexes* has come into general use.

XIII.25.1.13. PROPOSITION. Let C be a cell in $\mathbb{R}^m$ with vertices $\boldsymbol{v}_0, \dots, v_n$. Then the following are equivalent:

(i) C is an n-simplex.

(ii) The set $\{\boldsymbol{v}_1 - \boldsymbol{v}_0, \dots, \boldsymbol{v}_n - \boldsymbol{v}_0\}$ is linearly independent in $\mathbb{R}^m$.

(iii) Every $\boldsymbol{a} \in C$ can be uniquely written as a convex combination

$$\boldsymbol{a} = \sum_{k=0}^{n} \lambda_k \boldsymbol{v}_k$$

with all $\lambda_k \geq 0$ and $\sum_{k=0}^{n} \lambda_k = 1$.

PROOF: (i) $\iff$ (ii): This follows immediately from XIII.25.1.4., since C is n-dimensional if and only if $\{\boldsymbol{v}_1 - \boldsymbol{v}_0, \dots, \boldsymbol{v}_n - \boldsymbol{v}_0\}$ spans an n-dimensional subset of $\mathbb{R}^m$.
(ii) $\Rightarrow$ (iii): Since C is a cell, every element of C can be written as a convex combination of the $\boldsymbol{v}_k$. So if (iii) is false, there are $(\lambda_0, \dots, \lambda_n) \neq (\mu_0, \dots, \mu_n)$ satisfying the conditions of (iii) with

$$\sum_{k=0}^{n} \lambda_k \boldsymbol{v}_k = \sum_{k=0}^{n} \mu_k \boldsymbol{v}_k$$

$$\boldsymbol{0} = \sum_{k=0}^{n} (\lambda_k - \mu_k)\boldsymbol{v}_k = \sum_{k=1}^{n} (\lambda_k - \mu_k)(\boldsymbol{v}_k - \boldsymbol{v}_0) + \sum_{k=0}^{n} (\lambda_k - \mu_k)\boldsymbol{v}_0 = \sum_{k=1}^{n} (\lambda_k - \mu_k)(\boldsymbol{v}_k - \boldsymbol{v}_0)$$

since $\sum_{k=0}^{n} (\lambda_k - \mu_k) = 0$ because $\sum_{k=0}^{n} \lambda_k = \sum_{k=0}^{n} \mu_k = 1$. So $\{\boldsymbol{v}_1 - \boldsymbol{v}_0, \dots, \boldsymbol{v}_n - \boldsymbol{v}_0\}$ is not linearly independent.
(iii) $\Rightarrow$ (ii): 👍

XIII.25.1.14. If C is a simplex, the numbers λ_k for which $\boldsymbol{a} = \sum_{k=0}^{n} \lambda_k \boldsymbol{v}_k$ are called the *barycentric coordinates* of $\boldsymbol{a}$ in C. These are indexed by the vertices of C, i.e. λ_k is the barycentric coordinate of $\boldsymbol{a}$ with respect to $\boldsymbol{v}_k$ (in C). If C is not a simplex, then barycentric coordinates are not well defined since some points can be written in more than one way as convex combinations of the vertices.

XIII.25.1.15. In $\mathbb{R}^{n+1}$, let Σ_n be the convex hull of the set of standard basis vectors $\{\mathbf{e}_0, \ldots, \mathbf{e}_n\}$ (it is convenient here to index the coordinates from 0 to n rather than from 1 to $n+1$). Then

$$\Sigma_n = \left\{ (\lambda_0, \ldots, \lambda_n) \in \mathbb{R}^{n+1} : \lambda_k \geq 0, \sum_{k=0}^{n} \lambda_k = 1 \right\}$$

is called the *standard n-simplex*. If S is any n-simplex in $\mathbb{R}^m$ with vertices $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_n$, then there is a unique affine map $\phi : \mathbb{R}^{n+1} \to \mathbb{R}^m$ with $\phi(\mathbf{e}_k) = \boldsymbol{v}_k$ for each k, which maps Σ_n homeomorphically onto S. Thus every n-simplex "looks the same" and is affinely homeomorphic to the standard n-simplex.

We sometimes regard the simplex in $\mathbb{R}^n$ with vertices $0, \mathbf{e}_1, \ldots, \mathbf{e}_n$, i.e.

$$\left\{ (\lambda_1, \ldots, \lambda_n) : \lambda_k \geq 0, \sum_{k=1}^{n} \lambda_k \leq 1 \right\}$$

as the standard n-simplex. There is an obvious affine homeomorphism between this simplex and Σ_n.

A 0-simplex is a single point. A 1-simplex is a (closed) line segment. A 2-simplex is a (solid) triangle. A 3-simplex is a (solid) tetrahedron.

XIII.25.1.16. If S is an n-simplex with vertices $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_n$, and $\boldsymbol{w}_0, \ldots, \boldsymbol{w}_n$ are any $n+1$ points in $\mathbb{R}^m$, then there is a unique affine function f from S onto the convex hull of $\{\boldsymbol{w}_0, \ldots, \boldsymbol{w}_n\}$ with $f(\boldsymbol{v}_k) = \boldsymbol{w}_k$ for all k. The obvious formula is

$$f\left(\sum_{k=0}^{n} \lambda_k \boldsymbol{v}_k \right) = \sum_{k=0}^{n} \lambda_k \boldsymbol{w}_k$$

for $\lambda_k \geq 0$, $\sum_{k=0}^{n} \lambda_k = 1$.

XIII.25.1.17. Proposition. Every simplex is a finite intersection of (closed) half-spaces.

Polyhedra

XIII.25.1.18. Definition. A *polyhedron* is a subset of some $\mathbb{R}^n$ which is a finite union of cells.

This is not the way the definition of a polyhedron is usually expressed in PL topology, but is equivalent (in the compact case). The PL definition (XIII.25.1.34.) is hard to motivate geometrically, so we have chosen this alternative definition.

XIII.25.1.19. Definition. Let B be a (usually compact) subset of $\mathbb{R}^n$, and $\boldsymbol{a}$ a point of $\mathbb{R}^n$ not in B. The *join* $\boldsymbol{a}B$ of $\boldsymbol{a}$ and B is the union of all line segments in $\mathbb{R}^n$ between $\boldsymbol{a}$ and a point of B. The join is a *cone* with vertex $\boldsymbol{a}$ and base B if every $\boldsymbol{x} \in \boldsymbol{a}B \setminus \{\boldsymbol{a}\}$ can be uniquely written as $\lambda\boldsymbol{a}(1-\lambda)\boldsymbol{b}$ for some $\boldsymbol{b} \in B$ and $0 \leq \lambda < 1$.

XIII.25.1.20. It is easily seen that $\boldsymbol{a}B$ is a cone if and only if every ray beginning at $\boldsymbol{a}$ intersects B in at most one point. A cone $\boldsymbol{a}B$ is compact if B is compact.

XIII.25.1.21. DEFINITION. Let X be a subset of $\mathbb{R}^n$, $\boldsymbol{a} \in X$, and $\epsilon > 0$. Set

$$N_\epsilon(\boldsymbol{a}, X) = \{\boldsymbol{x} \in X : \rho_\infty(\boldsymbol{a}, \boldsymbol{x}) \leq \epsilon\}$$

$$\dot{N}_\epsilon(\boldsymbol{a}, X) = \{\boldsymbol{x} \in X : \rho_\infty(\boldsymbol{a}, \boldsymbol{x}) = \epsilon\}$$

$$\mathring{N}_\epsilon(\boldsymbol{a}, X) = \{\boldsymbol{x} \in X : \rho_\infty(\boldsymbol{a}, \boldsymbol{x}) < \epsilon\} \ .$$

Then $\mathring{N}_\epsilon(\boldsymbol{a}, X)$ is an open neighborhood of $\boldsymbol{a}$ in X, and $N_\epsilon(\boldsymbol{a}, X)$ is a closed neighborhood of $\boldsymbol{a}$ in X. If $N_\epsilon(\boldsymbol{a}, X)$ is a cone with vertex $\boldsymbol{a}$ and base $\dot{N}_\epsilon(\boldsymbol{a}, X)$, $N_\epsilon(\boldsymbol{a}, X)$ is called the *ϵ-star* of $\boldsymbol{a}$ in X; if not, the ϵ-star of $\boldsymbol{a}$ in X is not defined. X is *locally conical* if for every $\boldsymbol{a} \in X$ there is an ϵ-star of $\boldsymbol{a}$ in X for some $\epsilon > 0$.

The set $N_\epsilon(\boldsymbol{a}, X)$ is the intersection of X with a cube of side 2ϵ centered at $\boldsymbol{a}$, and $\dot{N}_\epsilon(\boldsymbol{a}, X)$ is the intersection of X with the boundary of this cube.

If $\boldsymbol{a}$ has an ϵ-star in X, it also has an ϵ'-star for any $\epsilon' < \epsilon$.

Roughly speaking, a set X in $\mathbb{R}^n$ is locally conical if every point $\boldsymbol{a}$ of X has a neighborhood U in $\mathbb{R}^n$ such that $X \cap U$ is a union of ray segments with endpoint $\boldsymbol{a}$ with lengths bounded away from zero.

XIII.25.1.22. EXAMPLES. (i) Any open set in $\mathbb{R}^n$ is locally conical: if U is open and $\boldsymbol{a} \in U$, then for sufficiently small $\epsilon > 0$ $N_\epsilon(\boldsymbol{a}, U)$ is an entire closed cube centered at $\boldsymbol{a}$.

(ii) Any affine subspace or half-space in $\mathbb{R}^n$ is locally conical: in fact, any point in such a subset has an ϵ-star for every $\epsilon > 0$.

(iii) A polygon in $\mathbb{R}^2$, with or without interior, is locally conical.

(iv) A circle is not locally conical: in fact, no point has an ϵ-star for any $\epsilon > 0$. Similarly, a closed disk is not locally conical.

The important fact justifying our nonstandard definition of polyhedron is:

XIII.25.1.23. THEOREM. Let X be a compact subset of $\mathbb{R}^m$. The following are equivalent:

(i) X is a polyhedron.

(ii) X is a finite union of simplexes.

(iii) X is locally conical.

The theorem will be proved using some lemmas.

XIII.25.1.24. LEMMA. Let $X_1, \ldots, X_n$ be subsets of $\mathbb{R}^m$, $X = \cap_{k=1}^n X_k$, and $\boldsymbol{a} \in X$. If $N_{\epsilon_k}(\boldsymbol{a}, X_k)$ is an ϵ_k-star of $\boldsymbol{a}$ in X_k for each k, and $\epsilon = \min(\epsilon_1, \ldots, \epsilon_n)$, then $N_\epsilon(\boldsymbol{a}, X) = \cap_{k=1}^n N_\epsilon(\boldsymbol{a}, X_k)$ is an ϵ-star of $\boldsymbol{a}$ in X.

PROOF: Obvious from the fact that $N_\epsilon(\boldsymbol{a}, X_k)$ is an ϵ-star of $\boldsymbol{a}$ in X_k for all k. 👍

XIII.25.1.25. COROLLARY. A finite intersection of locally conical subsets of $\mathbb{R}^n$ is locally conical.

XIII.25.1.26. LEMMA. Let $X_1, \ldots, X_n$ be compact locally conical subsets of $\mathbb{R}^m$. Then $X = \cup_{k=1}^n X_k$ is locally conical.

PROOF: Let $\boldsymbol{a} \in X$. By renumbering, suppose $\boldsymbol{a} \in \cap_{k=1}^r X_k$ and $x \notin \cup_{k=r+1}^n X_k$. Let $N_{\epsilon_k}(\boldsymbol{a}, X_k)$ be an ϵ_k-star of $\boldsymbol{a}$ in X_k for $1 \leq k \leq r$. Choose $\epsilon > 0$ so that $\epsilon \leq \epsilon_k$ for $1 \leq k \leq r$ and $\epsilon < \rho_\infty\left(\boldsymbol{a}, \cup_{k=r+1}^n X_k\right)$. Then it is easily checked that $N_\epsilon(\boldsymbol{a}, X)$ is an ϵ-star of $\boldsymbol{a}$ in X. 👍

XIII.25.1.27. LEMMA. Every simplex is locally conical.

PROOF: A half-space is locally conical (XIII.25.1.22.(ii)); thus any simplex is locally conical by XIII.25.1.17. and XIII.25.1.25.. 👍

XIII.25.1.28. LEMMA. Let C be a cell in $\mathbb{R}^n$, and $\boldsymbol{a} \in C$. Then C is the union of cones of the form $\boldsymbol{a}F$, where F is a face of C not containing $\boldsymbol{a}$.

PROOF: Let $\boldsymbol{x} \in C$, $\boldsymbol{x} \neq \boldsymbol{a}$. If R is the ray starting at $\boldsymbol{a}$ through $\boldsymbol{x}$, then $R \cap C$ is a segment from $\boldsymbol{a}$ to a point $\boldsymbol{y}$. The face F_y (XIII.25.1.8.) cannot contain $\boldsymbol{a}$ since there is no line segment in C through $\boldsymbol{a}$ and $\boldsymbol{y}$ containing $\boldsymbol{y}$ in its interior (and $\boldsymbol{y} \neq \boldsymbol{a}$). Thus $\boldsymbol{x}$ is in the union of the sets $\boldsymbol{a}F$. Such a set is a cone, since if a ray from $\boldsymbol{a}$ intersects a face F in points $\boldsymbol{z}$ and $\boldsymbol{z}'$ with $\boldsymbol{z}$ between $\boldsymbol{a}$ and $\boldsymbol{z}'$, by definition of a face F must contain $\boldsymbol{a}$. 👍

PROOF OF XIII.25.1.23.: We have (ii) ⇒ (iii) by XIII.25.1.27. and XIII.25.1.26., and (ii) ⇒ (i) is trivial.

(i) ⇒ (ii): By XIII.25.1.26., it suffices to show that every cell is a finite union of simplexes. We prove this by induction on the dimension of the cell. If C is a cell of dimension ≤ 1, then C is a point or line segment, hence a simplex. Now suppose all cells of dimension $< n$ are unions of simplexes, and let C be a cell of dimension n. If $\boldsymbol{a} \in C$, then by XIII.25.1.28. C is a (finite) union of cones of the form $\boldsymbol{a}F$, where F is a face of C not containing $\boldsymbol{a}$. Each such F is a cell of dimension $< n$, hence a finite union of simplexes $S_1, \ldots, S_r$. Then $\boldsymbol{a}F$ is the union of $\boldsymbol{a}S_1, \ldots, \boldsymbol{a}S_r$, each of which is a simplex by ().

It remains to show (iii) ⇒ (ii). This will be shown by induction on the dimension of the Euclidean space containing the set. A compact locally conical subset of $\mathbb{R}^1$ is a finite union of points and closed line segments, each of which is a simplex. Now suppose all compact locally conical subsets of Euclidean spaces of dimension less than n are finite unions of simplexes. Let X be a compact locally conical subset of $\mathbb{R}^n$. To show that X is a finite union of simplexes, it suffices to show that for every $\boldsymbol{a} \in X$, an ϵ-star of $\boldsymbol{a}$ in X is a finite union of simplexes for some $\epsilon = \epsilon_a$. For then the sets $\mathring{N}_{\epsilon_a}(\boldsymbol{a}, X)$ form an open cover of X which has a finite subcover. So suppose $N_\epsilon(\boldsymbol{a}, X)$ is an ϵ-star. Let F be a proper face of the closed cube $\{\boldsymbol{x} \in \mathbb{R}^n : \rho(\boldsymbol{a}, \boldsymbol{x}) \leq \epsilon\}$. Then F is compact and locally conical, and thus $X \cap F$ is compact and locally conical by XIII.25.1.25.. Since $X \cap F$ is contained in a Euclidean space of dimension less than n, it is a finite union of simplexes. Then $\dot{N}_\epsilon(\boldsymbol{a}, X)$,

which is a finite union of the $X \cap F$, is a finite union of simplexes $S_1, \ldots, S_r$. But then $N_\epsilon(\boldsymbol{a}, X)$ is the union of $\boldsymbol{a}S_1, \ldots, \boldsymbol{a}S_r$, each of which is a simplex by ().

This completes the proof of XIII.25.1.23.. □

Here is a consequence which is not at all obvious directly from the definition:

XIII.25.1.29. COROLLARY. A finite intersection of polyhedra is a polyhedron.

PROOF: A finite intersection of polyhedra is compact, and locally conical by XIII.25.1.25.. 👍

XIII.25.1.30. PROPOSITION. Let X be a polyhedron in $\mathbb{R}^n$. Then the convex hull of X is a cell. In particular, a convex polyhedron is a cell.

PROOF: X is a finite union of cells $C_1, \ldots, C_r$. Let S be the set of all extreme points of $C_1, \ldots, C_r$. Then S is a finite set and the convex hull of S is the convex hull of X. 👍

We then obtain one direction of XIII.25.1.11.:

XIII.25.1.31. COROLLARY. Let X be a compact set in $\mathbb{R}^n$ which is a finite intersection of half-spaces. Then X is a cell.

PROOF: By XIII.25.1.25., X is locally conical, hence a polyhedron, and X is also convex. 👍

XIII.25.1.32. Let X be a subset of $\mathbb{R}^n$. If $\boldsymbol{a} \in X$, a *star* of $\boldsymbol{a}$ in X is a closed neighborhood of $\boldsymbol{a}$ in X which is a cone of the form $\boldsymbol{a}L$, where L is a compact subset of X not containing $\boldsymbol{a}$ ($L = \emptyset$ is allowed, in which case $\boldsymbol{a}L = \{\boldsymbol{a}\}$). An ϵ-star of $\boldsymbol{a}$ in X is a star with $L = \dot{N}_\epsilon(\boldsymbol{a}, X)$ provided this set is compact (e.g. if X is compact).

XIII.25.1.33. PROPOSITION. Let $X \subseteq \mathbb{R}^n$. Then every point of X has a star if and only if X is locally compact and locally conical.

PROOF: If X is locally compact and locally conical, and $\boldsymbol{a} \in X$, then $\dot{N}_\epsilon(\boldsymbol{a}, X)$ is compact for sufficiently small ϵ, so $\boldsymbol{a}$ has a star. Conversely, if $\boldsymbol{a}L$ is a star of $\boldsymbol{a}$, it is a compact neighborhood of $\boldsymbol{a}$, and $N_\epsilon(\boldsymbol{a}, X)$ is an ϵ-star for any $\epsilon < \rho_\infty(\boldsymbol{a}, L)$, so if every point of X has a star, then X is locally compact and locally conical. 👍

XIII.25.1.34. In PL topology, a polyhedron is defined to simply be a subset of $\mathbb{R}^n$ in which every point has a star, i.e. a locally compact locally conical subset. While it is useful in PL topology to consider locally conical sets (polyhedra) which are not compact, we will not need to do so, so to avoid having to constantly write "compact polyhedron" we will require polyhedra to be compact.

Enclosing Compact Sets in Polyhedra

XIII.25.1.35. THEOREM. Let X be a compact subset of $\mathbb{R}^n$, and U an open set in $\mathbb{R}^n$ containing X. Then there is a polyhedron $P \subseteq U$ with $X \subseteq P^\circ$.

PROOF: For each $x \in X$ there is a closed coordinate rectangle R_x centered at x contained in U. The interiors of the R_x cover X, so there is a finite subcover $\{R_{x_1}^\circ, \ldots, R_{x_m}^\circ\}$. Then $P = \cup_{k=1}^m R_{x_k}$ is the desired polyhedron. 👍

The picture for general n is complicated, but in the case $n = 2$ it is fairly simple, since the boundary of P is then a finite union of simple closed polygonal curves with edges parallel to the coordinate axes (which may intersect at finitely many vertices). Thus we have, using the Jordan Curve Theorem (XIII.19.1.20.):

XIII.25.1.36. COROLLARY. Let X be a compact subset of $\mathbb{R}^2$ and U an open set in $\mathbb{R}^2$ containing X. Then there are finitely many simple closed polygonal curves $\{\gamma_1, \ldots, \gamma_r\}$ such that each γ_k and its inside V_k is contained in U and $X \subseteq \cup_{k=1}^r V_k$.

XIII.25.1.37. With a little more work, it can be shown that under the hypotheses of XIII.25.1.35. there are finitely many disjoint smooth copies $\{\Sigma_1, \ldots, \Sigma_m\}$ of S^{n-1} whose insides (XIII.19.1.23.) V_k are disjoint, with all Σ_k and V_k contained in U, and with X contained in the union of the V_k. In the case of XIII.25.1.36. this amounts to replacing the γ_k by pairwise disjoint smooth curves (e.g. by "rounding off corners").

XIII.25.2. Complexes

XIII.25.2.1. DEFINITION. A *cell complex* in $\mathbb{R}^n$ is a finite collection $\mathcal{K}$ of cells in $\mathbb{R}^n$ with the following properties:

(i) If $C \in \mathcal{K}$ and F is a face in C, then $F \in \mathcal{K}$.

(ii) If $C_1, C_2 \in \mathcal{K}$ then $C_1 \cap C_2$ is a face of both C_1 and C_2 (we make the convention that $\emptyset$ is a face of every cell, so $C_1 \cap C_2 = \emptyset$ is allowed).

The *dimension* of $\mathcal{K}$ is maximum of the dimensions of the cells of $\mathcal{K}$.

The *vertices* of $\mathcal{K}$ are the 0-cells of $\mathcal{K}$ and the *edges* of $\mathcal{K}$ are the 1-cells.

If $\mathcal{K}$ is a cell complex in $\mathbb{R}^n$, the union of the cells in $\mathcal{K}$ is a polyhedron in $\mathbb{R}^n$ denoted $|\mathcal{K}|$.

A cell complex $\mathcal{K}$ is a collection of cells that fit together nicely. The cell complex $\mathcal{K}$ can be thought of as the polyhedron $|\mathcal{K}|$ with a specified nice subdivision of $|\mathcal{K}|$ into cells. The vertices of $\mathcal{K}$ are the extreme points of the cells in $\mathcal{K}$.

Infinite cell complexes are often used in topology, but for simplicity we will make the convention that cell complexes are finite.

XIII.25.2.2. DEFINITION. A *simplicial complex* in $\mathbb{R}^n$ is a cell complex in which every cell is a simplex. A *triangulation* of a polyhedron X is a simplicial complex $\mathcal{K}$ with $|\mathcal{K}| = X$.

A simplicial complex in $\mathbb{R}^n$ is often called a *geometric simplicial complex* to distinguish it from an abstract simplicial complex (XIII.25.2.12.), which is a combinatorial object.

We will work almost exclusively with simplicial complexes.

XIII.25.2.3. EXAMPLES. (i) Any cell C can be regarded as a cell complex $\mathcal{C}$ consisting of all faces of C. Then $C = |\mathcal{C}|$. In particular, any simplex S can be regarded as a simplicial complex $\mathcal{S}$ consisting of the faces of S. $\mathcal{S}$ is a triangulation of S.

(ii) Let X be a finite union of line segments in $\mathbb{R}^n$, such that any two of the line segments are either disjoint or have only an endpoint in common. Then X is a polyhedron which is the underlying space of a simplicial complex $\mathcal{K}$ consisting of the line segments and their endpoints (plus $\emptyset$); $\mathcal{K}$ is a triangulation of X. In particular, any (simple) polygon in $\mathbb{R}^2$, not including interior, is the underlying space of a simplicial complex. (A simple closed polygon including interior is also the underlying space of a simplicial complex, although this is somewhat harder to show ().)

Subcomplexes

XIII.25.2.4. DEFINITION. Let $\mathcal{K}$ be a cell complex. A *subcomplex* of $\mathcal{K}$ is a subset of $\mathcal{K}$ which is itself a cell complex.

A subset $\mathcal{K}'$ of a complex $\mathcal{K}$ is a subcomplex if and only if, whenever $C \in \mathcal{K}'$, then all faces of C are also in $\mathcal{K}'$. A subcomplex of a simplicial complex is a simplicial complex.

There are important special subcomplexes of any complex:

XIII.25.2.5. DEFINITION. Let $\mathcal{K}$ be a cell complex, and $d \in \mathbb{N} \cup \{0\}$. The collection $\mathcal{K}_d$ of all cells in $\mathcal{K}$ of dimension $\leq d$ is a subcomplex of $\mathcal{K}$ called the *d-skeleton* of $\mathcal{K}$.

Thus the 0-skeleton consists of the vertices, the 1-skeleton the vertices and edges, etc.

Subdivisions

XIII.25.2.6. DEFINITION. Let $\mathcal{K}$ and $\mathcal{K}'$ be cell complexes. $\mathcal{K}'$ is a *subdivision* of $\mathcal{K}$ if $|\mathcal{K}'| = |\mathcal{K}|$ and every cell in $\mathcal{K}'$ is contained in a cell in $\mathcal{K}$.

It is easily seen that $\mathcal{K}'$ is a subdivision of $\mathcal{K}$ if and only if each cell in $\mathcal{K}$ is a union of cells in $\mathcal{K}'$.

Simplicial Maps

XIII.25.2.7. DEFINITION. (i) Let $\mathcal{K}$ be a cell complex in $\mathbb{R}^n$. A function $f : |\mathcal{K}| \to \mathbb{R}^m$ is *simplicial* (with respect to $\mathcal{K}$), or a *simplicial map*, if the restriction of f to each cell of $\mathcal{K}$ is affine.

(ii) Let $\mathcal{K}$ and $\mathcal{K}'$ be cell complexes. A function $f : |\mathcal{K}| \to |\mathcal{K}'|$ is *simplicial* (with respect to $\mathcal{K}$ and $\mathcal{K}'$) if the image of each cell in $\mathcal{K}$ is contained in a cell of $\mathcal{K}'$ and the restriction of f to each cell of $\mathcal{K}$ is affine.

A *simplicial isomorphism* between cell complexes $\mathcal{K}$ and $\mathcal{K}'$ is a bijection f between $|\mathcal{K}|$ and $|\mathcal{K}'|$ such that f and f^{-1} are simplicial.

XIII.25.2.8. A simplicial map is continuous by (), since each cell is closed in $|\mathcal{K}|$ and the restriction to each cell is continuous. A composition of simplicial maps is simplicial.

XIII.25.2.9. Examples. (i) If $\mathcal{K}'$ is a subcomplex of $\mathcal{K}$, the inclusion map from $|\mathcal{K}'|$ to $|\mathcal{K}|$ is simplicial for $\mathcal{K}'$ and $\mathcal{K}$.

(ii) If $\mathcal{K}'$ is a subdivision of $\mathcal{K}$, the identity map on $|\mathcal{K}'| = |\mathcal{K}|$ is simplicial for $\mathcal{K}'$ and $\mathcal{K}$. But it is not simplicial for $\mathcal{K}$ and $\mathcal{K}'$ unless $\mathcal{K}' = \mathcal{K}$.

(iii) Let P be a (simple closed) polygon in $\mathbb{R}^2$, not including interior, with n sides. Then P can be regarded as a simplicial complex $\mathcal{P}$ with n 0-simplexes (the vertices) and n 1-simplexes (the sides). If Q is another polygon with n sides, regarded as a simplicial complex $\mathcal{Q}$, then $\mathcal{P}$ and $\mathcal{Q}$ are simplicially isomorphic.

But if P and Q are convex polygons with n sides and the interiors are included, regarded as cell complexes $\mathcal{P}$, $\mathcal{Q}$ with one 2-cell and n 1-cells and 0-cells as above, then $\mathcal{P}$ and $\mathcal{Q}$ are not generally simplicially isomorphic, and the conditions for simplicial isomorphism are restrictive: for example, if P is a square, then $\mathcal{P}$ and $\mathcal{Q}$ are simplicially isomorphic if and only if the quadrilateral Q is a parallelogram.

For simplicial complexes, conditions for simplicial isomorphism are much less stringent (XIII.25.2.10.).

XIII.25.2.10. Proposition. Let $\mathcal{K}$ and $\mathcal{K}'$ be simplicial complexes, and let ϕ be a one-one correspondence between the simplexes in $\mathcal{K}$ and the simplexes in $\mathcal{K}'$ with the property that if $S \in \mathcal{K}$ and F is a face of S, then $\phi(F)$ is a face of $\phi(S)$. Then there is a unique simplicial isomorphism $f : |\mathcal{K}| \to |\mathcal{K}'|$ such that $f(S) = \phi(S)$ for every $S \in \mathcal{K}$.

Proof: By counting numbers of faces, it follows that ϕ must give a bijection between the d-simplexes in $\mathcal{K}$ and the d-simplexes of $\mathcal{K}'$ for each d. Since the 0-simplexes are just the vertices, ϕ gives a one-one correspondence between the vertices of $\mathcal{K}$ and the vertices of $\mathcal{K}'$. If $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_n$ are the vertices of $\mathcal{K}$, set $\boldsymbol{w}_k = \phi(\boldsymbol{v}_k)$ for each k. Then if S is a simplex in $\mathcal{K}$ with vertices $\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_r}$, then $\phi(S)$ must be a simplex in $\mathcal{K}'$ with vertices $\boldsymbol{w}_{i_i}, \ldots, \boldsymbol{w}_{i_r}$, and we may define $f : S \to \phi(S)$ by

$$f\left(\sum_{k=1}^{r} \lambda_k \boldsymbol{v}_{i_k}\right) = \sum_{k=1}^{r} \lambda_k \boldsymbol{w}_{i_k} \ .$$

It is easily checked that if $\boldsymbol{x} \in |\mathcal{K}|$, this definition of $f(\boldsymbol{x})$ is independent of which of the simplexes of $\mathcal{K}$ containing $\boldsymbol{x}$ is used, and thus we obtain a well-defined function $f : |\mathcal{K}| \to |\mathcal{K}'|$ which obviously has the right properties. 👍

Barycentric Coordinates

XIII.25.2.11. The construction in the proof of XIII.25.2.10. can be thought of in another way. If $\mathcal{K}$ is a simplicial complex and $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_n$ are the vertices of $\mathcal{K}$, and $\boldsymbol{x} \in |\mathcal{K}|$, then for any simplex S of $\mathcal{K}$ containing $\boldsymbol{x}$ with vertices $\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_r}$, $\boldsymbol{x}$ has barycentric coordinates () λ_{i_k} with respect to $\boldsymbol{v}_{i_k}$. These barycentric coordinates are defined for any vertex of any simplex of $\mathcal{K}$ containing $\boldsymbol{x}$, and do not depend on which simplex containing $\boldsymbol{x}$ and that vertex is used. If $\boldsymbol{v}_j$ is a vertex of $\mathcal{K}$ and $\boldsymbol{x}$ is not contained in any simplex of $\mathcal{K}$ with $\boldsymbol{v}_j$ a vertex, set the barycentric coordinate $\lambda_j(\boldsymbol{x})$ equal to 0. Thus the barycentric coordinates

$\lambda_0(\boldsymbol{x}), \ldots, \lambda_n(\boldsymbol{x})$ of $\boldsymbol{x}$ are well defined for every vertex, and satisfy $\lambda_k(\boldsymbol{x}) \geq 0$ for all k and $\sum_{k=0}^{n} \lambda_k(\boldsymbol{x}) = 1$. We may symbolically write

$$\boldsymbol{x} = \sum_{k=0}^{n} \lambda_k(\boldsymbol{x}) \boldsymbol{v}_k$$

interpreted as follows: if the set of vertices $\boldsymbol{v}_k$ for which $\lambda_k(\boldsymbol{x}) > 0$ is $\{\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_r}\}$, then the simplex S with vertices $\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_r}$ is in $\mathcal{K}$ and $\boldsymbol{x}$ is the point of S with barycentric coordinates $\lambda_{i_1}(\boldsymbol{x}), \ldots, \lambda_{i_r}(\boldsymbol{x})$ with respect to $\boldsymbol{v}_{i_1}, \ldots, \boldsymbol{v}_{i_r}$.

Abstract Simplicial Complexes

The existence of barycentric coordinates (XIII.25.2.11.) in any simplicial complex allows the definition of simplicial complexes as combinatorial objects:

XIII.25.2.12. DEFINITION. An *abstract simplicial complex* is a collection $\mathcal{K}$ of subsets of a finite set $\{v_0, \ldots, v_n\}$ such that $\{v_k\} \in \mathcal{K}$ for all k and if $S \in \mathcal{K}$, then all subsets of S are also in $\mathcal{K}$. The v_k are called the *vertices* of $\mathcal{K}$.

If $\mathcal{K}$ and $\mathcal{K}'$ are abstract simplicial complexes with sets of vertices V and V' respectively, a *simplicial isomorphism* between $\mathcal{K}$ and $\mathcal{K}'$ is a bijection ϕ from V to V' such that $F \subseteq V$ is in $\mathcal{K}$ if and only if $\phi(F) \in \mathcal{K}'$.

A *geometric realization* of an abstract simplicial complex $\mathcal{K}$ is a (geometric) simplicial complex $\mathcal{K}'$ and a bijection ϕ from the set $\{v_0, \ldots, v_n\}$ of vertices of $\mathcal{K}$ to the vertices of $\mathcal{K}'$, say $\boldsymbol{w}_k = \phi(v_k)$, such that the simplex with vertices $\{\boldsymbol{w}_{i_1}, \ldots, \boldsymbol{w}_{i_r}\}$ is in $\mathcal{K}'$ if and only if $\{v_{i_1}, \ldots, v_{i_r}\} \in \mathcal{K}$.

The next result is obvious from XIII.25.2.10.:

XIII.25.2.13. PROPOSITION. Let $\mathcal{K}$ be an abstract simplicial complex. Then there is a unique simplicial isomorphism between any two geometric realizations of $\mathcal{K}$ which extends the identification of the vertices by the realizations. More generally, if $\mathcal{K}$ and $\mathcal{K}'$ are abstract simplicial complexes and ϕ is a simplicial isomorphism from $\mathcal{K}$ to $\mathcal{K}'$, then there is a unique simplicial isomorphism from any geometric realization of $\mathcal{K}$ to any geometric realization of $\mathcal{K}'$ extending the identification of the vertices via ϕ.

XIII.25.2.14. Every abstract simplicial complex has a geometric realization. In fact, if an ordering $v_0, v_1, \ldots, v_n$ on the vertices of an abstract simplicial complex $\mathcal{K}$ is fixed, there is a corresponding *standard realization* of $\mathcal{K}$ as a subcomplex $\mathcal{K}'$ of the standard n-simplex Σ_n: just let $\mathcal{K}'$ be the collection of all faces of Σ_n with vertices $\{\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_r}\}$ for sets $\{v_{i_1}, \ldots, v_{i_r}\}$ in $\mathcal{K}$.

Thus there is a natural one-one correspondence between simplicial isomorphism classes of abstract simplicial complexes and simplicial isomorphism classes of geometric simplicial complexes.

XIII.25.3. Triangulations

Recall (XIII.25.2.2.) that a *triangulation* of a polyhedron P is a simplicial complex $\mathcal{K}$ with $|\mathcal{K}| = P$.

XIII.25.3.1. THEOREM. Every polyhedron has a triangulation.

Triangulations of Topological Spaces

XIII.25.3.2. DEFINITION. Let X be a topological space (usually a compact Hausdorff space). A *weak triangulation* of X is a simplicial complex $\mathcal{K}$ and a continuous function $f : X \to |\mathcal{K}|$ such that every vertex of $\mathcal{K}$ is in the range of f. A weak triangulation f of X is a *triangulation* of X if f is a homeomorphism from X onto $|\mathcal{K}|$.

This terminology is consistent: if P is a polyhedron, then a triangulation of P in the simplicial sense (XIII.25.2.2.) is a triangulation in the topological sense (XIII.25.3.2.) when P is regarded as a topological space. The other direction is not well defined, since a polyhedron is more than a topological space: it also has a "local affine structure" (XIII.25.4.1.) which is not part of the topological structure. See () for further comments.

Note that a topological space with a triangulation must be homeomorphic to a polyhedron, hence a compact metrizable space. But spaces which are not compact and/or metrizable can have weak triangulations: for example, any constant function from a topological space X to $\mathbb{R}^n$ is a weak triangulation.

Triangulations and Partitions of Unity

Weak triangulations of a topological space X are effectively the same thing as finite partitions of unity on X.

XIII.25.3.3. Let X be a topological space and $\{f_0, \ldots, f_n\}$ a finite partition of unity on X. Define a continuous function f from X to the standard n-simplex Σ_n by

$$f(x) = \sum_{k=0}^{n} f_k(x)\mathbf{e}_k \ .$$

If Σ_n is regarded as a simplicial complex, then f is a weak triangulation of X. (The requirement that each f_k take the value 1 is precisely the same as the requirement that the range of f contain all vertices of Σ_n.) Note that the range of f may well be contained in a proper subcomplex of Σ_n containing all the vertices.

XIII.25.3.4. Conversely, let f be a weak triangulation from X to a simplicial complex $\mathcal{K}$, and let $\{\boldsymbol{v}_0, \ldots, \boldsymbol{v}_n\}$ be the vertices of $\mathcal{K}$. If $\boldsymbol{a} \in |\mathcal{K}|$, let $\lambda_k(\boldsymbol{a})$ be the barycentric coordinate of $\boldsymbol{a}$ with respect to $\boldsymbol{v}_k$ in $\mathcal{K}$ (). For each k set $f_k = \lambda_k \circ f$, i.e. for $x \in X$ set $f_k(x) = \lambda_k(f(x))$. Then the f_k give a finite partition of unity on X.

XIII.25.3.5. Suppose $f : X \to |\mathcal{K}|$ is a weak triangulation, and let $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_n$ be the vertices of $\mathcal{K}$ and $f_0, \ldots, f_n$ the corresponding partition of unity. For each k let V_k be the set of $\boldsymbol{a} \in |\mathcal{K}|$ with $\lambda_k(\boldsymbol{a}) > 0$; V_k is the set of points $\boldsymbol{a}$ of $|\mathcal{K}|$ such that every simplex in $\mathcal{K}$ containing $\boldsymbol{a}$ has $\boldsymbol{v}_k$ as a vertex. Then V_k is open in $|\mathcal{K}|$, and if $U_k = f^{-1}(V_k) = \{x \in X : f_k(x) > 0\}$, then $\mathcal{U} = \{U_0, \ldots, U_n\}$ is an open cover of X, and $\{f_0, \ldots, f_n\}$ is a partition of unity subordinate to $\mathcal{U}$.

XIII.25.3.6. So if X is a compact Hausdorff space, the theory of partitions of unity assures the existence of "fine" weak triangulations. In fact, we have:

XIII.25.3.7. THEOREM. Let X be a compact Hausdorff space and $\mathcal{W}$ an open cover of X. Then there is a surjective weak triangulation $f : X \to |\mathcal{K}|$ such that the cover $\mathcal{U}$ of XIII.25.3.5. is a refinement of $\mathcal{W}$.

PROOF: By passing to a subcover, we may assume $\mathcal{W}$ is finite. Let $\mathcal{V}$ be an open star-refinement of $\mathcal{W}$, which we may also assume finite, and $\{f_0, \ldots, f_n\}$ a partition of unity subordinate to $\mathcal{V}$. The corresponding weak triangulation $f : X \to \mathcal{K}' = \Sigma_n$ works except that f may not be surjective.

We modify f in stages and reduce $\mathcal{K}'$ to a subcomplex, beginning with the largest simplex. To do the modification on a simplex S which is not a face of another simplex in $\mathcal{K}'$, suppose the interior of S is not entirely contained in the range of f (if it is, no modification for this S is necessary). If $\boldsymbol{a}$ is an interior point of S not contained in the range, there is a retraction r (e.g. radial projection from $\boldsymbol{a}$) from $S \setminus \{\boldsymbol{a}\}$ onto its boundary. Replace $\mathcal{K}'$ by $\mathcal{K}' \setminus \{S\}$ and f by $r \circ f$. Then the inverse image of the star of any vertex in the reduced complex under $r \circ f$ is contained in the inverse image of the star of that vertex in $\mathcal{K}'$ under f, so is contained in one of the sets of $\mathcal{W}$. The modified $r \circ f$ agrees with f on any $x \in X$ for which $f(x)$ is not in the interior of S. Repeat the reduction through all simplexes in $\mathcal{K}'$ which are not faces of remaining simplexes after the previous reduction. Since all vertices of $\mathcal{K}'$ are in the range of f at every stage, after a finite number of steps a subcomplex $\mathcal{K}$ and modified f will be surjective. 👍

XIII.25.3.8. COROLLARY. Let X be a compact metrizable space, ρ a fixed metric on X, and $\epsilon > 0$. Then there is a surjective weak triangulation $f : X \to |\mathcal{K}|$ of X, such that for every vertex $\boldsymbol{v}_k$ of $|\mathcal{K}|$, the ρ-diameter of $f^{-1}(V_k)$ is less than ϵ.

PROOF: Let $\mathcal{W}$ be the open cover of X by balls of ρ-diameter $< \epsilon$. 👍

The Nerve of an Open Cover

XIII.25.4. Piecewise-Linear Maps

XIII.25.4.1. If X is a polyhedron and $\boldsymbol{a} \in X$, then for any $\boldsymbol{b} \in X$ sufficiently close to $\boldsymbol{a}$ (within a star of $\boldsymbol{a}$ in X) there is an unambiguous notion of the line segment from $\boldsymbol{a}$ to $\boldsymbol{b}$ in X, and of $\lambda\boldsymbol{a} + (1-\lambda)\boldsymbol{b}$ for $0 \le \lambda \le 1$, i.e. X has a "local affine structure." (Existence of such a structure does not characterize polyhedra, however; for example, any convex set also has such a local affine structure.)

The local affine structure is "continuous":

XIII.25.4.2. PROPOSITION. Let X be a subset of $\mathbb{R}^n$, $\boldsymbol{a} \in X$, and $\boldsymbol{a}L$ a star of $\boldsymbol{a}$ in X (with L compact). If $(\lambda_n\boldsymbol{a} + (1-\lambda_n)\boldsymbol{b}_n)$ is a sequence in $\boldsymbol{a}L$ (with $0 \le \lambda_n \le 1$ and $\boldsymbol{b}_n \in L$ for all n), then $\lambda_n\boldsymbol{a} + (1-\lambda_n)\boldsymbol{b}_n \to \lambda\boldsymbol{a} + (1-\lambda)\boldsymbol{b} \in \mathbb{R}^n$ with $\boldsymbol{b} \in L$ if and only if $\lambda_n \to \lambda$ and, if $\lambda \ne 1$, $\boldsymbol{b}_n \to \boldsymbol{b}$.

PROOF: If $\lambda_n \to \lambda$ and $\boldsymbol{b}_n \to \boldsymbol{b}$, then $\lambda_n\boldsymbol{a} + (1-\lambda_n)\boldsymbol{b}_n \to \lambda\boldsymbol{a} + (1-\lambda)\boldsymbol{b}$, and $\boldsymbol{b} \in L$ since L is closed in $\mathbb{R}^n$. Conversely, suppose $\lambda_n\boldsymbol{a} + (1-\lambda_n)\boldsymbol{b}_n \to \lambda\boldsymbol{a} + (1-\lambda)\boldsymbol{b}$ with $\boldsymbol{b} \in L$. Since $[0,1]$ and L are compact, any subsequence of $(\lambda_n\boldsymbol{a}+(1-\lambda_n)\boldsymbol{b}_n)$ has a subsubsequence $(\lambda_{k_n}\boldsymbol{a}+(1-\lambda_{k_n})\boldsymbol{b}_{k_n})$ such that (λ_{k_n}) and $(\boldsymbol{b}_{k_n})$ both converge, say to μ and $\boldsymbol{c}$ respectively, with $0 \le \mu \le 1$ and $\boldsymbol{c} \in L$. Then $(\lambda_{k_n}\boldsymbol{a}+(1-\lambda_{k_n})\boldsymbol{b}_{k_n}) \to \mu\boldsymbol{a}+(1-\mu)\boldsymbol{c}$. By uniqueness of limits in $\mathbb{R}^n$, $\mu\boldsymbol{a} + (1-\mu)\boldsymbol{c} = \lambda\boldsymbol{a} + (1-\lambda)\boldsymbol{b}$, and by uniqueness of representation of points of $\boldsymbol{a}L$, we must have $\mu = \lambda$ and, if $\lambda \ne 1$, $\boldsymbol{c} = \boldsymbol{b}$. Thus every subsequence of (λ_n) has a subsubsequence

converging to λ, and if $\lambda \neq 1$ every subsequence of $(\boldsymbol{b}_n)$ has a subsubsequence converging to $\boldsymbol{b}$, so $\lambda_n \to \lambda$ and, if $\lambda \neq 1$, $\boldsymbol{b}_n \to \boldsymbol{b}$ by ().

We can use the local affine structure to define piecewise-linear maps between polyhedra:

XIII.25.4.3. DEFINITION. Let X and Y be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. A function $f : X \to Y$ is *piecewise-linear (PL)* if for every $\boldsymbol{a} \in X$ there are $\delta > 0$, $\epsilon > 0$ such that

(i) $\boldsymbol{a}$ has a δ-star in X and $f(\boldsymbol{a})$ has an ϵ-star in Y

(ii) $f(N_\delta(\boldsymbol{a}, X)) \subseteq N_\epsilon(f(\boldsymbol{a}), Y)$

(iii) $f(\lambda\boldsymbol{a} + (1-\lambda)\boldsymbol{b}) = \lambda f(\boldsymbol{a}) + (1-\lambda)f(\boldsymbol{b})$ for all $\boldsymbol{b} \in N_\delta(\boldsymbol{a}, X)$, $0 \leq \lambda \leq 1$.

Such a function might be more properly called something like *locally affine*, but the name "piecewise-linear" is standard, and is appropriate since X can indeed be broken up into pieces on which f is linear (affine) (XIII.25.4.6.).

The next fact may seem obvious, but requires proof:

XIII.25.4.4. PROPOSITION. Let X and Y be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f : X \to Y$ a PL function. Then f is continuous.

PROOF: Let $(\boldsymbol{c}_n)$ be a sequence in X with $\boldsymbol{c}_n \to \boldsymbol{a}$. Let $\delta > 0$ and $\epsilon > 0$ be as in XIII.25.4.3. for $\boldsymbol{a}$. Passing to a tail, we may assume $\boldsymbol{c}_n \in N_\delta(\boldsymbol{a}, X)$ for all n. Write $\boldsymbol{c}_n = \lambda_n\boldsymbol{a} + (1-\lambda_n)\boldsymbol{b}_n$ for $0 \leq \lambda_n \leq 1$ and $\boldsymbol{b}_n \in N_\delta(\boldsymbol{a}, X)$. Since $\boldsymbol{c}_n \to \boldsymbol{a}$, we have $\lambda_n \to 1$ by XIII.25.4.2.. We have

$$f(\boldsymbol{c}_n) = \lambda_n f(\boldsymbol{a}) + (1-\lambda_n)f(\boldsymbol{b}_n) \to f(\boldsymbol{a}) + \boldsymbol{0} = f(\boldsymbol{a})$$

since $\lambda_n \to 1$ and $(f(\boldsymbol{b}_n))$ is a bounded sequence.

XIII.25.4.5. PROPOSITION. Let X and Y be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f : X \to Y$ a function. Then f is PL if and only if the graph $\Gamma(f)$ is a polyhedron in $\mathbb{R}^{n+m}$.

PROOF: It is easy to verify that $\Gamma(f)$ is compact and locally conical if and only if f is PL.

XIII.25.4.6. COROLLARY. Let X and Y be polyhedra in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and $f : X \to Y$ a PL function. Then there is a triangulation of X such that f is simplicial for the triangulation. Conversely, such a map is PL.

PROOF: If f is PL, triangulate the polyhedron $\Gamma(f)$. The projection map from $\Gamma(f)$ to X is affine, and is a homeomorphism, so the projections of the simplexes in the triangulation of $\Gamma(f)$ are simplexes in X and give a triangulation of X as desired. The converse is obvious.

This result justifies the name "piecewise-linear."

XIII.25.5. Exercises

XIII.25.5.1. (a) Show that every affine subspace of $\mathbb{R}^n$ is a finite intersection of (closed) half-spaces.

(b) Let C be a cell in $\mathbb{R}^n$ with vertices $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$. Suppose C generates $\mathbb{R}^n$ as an affine subspace. For $1 \leq k \leq m$, let C_k be the cone (in the vector space sense) generated by C with vertex $\boldsymbol{v}_k$, i.e. the union of all rays from $\boldsymbol{v}_k$ passing through other points of C. Show that $C = C_1 \cap \cdots \cap C_m$. [Use an argument like the one in the proof of XIII.25.1.28..]

(c) Show that each C_k is a finite intersection of half-spaces.

(d) Conclude that every cell is a finite intersection of half-spaces. Use this to give a direct proof of (i) $\Rightarrow$ (iii) of XIII.25.1.23..

XIII.25.5.2. Let T be an open triangle in $\mathbb{R}^2$ with one vertex added. Show that T is locally conical, but not a polyhedron in the PL sense (XIII.25.1.34.) since it is not locally compact.

XIII.26. Absolute Retracts and Absolute Neighborhood Retracts

There is an important class of nicely behaved spaces called *absolute retracts* and, more generally, *absolute neighborhood retracts*. We can work within various classes of spaces; we work first with compact Hausdorff spaces (compacta), and in Section () with more general classes.

The two standard references for this subject are [Bor67] and [**?**]. [vM89] is also a good reference. These references contain much more than we present here.

XIII.26.1. Retractions

XIII.26.1.1. Definition. Let X be a topological space and $Y \subseteq X$. A *retraction* of X to Y is a map (continuous function) $r : X \to Y$ with $r(y) = y$ for all $y \in Y$. Y is a *retract* of X if there is a retraction from X to Y.

XIII.26.1.2. Examples. (i) If X is any topological space, the identity map ι_X is a retraction from X to X.

(ii) Let X be a topological space, and $p \in X$. Then the constant function $f(x) = p$ for all $x \in X$ is a retraction of X to $\{p\}$. Thus every singleton subset of X is a retract of X.

(iii) Let X and Y be topological spaces, and $x_0 \in X$, $y_0 \in Y$. Then the projection functions $f((x, y)) = (x, y_0)$ and $g((x, y)) = (x_0, y)$ are retractions of $X \times Y$ (with the product topology) to $X \times \{y_0\}$ and $\{x_0\} \times Y$ respectively. Thus the horizontal and vertical slices of $X \times Y$ are retracts of $X \times Y$.

(iv) Let X be the closed unit ball in $\mathbb{R}^n$ ($n \geq 1$), and Y the unit sphere. The Brouwer Fixed-point Theorem implies that there is no retraction from X to Y, i.e. Y is not a retract of X.

XIII.26.1.3. A retraction from X to a subspace is just an idempotent map, a map $f : X \to X$ with $f \circ f = f$: such a map is a retraction onto its range.

XIII.26.1.4. Proposition. If X is Hausdorff, any retract of X is a closed subset.

Proof: Let Y be a retract of X, and $r : X \to Y$ a retraction. Let $x_0 \in X \setminus Y$, and set $y_0 = r(x_0)$. Then $y_0 \neq x_0$ since $y_0 \in Y$. Let U and V be disjoint open neighborhoods of x_0 and y_0 respectively. Then by continuity of r there is an open neighborhood W of x_0, contained in U, such that $r(W) \subseteq V$. If $x \in W$, then $r(x) \in V$, which is disjoint from W, so $r(x) \neq x$ and $x \notin Y$. Thus $X \setminus Y$ is open. 👍

This result can fail if X is not Hausdorff (e.g. XIII.26.1.2.(ii) if X is not T_1; see also Exercise XIII.26.6.1.).

XIII.26.1.5. One can more generally define an *r-map* between topological spaces X and Y to be a map $f : X \to Y$ which has a (continuous) right inverse (cross-section), a map $g : Y \to X$ such that $f \circ g = \iota_Y$. A retraction is an r-map. If f is an r-map from X to Y, f must be surjective, g must be a homeomorphism from Y to $g(Y)$, and $g \circ f$ a retraction from X to $g(Y)$. Thus an r-map is precisely a retraction composed with a homeomorphism.

A composition of r-maps is an r-map. A homeomorphism is an r-map. Thus the class of topological spaces, or any subclass, with r-maps as morphisms, forms a category.

Deformation Retractions

Retractions must be carefully distinguished from deformation retractions:

XIII.26.1.6. Definition. Let X be a topological space, and $Y \subseteq X$. A *deformation retraction* from X to Y is a map $h : X \times [0,1] \to Y$ such that $h(x,0) = x$ and $h(x,1) \in Y$ for all $x \in X$ and $h(y,1) = y$ for all $y \in Y$. The map h is a *strong deformation retraction* if $h(y,t) = y$ for all $y \in Y$ and all $t \in [0,1]$. Y is a [strong] deformation retract of X if there is a [strong] deformation retraction form X to Y.

XIII.26.1.7. A deformation retraction from X to Y "continuously moves" all points of X into Y, with points of Y ending up unmoved. A strong deformation retraction is one where the points of Y are never moved at all. If we set $f_t : X \to Y$ by $f_t(x) = h(x,t)$, then $f_0 = \iota_X$ and f_1 is a retraction of X to Y. Thus a deformation retraction from X to Y can be thought of as a retraction from X to Y which is homotopic to the identity on X, and a strong deformation retraction is a retraction which is homotopic to the identity on X relative to the points of Y ().

XIII.26.1.8. A deformation retract of X is a retract of X. But being a deformation retract is much more restrictive than being a retract. For example, if X is a space which is not contractible (e.g. a space which is not connected, or S^1), and $p \in X$, then $\{p\}$ is a retract of X (XIII.26.1.2.(ii)) but not a deformation retract. In fact, a singleton $\{p\}$ in a topological space X is a [strong] deformation retract of X if and only if X is [strongly] contractible to p.

XIII.26.2. Absolute Retracts

In this section, we will deal only with compacta (compact Hausdorff spaces). The primary interest will be the metrizable case, but this restriction is not necessary for the basic theory.

XIII.26.2.1. Definition. A compactum X is an *absolute retract (AR)* if, whenever X is embedded as a (necessarily closed) subset of a compactum Z, X is a retract of Z (i.e. there is a retraction of Z onto X).

XIII.26.2.2. Examples. (i) By XIII.26.1.2.(ii), a singleton space is an AR.

(ii) If $[0,1]$ is embedded in a compactum Z, the identity map on $[0,1]$ extends to a map from Z to $[0,1]$ by the Tietze Extension Theorem (), giving a retraction of Z to $[0,1]$. Thus $[0,1]$ is an AR.

(iii) A two-point space with the discrete topology is not an AR: embed $\{0,1\}$ into $[0,1]$.

(iv) A circle is not an AR: it is not a retract of a disk.

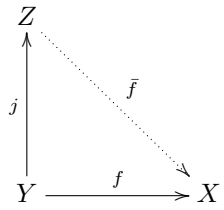
We will obtain many more examples later.

Absolute Extensors

Here is a closely related (*extremely* closely related – see XIII.26.2.6.) but seemingly more restrictive notion:

XIII.26.2.3. DEFINITION. A compactum X is an *absolute extensor (AE)* if, whenever Z is a compactum and Y a closed subspace of Z, every map $f : Y \to X$ extends to a map $\bar{f} : Z \to X$.

In other words, the following commutative diagram can always be completed in the category of compacta, where j is the inclusion map:



i.e. X is an injective object in this category.

XIII.26.2.4. An AE is an AR: if X is an AE, embedded in Z, extend ι_X to a map from Z to X.

XIII.26.2.5. EXAMPLES. (i) A singleton is trivially an AE.

(ii) The Tietze Extension Theorem implies that $[0, 1]$ is an AE.

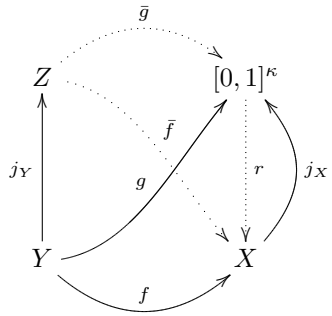
(iii) An arbitrary product of AEs is an AE (just extend every coordinate function; Tikhonov's Theorem () is needed to conclude that the product is a compactum). In particular, any cube (product of $[0, 1]$'s) is an AE.

The main theorem is:

XIII.26.2.6. THEOREM. A compactum is an AE if and only if it is an AR.

PROOF: We have already observed that an AE is an AR. Conversely, let X be an AR. Then X embeds in a cube $[0, 1]^\kappa$ (), and there is a retraction $r : [0, 1]^\kappa \to X$. Let Z be a compactum and Y a closed subspace of Z, and $f : Y \to X$ a map. Regard f as a function g from Y to $[0, 1]^\kappa$. Since $[0, 1]^\kappa$ is an AE, g extends to a $\bar{g}$ from Z to $[0, 1]^\kappa$. Then $\bar{f} = r \circ \bar{g} : Z \to X$ is an extension of f to Z.

The following commutative diagram summarizes the construction, where j_X and j_Y are the embeddings:



The next result has a simple direct proof, but follows immediately from XIII.26.2.6. and XIII.26.2.5.(iii):

XIII.26.2.7. Corollary. An arbitrary product of ARs is an AR.

A proof very similar to the proof of XIII.26.2.6. shows:

XIII.26.2.8. Proposition. A retract of an AR is an AR.

XIII.26.2.9. Corollary. Let X be a compactum. Then X is an AR if and only if it is (homeomorphic to) a retract of a cube.

Proof: Since any cube is an AR, a retract of a cube is an AR. Conversely, any compactum X embeds in a cube, and if X is an AR there is a retraction onto the image. 👍

Strong Contractibility

An important property of ARs is that they are strongly contractible:

XIII.26.2.10. Theorem. Let X be an AR and $p \in X$. Then X is strongly contractible to p.

Proof: Let Z be the cone over X: $Z = X \times [0,1]$ with $X \times \{1\}$ collapsed to a point. Let

$$Y = (X \times \{0\}) \cup (\{p\} \times [0,1]) \subseteq Z \ .$$

Y consists of the base of the cone plus a vertical spike from p to the vertex (Fig. ()). There is an obvious map $f : Y \to X$ given by $f((x,0)) = x$ for all $x \in X$ and $f((p,t)) = p$ for $t \in [0,1]$. An extension of f to Z is a strong contraction of X to p. 👍

A similar proof shows more generally [take $X_0 = \{p\}$ and Z_0 the vertical spike]:

XIII.26.2.11. Proposition. Let X be an AR, and X_0 a closed subspace. Let Z be a compactum containing X, and Z_0 a closed set in Z with $Z_0 \cap X = X_0$. Then any retraction from Z_0 to X_0 extends to a retraction from Z to X. In particular, if X_0 is also an AR, there is a retraction r from Z to X with $r(Z_0) = X_0$.

Proof: Let r_0 be a retraction from Z_0 to X_0. Let $Y = X \cup Z_0$; then Y is a closed subset of Z. Define $f : Y \to X$ by

$$f(y) = \begin{cases} y & \text{if} \quad y \in X \\ r_0(y) & \text{if} \quad y \in Z_0 \end{cases} \ .$$

It is easily checked that f is continuous, and an extension of f to Z is the desired retraction. 👍

The next property follows from XIII.26.2.10. and XV.1.1.21.:

XIII.26.2.12. PROPOSITION. An AR X is locally strongly contractible to any point of X.

Since a [locally] contractible space is [locally] path-connected, we have:

XIII.26.2.13. COROLLARY. An AR is path-connected and locally path-connected.

Homotopies

Any two maps from any compactum into an AR are homotopic. In fact, any two maps from any space into a contractible space are homotopic. But ARs have a much stronger property as a target space:

XIII.26.2.14. PROPOSITION. Let X be an AR, Z a compactum, Y a closed subspace of Z, and f_0, f_1 maps from Z to X. Suppose $\theta : Y \times [0,1]$ is a homotopy from $f_0|_Y$ to $f_1|_Y$. Then there is a homotopy $h : Z \times [0,1] \to X$ from f_0 to f_1 extending θ.

PROOF: Let $W = (Z \times \{0\}) \cup (Z \times \{1\}) \cup (Y \times [0,1])$. W is a closed subset of $Z \times [0,1]$, and there is an obvious map $\phi : W \to X$ which is f_0 on $Z \times \{0\}$, f_1 on $Z \times \{1\}$, and θ on $Y \times [0,1]$. Extend ϕ to $h : Z \times [0,1] \to X$. 👍

The following special case is especially important:

XIII.26.2.15. COROLLARY. Let X be an AR, Z a compactum, Y a closed subspace of Z, and f_0, f_1 maps from Z to X. Suppose f_0 and f_1 agree on Y. Then there is a homotopy $h : Z \times [0,1] \to X$ from f_0 to f_1 with $h(y,t) = f_0(y) = f_1(y)$ for all $y \in Y$, $t \in [0,1]$.

Continuous Strong Contractions

XIII.26.2.16. XIII.26.2.15. can be used to extend XIII.26.2.10.: not only can an AR X be strongly contracted to any $p \in X$, but the strong contractions can be made to "vary continuously" as p varies, as follows.

Let X be an AR. Apply XIII.26.2.15. with $Z = X \times X$ and Y the diagonal. Define f_0, f_1 by $f_0(x,y) = x$ and $f_1(x,y) = y$ (coordinate projections from $X \times X$ to X). Then f_0 and f_1 agree on Y. Let h be a homotopy from f_0 to f_1 leaving Y pointwise fixed, i.e. $h(x,y,0) = x$ and $h(x,y,1) = y$ for all $x, y \in X$, and $h(y,y,t) = y$ for all $y \in X$ and $t \in [0,1]$. A space with such an h is called *equiconnected* ([Fox43], [Bor69]).

If y is fixed and h is regarded as a function of x and t, i.e. $h_y(x,t) = h(x,y,t)$, then h_y is a strong contraction of X to y; and the h_y vary continuously in the strongest reasonable sense.

The Fixed-Point Property

Every AR has the fixed-point property (XIII.21.6.1.). This follows from:

XIII.26.2.17. PROPOSITION. Let Z be a topological space, and X a retract of Z. If Z has the fixed-point property, then so does X.

PROOF: Let j be the inclusion map from X to Z, and r a retraction from Z to X. Let $\phi : X \to X$ be a map. Define $\psi = j \circ \phi \circ r : Z \to Z$. Then ψ has a fixed point x. Since the range of ψ is contained in X, $x \in X$; and $\psi(x) = \phi(x)$ since $r(x) = x$. 👍

XIII.26.2.18. COROLLARY. Every AR has the fixed-point property.

PROOF: Every AR is a retract of a cube, which has the fixed-point property by XIII.21.6.3.. 👍

The Metrizable Case

Not all ARs by definition XIII.26.2.1. are metrizable: for example, any cube, no matter how large, is an AR. But metrizable ARs are particularly nice; in fact, many references require ARs to be metrizable. So for the rest of this section, "AR" will mean "metrizable AR".

The first observation is that we can work exclusively with metrizable compacta:

XIII.26.2.19. PROPOSITION. Let X be a metrizable compactum. The following are equivalent:
(i) X is an AR (by definition XIII.26.2.1.).
(ii) X is (homeomorphic to) a retract of the Hilbert cube.
(iii) Whenever X is embedded in a metrizable compactum Z, there is a retraction from Z to X.

PROOF: (i) ⇒ (iii) is trivial. (iii) ⇒ (ii) since X embeds in the Hilbert cube (), which is a metrizable compactum. And (ii) ⇒ (i) by XIII.26.2.8. since the Hilbert cube is an AR. 👍

As a sample of what can be proved in the metrizable case, we have the following assembly result:

XIII.26.2.20. THEOREM. Let X be a metrizable compactum, X_1 and X_2 closed subspaces with $X_0 = X_1 \cap X_2$.
(i) If X_0, X_1, X_2 are ARs, then X is an AR.
(ii) If X and X_0 are ARs, then X_1 and X_2 are ARs.

PROOF: (i): Embed X in a metrizable compactum Z. Since Z is completely normal (XIII.10.1.1.), there are closed subspaces Z_1, Z_2 of Z such that $Z_j \cap X = X_j$ for $j = 1, 2$ (XIII.10.6.12.). Then $Z_0 = Z_1 \cap Z_2$ satisfies $Z_0 \cap X = X_0$. There is a retraction r_0 from Z_0 to X_0. This retraction extends to retractions r_1 from Z_1 to X_1 and r_2 from Z_2 to X_2 by XIII.26.2.11.. These retractions combine to give a retraction from Z to X.
(ii): There are retractions r_1 from X_1 to X_0 and r_2 from X_2 to X_0. Set

$$r(x) = \begin{cases} x & \text{if} \quad x \in X_1 \\ r_2(x) & \text{if} \quad x \in X_2 \end{cases} .$$

Then r is a retraction from X to X_1. Since X is an AR, X_1 is an AR by XIII.26.2.8.. Similarly X_2 is an AR. 👍

XIII.26.2.21. The proof of (ii) does not use metrizability and works equally well in the nonmetrizable case. But the proof of (i) uses complete normality, which does not hold in general in nonmetrizable compacta (XIII.10.6.11.(d)). In fact, (i) can fail in the nonmetrizable case (Exercise XIII.26.6.7.).

XIII.26.2.22. This result can be rephrased to say that if two (metrizable) ARs are joined by identifying sub-ARs of each, the join is also an AR. See XIII.26.2.23. for a generalization. An important special case is that if two ARs are joined by a single point, the join is an AR. Applying this repeatedly, we obtain that any finite tree is an AR. Actually, one-dimensional (metrizable) ARs can be characterized: they are precisely the *dendrites*, metrizable compacta which are connected and locally connected (hence path-connected and locally path-connected by the Hahn-Mazurkiewicz Theorem) and contain no simple closed curves, or alternatively are unicoherent (). See XIII.19.5.1., XIII.19.6.10., and XIII.19.6.11. for some interesting examples; another example is a *countable star*, the one-point compactification of a sequence of copies of $(0, 1]$. Higher-dimensional ARs are much more complicated, and no complete description is known. Some pathological ones can be found in [Bor67]; cf. XIII.19.5.3..

A harder and stranger generalization of XIII.26.2.20.(i), due to J. H. C. Whitehead [], concerns identification spaces. See e.g. [Bor67, V.9.1] for the proof.

XIII.26.2.23. Theorem. Let X_1 and X_2 be metrizable ARs, and X_0 a subset of X_1 which is also an AR. Let $f : X_0 \to X_2$ be a continuous function. Then $X_1 \cup_f X_2$ () is an AR.

XIII.26.2.24. Note that ϕ need not be injective or surjective (the injective case is XIII.26.2.20.(i)), and $f(X_0)$ need not be an AR; in fact, it can be a very complicated set (cf. XIII.26.2.25.(ii)). Even if $f(X_0)$ is an AR, the identification space can be bizarre (cf. XIII.26.2.25.(i), XIII.26.2.25.(iii)).

XIII.26.2.25. Examples. (i) Let X_1 and X_2 be copies of $[0, 1]^2$. Let X_0 be an arc, say one edge, in X_1. Let f be a space-filling curve, a surjective map of X_0 onto X_2. Then $X = X_1 \cup_f X_2$ is an AR. X is a square in which X_0 has been "expanded" into a second square. Actually X_2 can be any metrizable AR, since any metrizable AR is a continuous image of $[0, 1]$ by XIII.26.2.13. and the Hahn-Mazurkiewicz Theorem ().

(ii) In (i), $f(X_0)$ is an AR. Instead, let Y be a copy of the Sierpiński carpet () inside X_2. There is a surjective function f from X_0 onto Y by the Hahn-Mazurkiewicz Theorem. The identification space is again an AR, although X_0 has become a copy of the Sierpiński carpet, which is far from an AR.

(iii) An interesting special case is when X_2 is a single point and f is a constant function. The identification space then amounts to taking the quotient of X_1 by collapsing X_0 to a point. So a quotient of an AR by collapsing a sub-AR to a point is again an AR. An interesting example is to consider a wild arc (XIII.19.5.39.) embedded in a 3-ball. Collapsing the arc to a point p gives a 3-dimensional AR X with complicated local topology around p: for example, no deleted neighborhood of p is simply connected; thus X is not locally Euclidean at p.

Example (iii) can be generalized, due to the following consequence of a theorem of S. Smale (see [Bor67, V.12.1] for a proof):

XIII.26.2.26. THEOREM. Let Z be a (metrizable) AR, X a (necessarily metrizable) finite-dimensional compactum, and $f : Z \to X$ a surjective map. If $f^{-1}(\{x\})$ is an AR for every $x \in X$, then X is also an AR.

Using (), this can be rephrased:

XIII.26.2.27. COROLLARY. Let Z be a (metrizable) AR, and $\sim$ a usc decomposition of Z into sub-ARs. Then the quotient space $X = Z/\sim$, if finite-dimensional, is also an AR.

Using this, an enormous supply of pathological ARs can be constructed; see e.g. Exercises ().

XIII.26.3. Absolute Neighborhood Retracts

Absolute retracts are nice, but the class is too restrictive to include many spaces of interest even though there are many pathological examples (e.g. ARs must be contractible). There is a much larger class of absolute neighborhood retracts (ANRs), spaces which behave locally like ARs, which include most compacta commonly considered in topology, including all finite CW-complexes (in particular all polyhedra) and all compact topological manifolds with or without boundary, and more generally all finite-dimensional locally contractible metrizable compacta.

Much of the theory of ANRs is parallel to the theory of ARs, with similar but often somewhat more complicated proofs.

XIII.26.3.1. DEFINITION. Let X be a compactum. X is an *absolute neighborhood retract (ANR)* if, whenever X is embedded in a compactum Z, there is a retraction from an open neighborhood U of X in Z to X.

XIII.26.3.2. EXAMPLES. (i) Any AR is an ANR.

(ii) A finite separated union of ANRs is an ANR. In particular, a finite separated union of ARs (e.g. a finite discrete space) is an ANR.

XIII.26.3.3. Since any compactum embeds in a cube, any ANR is a retract of an open set in a cube. The converse also holds ().

Since every open neighborhood of a closed set in a compactum contains a compact neighborhood, the definition can be rephrased:

XIII.26.3.4. PROPOSITION. A compactum X is an ANR if and only if, whenever X is embedded in a compactum Z, there is a retraction from a compact neighborhood U of X in Z to X.

Absolute Neighborhood Extensors

As with ARs, we can extend the definition:

XIII.26.3.5. DEFINITION. A compactum X is an *absolute neighborhood extensor (ANE)* if, whenever Z is a compactum and Y a closed subspace, and $f : Y \to X$ is a map, then f can be extended to an open neighborhood of Y in Z.

XIII.26.3.6. THEOREM. A compactum is an ANR if and only if it is an ANE.

PROOF: If X is an ANE, embedded in a compactum Z, extend the identity map of X to a map $\bar{f}$ from some open neighborhood U of X in Z to X. Then g is a retraction from U to X. So X is an ANR.

Conversely, suppose X is an ANR. Embed X in a cube K. There is a retraction r of an open set V in K containing X onto X. If Z is a compactum and Y a closed subspace of Z, and $f : Y \to X$ is a map, regard f as a map from Y to K. Since K is an AR, f extends to $g : Z \to K$. Set $U = g^{-1}(V)$; then U is an open set in Z containing Y. Set $\bar{f} = r \circ g|_U$; then $\bar{f}$ maps U into X and extends f to U. 👍

A similar argument shows:

XIII.26.3.7. PROPOSITION. A compactum which is a retract of an open set in an ANR is an ANR.

PROOF: Let X be an ANR, V an open set in X, and r a retraction from V to a closed subset X_0 of X. Let Z be a compactum, Y a closed subset of Z, and $f : Y \to X_0$ a map. Regard f as a map from Y to X. Then f extends to a map g from an open set W in Z containing Y to X. Set $U = g^{-1}(V)$; then U is open in W, hence in Z, and g maps U into V. Then $\bar{f} = r \circ g|_U$ maps U into X_0 and extends f. 👍

XIII.26.3.8. EXAMPLE. The n-sphere S^n is a radial retract of $B^{n+1} \setminus \{0\}$, and B^{n+1} is an AR; thus S^n is an ANR.

This will be greatly extended ().

Local Strong Contractibility

An ANR need not be contractible, or even connected. However:

XIII.26.3.9. THEOREM. An ANR is locally strongly contractible.

PROOF: We argue similarly to the proofs of XIII.26.2.10. and XV.1.1.21.. Let X be an ANR, and $p \in X$. Let Y be the subspace $(X \times \{0\}) \cup (\{p\} \times [0,1])$ of CX. The obvious map $f : Y \to X$ extends to a map g from an open neighborhood A of Y to X. By compactness of $[0,1]$, there is an open neighborhood W of p in X with $CW \subseteq A$. Let h be the restriction of g to CW; then h gives a strong contraction of W to p within X.

Let U be an open neighborhood of p in X; we may assume $U \subseteq W$ without loss of generality. Then $B = h^{-1}(U)$ is an open neighborhood of $(U \times \{0\}) \cup (\{p\} \times [0,1])$ in CW, so there is an open neighborhood V of p in X with $CV \subseteq B$. Then $h|_{CV}$ gives a strong contraction of V to p within U. 👍

XIII.26.3.10. Corollary. The path components of an ANR are clopen sets. In particular, an ANR has only finitely many connected components.

Thus, for example, an infinite totally disconnected compactum (e.g. the Cantor set, or even the one-point compactification of $\mathbb{N}$) is not an ANR. This can easily be seen directly.

The Homotopy Extension Theorem

One of the most important facts about ANRs is that homotopies can be extended:

XIII.26.3.11. Theorem. [Homotopy Extension Theorem] Let X be an ANR, Z a compactum, Y a closed subspace of Z. Let f_0, f_1 be maps from Y to X, and $h : Y \times [0,1] \to X$ a homotopy from f_0 to f_1. If f_0 extends to a map $\bar{f}_0$ from Z to X, then the entire homotopy h extends to a homotopy $\bar{h} : Z \times [0,1] \to X$. In particular, f_1 extends to a function $\bar{f}_1 : Z \to X$ which is homotopic to $\bar{f}_0$.

For the proof, we need a lemma:

XIII.26.3.12. Lemma. Let Z be a normal space, and Y a closed subspace of Z. For every open neighborhood V of $A = (Z \times \{0\}) \cup (Y \times [0,1])$ in $Z \times [0,1]$ there is a map $\phi : Z \times [0,1] \to V$ which is the identity on A.

Proof: By compactness of $[0,1]$ there is an open neighborhood U of Y in Z such that $U \times [0,1] \subseteq V$. By Urysohn's Lemma there is a map $g : Z \to [0,1]$ which is 1 on Y and 0 on $Z \setminus U$. Set

$$\phi(z,t) = (z, g(z)t)$$

for $z \in Z$, $t \in [0,1]$. Then ϕ is the identity on A and the range of ϕ is contained in $(Z \times \{0\}) \cup (U \times [0,1]) \subseteq V$. 👍

We now give the proof of Theorem XIII.26.3.11..

Proof: Let $A = (Z \times \{0\}) \cup (Y \times [0,1])$, and define $g : A \to X$ by

$$g((z,t)) = \begin{cases} \bar{f}_0(z) & \text{if} \quad t = 0 \\ h(z,t) & \text{if} \quad z \in Y \end{cases} .$$

Then g extends to an open neighborhood V of A in $Z \times [0,1]$. Since Z is normal, there is a map $\phi : Z \times [0,1] \to V$ which is the identity on A. Set

$$\bar{h}(z,t) = g(\phi(z,t)$$

for $z \in Z$, $t \in [0,1]$ is the desired extension. 👍

If X is an AR, the proof simplifies greatly and Lemma XIII.26.3.12. is not needed, since we can take V to be all of $Z \times [0,1]$.

The Homotopy Extension Theorem has many applications, such as:

XIII.26.3.13. COROLLARY. An ANR is an AR if and only if it is contractible.

PROOF: An AR is a contractible ANR. Conversely, suppose X is a contractible ANR, embedded in a compactum Z. If h is a contraction of X, written as a homotopy from a constant function $f_0 : X \to \{p\} \subseteq X$ to $f_1 = \iota_x$, then f_0 extends to a (constant) function from Z to X, and hence ι_X extends to a map from Z to X, which is a retraction. Thus X is an AR. 👍

XIII.26.3.14. In fact, a nearly identical proof shows that the homotopy extension property characterizes ARs among all contractible compacta. However, it does not characterize ANRs: a zero-dimensional compactum like the Cantor set trivially has the homotopy extension property. It is much more difficult to say whether it characterizes ANRs among locally path-connected or locally contractible compacta.

Homotopy of Close Maps

Another property related to the Homotopy Extension Theorem is that "sufficiently close" maps from a space into an ANR are homotopic, even homotopic by maps that remain "close."

XIII.26.3.15. DEFINITION. Let X be a topological space and $\mathcal{U}$ an open cover of X. Two functions f and g from a set Z into X are *$\mathcal{U}$-close* if for every $z \in Z$, there is a $U \in \mathcal{U}$ with $f(z)$ and $g(z)$ both in $\mathcal{U}$. If Z is a topological space, a homotopy (f_t) from f_0 to f_1 is a *$\mathcal{U}$-homotopy* if for every $z \in Z$ there is a $U \in \mathcal{U}$ with $f_t(z) \in \mathcal{U}$ for all $t \in [0,1]$. If (X, ρ) is a metric space, two functions f and g are *ϵ-close* ($\epsilon > 0$) if $\rho(f(z), g(z)) < \epsilon$ for all $z \in Z$. An ϵ-homotopy from f_0 to f_1 is a homotopy for which $\rho(f_t(z), f_s(z)) < \epsilon$ for all $z \in Z$, $s, t \in [0,1]$.

If (X, ρ) is a metric space and $\epsilon > 0$, ϵ-close functions are $\mathcal{U}$-close and ϵ-homotopies are $\mathcal{U}$-homotopies, where $\mathcal{U}$ is the covering of X by open balls of radius $\frac{\epsilon}{2}$.

XIII.26.3.16. THEOREM. Let X be an ANR. Then for every open cover $\mathcal{U}$ of X there is a refinement $\mathcal{V}$ such that, whenever Z is a topological space and f and g are $\mathcal{V}$-close maps from Z to X, then f and g are $\mathcal{U}$-homotopic. If X is metrizable by a metric ρ, then for any $\epsilon > 0$ there is a $\delta > 0$ such that if Z is a topological space and f and g are δ-close maps from Z to X, then f and g are ϵ-homotopic.

Note that $\mathcal{V}$ depends on $\mathcal{U}$ but not on Z, f, g, and in the metrizable case δ depends on ρ and ϵ but not on Z, f, g.

PROOF: Embed X in a cube K. There is a retraction r from an open neighborhood W of X to X. Let $\tilde{\mathcal{U}} = \{r^{-1}(U) : U \in \mathcal{U}\}$; then $\tilde{\mathcal{U}}$ is an open cover of W. Since K has a convex structure and a base for its topology consisting of convex open sets, there is a refinement $\tilde{\mathcal{V}}$ of $\tilde{\mathcal{U}}$ consisting of convex open sets. Set $\mathcal{V} = \{\tilde{V} \cap X : \tilde{V} \in \tilde{\mathcal{V}}\}$; then $\mathcal{V}$ is an open cover of X refining $\mathcal{U}$.

Let Z be a topological space and f, g $\mathcal{V}$-close maps from Z to X. If f and g are regarded as maps into W, they are $\tilde{\mathcal{V}}$-close. Since the sets in $\tilde{\mathcal{V}}$ are convex, we may define, for $z \in Z$ and $t \in [0,1]$,

$$\tilde{f}_t(z) = \tilde{h}(z,t) = tg(z) + (1-t)f(z) \ .$$

Then $\tilde{h}$ is a $\mathcal{V}$-homotopy from $\tilde{f}_0 = f$ to $\tilde{f}_1 = g$ as maps from Z to W. Set $h = r \circ \tilde{h}$, i.e. $f_t = r \circ \tilde{f}_t$; then h is a $\mathcal{U}$-homotopy from f to g as functions from Z to X.

For the metrizable case, see Exercise XIII.26.6.8.. 👍

This argument is a nice application of convexity in topology.

Products of ANRs

XIII.26.3.17. An arbitrary product of ARs is an AR. It is easily seen that a finite product of ANRs is an ANR, but an arbitrary product of ANRs is not an ANR in general: for example, neither the Cantor set (an infinite product of two-point spaces) nor an infinite product of circles is an ANR (for one thing, they are not locally contractible). In fact, we have

XIII.26.3.18. PROPOSITION. If $\{X_i : i \in I\}$ is a set of compacta, then $\prod_i X_i$ is an ANR if and only if each X_i is an ANR and all but finitely many are ARs.

Metrizable ANRs

We have, in analogy with ():

XIII.26.3.19. PROPOSITION. Let X be a metrizable compactum. The following are equivalent:
(i) X is an ANR.
(ii) X is (homeomorphic to a) retract of an open set in the Hilbert cube.
(iii) Whenever X is embedded in a metrizable compactum Z, there is a retraction from an open neighborhood U of X in Z to X.

One unsatisfying thing about the definition of an ANR (even the variation in XIII.26.3.4.) is that, although it deals with compacta, it involves open sets (neighborhoods) in compacta and thus is not phrased entirely in terms of compacta. There is an alternate form of the definition in the metrizable case, which not only remedies this but also has a straightforward translation to a useful condition (semiprojectivity) for noncommutative C*-algebras [**?**]:

XIII.26.3.20. PROPOSITION. Let X be a metrizable compactum. The following are equivalent:
(i) X is an ANR.
(ii) Whenever X is embedded in a metrizable compactum Z, and (Y_n) is a decreasing sequence of closed sets in Z with $\cap_n Y_n = X$, there is a retraction from Y_n to X for some n (hence for all sufficiently large n).
(iii) Whenever Z is a metrizable compactum and (Y_n) is a decreasing sequence of closed sets in Z with $Y = \cap_n Y_n$, and f is a map from Y to X, there is an extension of f to a map from Y_n to X for some n (hence for all sufficiently large n).

PROOF: (i) $\Rightarrow$ (iii): Extend f to an open set U containing Y in Z. The open sets $(Z \setminus Y_n)$ increase and cover the compact set $Z \setminus U$, so $Z \setminus U \subseteq X \setminus Y_n$ and thus $Y_n \subseteq U$ for some (all sufficiently large) n.
(ii) $\Rightarrow$ (i): Fix a metric ρ on Z. If $X \subseteq Z$, for each n let $U_n = \left\{z \in Z : \rho(z, X) < \frac{1}{n}\right\}$ and $Y_n = \bar{U}_n$. Then $\cap_n Y_n = X$. Extend ι_X to Y_n for some n; the extension gives a retraction of U_n to X.
(iii) $\Rightarrow$ (ii) is trivial. 👍

Assembly

There is an analog of XIII.26.2.20.. The proof is similar but slightly more complicated; see e.g. [Bor67,].

XIII.26.3.21. THEOREM. Let X be a metrizable compactum, X_1 and X_2 closed subspaces with $X_0 = X_1 \cap X_2$.
(i) If X_0, X_1, X_2 are ANRs, then X is an ANR.
(ii) If X and X_0 are ANRs, then X_1 and X_2 are ANRs.

Similarly, XIII.26.2.23. has a direct analog:

XIII.26.3.22. THEOREM. Let X_1 and X_2 be metrizable ANRs, and X_0 a subset of X_1 which is also an ANR. Let $f : X_0 \to X_2$ be a continuous function. Then $X_1 \cup_f X_2$ () is an ANR.

This has an important consequence:

XIII.26.3.23. COROLLARY. Every finite CW-complex (in particular, any polyhedron) is an ANR.

XIII.26.3.24. There are many important results about approximation of ANRs by polytopes which can be found in the standard references.

A Characterization of Metrizable ANRs

ANRs are locally strongly contractible. Conversely:

XIII.26.3.25. THEOREM. Let X be a finite-dimensional metrizable compactum. Then X is an ANR if and only if it is locally contractible.

See e.g. [Bor67,] for a proof.

XIII.26.3.26. Finite-dimensionality cannot be eliminated here: there is an example, due to Borsuk, of an infinite-dimensional metrizable compactum which is locally contractible but not an ANR (see e.g. [Bor67,]). However, the following seems to be a reasonable conjecture:

XIII.26.3.27. **Conjecture:** A metrizable compactum is an ANR if and only if it is locally strongly contractible.

XIII.26.3.28. It is surprising that this conjecture does not seem to have appeared in the literature. No counterexamples appear to be known. In fact, topologists do not seem to have appreciated the notion of local strong contractibility, which is not generally mentioned in the literature; when XIII.26.3.25. is discussed, only Borsuk's counterexample in the infinite-dimensional case is mentioned. Indeed, the reason Borsuk's example is not an ANR is essentially precisely because it is *not* locally strongly contractible. Actually, one interesting consequence of XIII.26.3.25. is that local contractibility and local strong contractibility coincide for finite-dimensional metrizable compacta. See XIII.26.6.5., [Bor69], and [vM89, 5.1.2] for closely related conjectures.

A simple consequence is:

XIII.26.3.29. Corollary. Any compact topological manifold, with or without boundary, is an ANR.

Proof: If X is a compact topological manifold and $p \in X$, then there is a local base for the topology at p consisting of open sets strongly contractible to p. 👍

The Countable Chain Condition

Although an ANR, even an AR, need not be metrizable or even separable (a large enough cube is not), we have:

XIII.26.3.30. Proposition. Every ANR is measurably separable (XIII.24.1.7.), and hence has the Countable Chain Condition (XIII.24.1.9.).

Proof: Every cube is MS (XIII.24.1.12.); hence any open set in a cube is MS (XIII.24.1.11.), and any continuous image of an open set in a cube is MS (XIII.24.1.10.). 👍

This result requires some version of the AC.

XIII.26.4. The Noncompact Case

The theory of ARs and ANRs can be extended to more general classes of spaces, with some restrictions and technical complications. This increased generality not only has some direct applications, but is even relevant to the structure of compact ARs and ANRs. We will only survey this extended theory, with some relevant proofs.

Actually, from a purely logical standpoint we have probably done things backwards: the standard references do the general or metrizable case first and then go on to the compact or compact metrizable case. But some things, notably the equivalence of retracts and extensors, are much easier in the compact case, and the compact case is the primary one of interest anyway, so we chose to do it first.

XIII.26.4.1. We will work only with Hausdorff spaces, so "space" will mean "Hausdorff space." We consider various classes $\mathcal{C}$ of spaces which are *weakly hereditary*: if a space X is in $\mathcal{C}$, so is every closed subspace (in particular, every singleton is in $\mathcal{C}$). Some important examples are:

$\mathcal{N}$, the class of all normal spaces.

$\mathcal{M}$, the class of all metrizable spaces.

$\mathcal{K}$, the class of all compacta.

$\mathcal{KM}$, the class of all metrizable compacta.

There are many other weakly hereditary classes such as the class of perfectly normal spaces or locally compact metrizable spaces. It turns out that only classes of normal spaces give anything interesting in this theory (XIII.26.4.4.).

XIII.26.4.2. Definition. Let $\mathcal{C}$ be a weakly hereditary class of spaces.

(i) A space X is an *AR($\mathcal{C}$)* if, whenever X is embedded as a closed subspace of a $Z \in \mathcal{C}$, there is a retraction from Z to X.

(ii) A space X is an *AE($\mathcal{C}$)* if, whenever $Z \in \mathcal{C}$ and Y is a closed subspace of Z, any map from Y to X extends to a map from Z to X.

(iii) A space X is an *ANR($\mathcal{C}$)* if, whenever X is embedded as a closed subspace of a $Z \in \mathcal{C}$, there is a retraction from an open neighborhood of X in Z to X.

(iv) A space X is an *ANE($\mathcal{C}$)* if, whenever $Z \in \mathcal{C}$ and Y is a closed subspace of Z, any map from Y to X extends to a map from an open neighborhood of Y in Z to X.

XIII.26.4.3. An AR($\mathcal{C}$) is an ANR($\mathcal{C}$), and an AE($\mathcal{C}$) is an ANE($\mathcal{C}$). A space which is an ANR($\mathcal{C}$) must be in $\mathcal{C}$ since $\mathcal{C}$ is weakly hereditary; but by this definition an AE($\mathcal{C}$) or ANE($\mathcal{C}$) is not required to be in $\mathcal{C}$. An AE($\mathcal{C}$) [ANE($\mathcal{C}$)] which is in $\mathcal{C}$ is an AR($\mathcal{C}$) [ANR($\mathcal{C}$)] by the standard argument. The converse is trickier than in the compact case. An AE($\mathcal{C}$) [ANE($\mathcal{C}$)] is also an AE($\mathcal{D}$) [ANE($\mathcal{D}$)] for any subclass $\mathcal{D}$ of $\mathcal{C}$; the same is true for an AR($\mathcal{C}$) [ANR($\mathcal{C}$)] if the space is in $\mathcal{D}$.

An AR in the sense of () is precisely an AR($\mathcal{K}$), and an ANR in the sense of () is precisely an ANR($\mathcal{K}$). A metrizable AR [ANR] is precisely an AR($\mathcal{KM}$) [ANR($\mathcal{KM}$)]. The terms AR and ANR with no class qualification will continue to have their previous meanings.

If $\mathcal{C}$ is too large, there are no nontrivial ANE($\mathcal{C}$)s:

XIII.26.4.4. Proposition. [**?**, II.2.3] If $\mathcal{C}$ is a weakly hereditary class containing a nonnormal space, then no space with more than one point is an ANE($\mathcal{C}$).

Proof: Let X be a (Hausdorff) space and p and q distinct points of X. Let V and W be disjoint open neighborhoods of p and q. Let Z be a nonnormal space in $\mathcal{C}$, and A and B disjoint closed sets in Z which do not have disjoint open neighborhoods. Let $Y = A \cup B$, and define $f : Y \to X$ by $f(y) = p$ if $y \in A$ and $f(y) = q$ if $y \in B$. Suppose f extends to a map $\bar{f} : U \to X$ for some open neighborhood U of Y. Then $\bar{f}^{-1}(V)$ and $\bar{f}^{-1}(W)$ are disjoint open neighborhoods of A and B, a contradiction. 👍

Thus from now on we will consider only normal spaces.

Embedding in Convex Sets

In the compact case, the proof that an AR [ANR] is an AE [ANE] is a simple consequence of the Tietze Extension Theorem and the fact that a compactum embeds as a closed subset of a cube. A different argument is needed in the noncompact case: the Tietze Extension Theorem is replaced by the Dugundji Extension Theorem (), and the embedding in a cube must be replaced by an embedding as a closed subset of a convex set in a locally convex topological vector space.

XIII.26.4.5. THEOREM. [DUGUNDJI EXTENSION THEOREM] Let V be a locally convex topological vector space, Z a metrizable space, Y a closed subspace of Z. If f is a map from Y to V, then f extends to a map $\bar{f}$ from Z to V with $\bar{f}(Z)$ contained in the convex hull of $f(X)$.

The proof is complicated and beyond the scope of this book, and apparently only works for metrizable Z. See the standard references.

XIII.26.4.6. COROLLARY. Any convex set in a locally convex topological vector space is an AE($\mathcal{M}$).

Note that such a convex set need not itself be metrizable, hence need not be an AR($\mathcal{M}$); but it is an AR($\mathcal{M}$) if it is metrizable.

The metrizable version of the Tietze Extension is the special case of XIII.26.4.6. for a convex set (interval) in $\mathbb{R}$. But the Tietze Extension Theorem also applies in the (normal) metrizable case, i.e. an interval in $\mathbb{R}$ is an AE($\mathcal{N}$).

The proof of the next result is simple but uses some heavy functional analysis machinery.

XIII.26.4.7. PROPOSITION. Let X be a completely regular space. Then there is a precompact convex set Z in a locally convex topological vector space whose set of extreme points is closed in Z and homeomorphic to X. If X is second countable, Z can be chosen metrizable. (Conversely, the completely regular condition is necessary since any subset of a Hausdorff topological group is completely regular.)

PROOF: X embeds in a cube K. The Bauer simplex Δ_K () is a weak-* compact convex subset of $C(K)^*$, where $C(K)$ is the Banach space of real-valued continuous functions on K, and the set of extreme points of Δ_K is closed and homeomorphic to K. Let Z be the convex hull of X in Δ_K. We have $K \cap Z = X$ since every point of K is an extreme point of Δ_K; thus X is closed in Z, and the points of X are extreme points (all the extreme points) of Z. If X is second countable, K can be chosen to be the Hilbert cube. 👍

XIII.26.4.8. This result is a variation of a theorem of KURATOWSKI and WOJDYSŁAWSKI, which gives a different embedding of a metrizable space as a closed subset of a convex set in a Banach space. In our construction the convex set Z is a "noncompact Bauer simplex": every point of Z is uniquely a convex combination of points of X. (The closure of Z in Δ_K is isomorphic to $\Delta_{\bar{X}}$.)

With these two results, the proof of the following result is essentially identical to the proofs of () and ():

XIII.26.4.9. THEOREM. An AR($\mathcal{M}$) [ANR($\mathcal{M}$)] is an AE($\mathcal{M}$) [ANE($\mathcal{M}$).

This result can be extended somewhat to nonmetrizable spaces; cf. [**?**, II.16.1].

Retracts of ANRs

Proposition () has a considerable extension in the noncompact case, with an identical proof:

XIII.26.4.10. PROPOSITION. Let $\mathcal{C}$ be a weakly hereditary class. Then any retract of an open set (in particular, any open set) in an ANE($\mathcal{C}$) is an ANE($\mathcal{C}$).

XIII.26.4.11. COROLLARY. Any retract of an open set (in particular, any open set) in an ANR($\mathcal{M}$) is an ANR($\mathcal{M}$).

The Homotopy Extension Theorem

The Homotopy Extension Theorem (XIII.26.3.11.) works identically with the same proof in the general case of ANEs, since the proof only uses that Z is normal:

XIII.26.4.12. THEOREM. [GENERAL HOMOTOPY EXTENSION THEOREM] Let $\mathcal{C}$ be a weakly hereditary class of normal spaces, X be an ANE($\mathcal{C}$), $Z \in \mathcal{C}$, Y a closed subspace of Z. Let f_0, f_1 be maps from Y to X, and $h : Y \times [0,1] \to X$ a homotopy from f_0 to f_1. If f_0 extends to a map $\bar{f}_0$ from Z to X, then the entire homotopy h extends to a homotopy $\bar{h} : Z \times [0,1] \to X$. In particular, f_1 extends to a function $\bar{f}_1 : Z \to X$ which is homotopic to $\bar{f}_0$. Taking $\mathcal{C} = \mathcal{M}$, ANE($\mathcal{C}$) can be replaced by ANR($\mathcal{M}$).

XIII.26.4.13. COROLLARY. Let $\mathcal{C}$ be a weakly hereditary class of normal spaces. An ANE($\mathcal{C}$) is an AE($\mathcal{C}$) if and only if it is contractible. An ANR($\mathcal{M}$) is an AR($\mathcal{M}$) if and only if it is contractible.

The Close Homotopy Theorem XIII.26.3.16. has a version for ANR($\mathcal{M}$)s. The proof is identical except that XIII.26.4.7. is used in place of embedding in a cube.

XIII.26.4.14. THEOREM. Let X be an ANR($\mathcal{M}$). Then for every open cover $\mathcal{U}$ of X there is a refinement $\mathcal{V}$ such that, whenever Z is a topological space and f and g are $\mathcal{V}$-close maps from Z to X, then f and g are $\mathcal{U}$-homotopic.

XIII.26.4.15. There appears to be a possibility of extending this result to nonmetrizable $\mathcal{C}$, since only XIII.26.4.7., which holds for arbitrary $\mathcal{C}$, and not the Dugundji Extension Theorem, is used in the proof. However, there is a subtlety: for the proof to work, X must be embeddable as a closed subset of a convex set *which is in* $\mathcal{C}$. It is unclear what classes besides $\mathcal{K}$, $\mathcal{M}$, and $\mathcal{KM}$ have this property. And even in the metrizable case, the $\epsilon - \delta$ version does not follow and is likely false in general, since an open cover of a noncompact metric space does not generally have a Lebesgue number.

Applications to the Compact Case

XIII.26.4.16. Proposition. Every AR [ANR] is an AE($\mathcal{N}$) [ANE($\mathcal{N}$)], and hence an AR($\mathcal{N}$) [ANR($\mathcal{N}$), A metrizable AR [ANR] is an AR($\mathcal{M}$) [ANR($\mathcal{M}$)].

This is a simple consequence of XIII.26.4.10. and the fact that any cube is an AE($\mathcal{N}$) by the Tietze Extension Theorem.

There is a local characterization of ANR($\mathcal{M}$)s:

XIII.26.4.17. Theorem. Let X be a metrizable space. Then X is an ANR($\mathcal{M}$) if (and only if) every point of X has an open neighborhood which is an ANR($\mathcal{M}$).

XIII.26.4.18. Corollary. Let X be a metrizable compactum. If every point of X has a compact neighborhood which is an ANR, then X is an ANR.

XIII.26.4.19. It is a consequence of XIII.26.4.11. that any ANR($\mathcal{M}$) (e.g. metrizable ANR) can be covered by open ANR($\mathcal{M}$)s of arbitrarily small diameter, i.e. there is a base for the topology consisting of ANR($\mathcal{M}$)s.

It might be asked whether a metrizable ANR can be covered by open sets of arbitrarily small diameter whose closures are ANRs, or even ARs. However, this is false in general: An example is given in [Bor67, VI.4] of a two-dimensional metrizable AR X such that every proper sub-AR of X is nowhere dense; thus X cannot be written as a countable union of proper sub-ARs, and hence there are points of x with no local base of ARs.

XIII.26.5. Affine Structure in ARs and ANRs

Whenever an AR X is embedded in a compact convex set K, the affine structure on K can be transferred to X via a retraction. This can also be done locally for an ANR. There is a particularly nice way to do this which is "almost canonical."

XIII.26.5.1. If X is a compactum, we can form the Bauer simplex Δ_X (), which is a compact convex set in the dual of the Banach space $C(X)$ of real-valued continuous functions on X. The elements of Δ_X can (at least if X is metrizable) be identified with the Borel probability measures on X, and the extreme points are the point masses at points of X, so the set of extreme points is closed and can be identified topologically with X.

A retraction from Δ_X to X transfers the affine structure of Δ_X to X. One gets not only a notion of convex combinations of points of X, but even a well-defined barycenter for any Borel probability measure on X. The affine structure on Δ_X does not depend on how X is identified with the extreme points of Δ_X, since a homeomorphism of X induces an affine homeomorphism of Δ_X. The induced affine structure on X does depend on what retraction is chosen, however, and is not invariant under homeomorphisms of X in general.

XIII.26.5.2. If X is an ANR, there is only a retraction of an open neighborhood U of X in Δ_X to X. If $p \in X$, there is a compact neighborhood A_p of p such that the closed convex hull of A_p is contained in U (); this closed convex hull is isomorphic to Δ_{A_p} (). Thus the affine structure of Δ_{A_p} can be transferred to X, so in a sense X has a "local affine structure." These are in a way even nicer than local coordinate systems on a manifold: when two of them overlap they agree on the overlap.

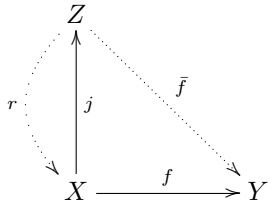
However, the local affine structure obtained from A_p is not an affine structure on A_p in general (e.g. a point of an ANR need not have a contractible neighborhood, cf. ()): a convex combination of points of A_p may lie outside p, although will lie in a "small" neighborhood of A_p in X (e.g. if V is an open neighborhood of p in X, A_p can be chosen so that convex combinations of points in A_p, or barycenters of measures, are always in V).

XIII.26.5.3. It is unclear whether these affine structures are well-behaved enough to be useful, but they are intriguing.

XIII.26.6. Exercises

XIII.26.6.1. Let X be an infinite set. Define $r : X \times \{0,1\} \to X \times \{0,1\}$ by $r((x,n)) = (x,0)$ for $x \in X$, $n \in \{0,1\}$. Show that if $X \times \{0,1\}$ has the finite complement topology (), then r is a retraction whose range is not closed.

XIII.26.6.2. Here is a variation on XIII.26.2.6.. Let X be an AR, embedded in a compactum Z. If Y is any topological space and $f : X \to Y$ a map, then f extends to a map $\bar{f} : Z \to Y$. [Let r be a retraction from Z to X, and set $\bar{f} = f \circ r$.]



XIII.26.6.3. Prove the following generalization of XIII.26.2.10.:

THEOREM. Let X be an AR. If X is embedded as a subspace of another AR Z, there is a strong deformation retraction of Z onto X.

(a) Let r be a retraction from Z to X, and let W be the mapping cylinder (XIII.9.5.10.) of r. Let Y be the subspace $(Z \times \{0\}) \cup (X \times [0,1])$ of W. There is an obvious function $f : Y \to Z$.

(b) An extension of f to W gives a strong deformation retraction of Z to X.

(c) Use the theorem to show that a compactum is an AR if and only if it is a strong deformation retract of a cube.

(d) Obtain XIII.26.2.10. as a special case of this theorem.

(e) Similarly show:

THEOREM. Let Z be an AR, and X a (necessarily closed) subset of Z. Any retraction from Z to X is strongly homotopic to the identity map ι_Z (i.e. homotopic with the points of X pointwise fixed).

XIII.26.6.4. Let Z be a finite-dimensional metrizable AR, and (Z_n) a sequence of pairwise disjoint sub-ARs of Z.

(a) Consider the decomposition $\sim$ of Z into the Z_n and singletons. Show that if the diameters of the Z_n (in some, hence any, metric on Z) go to zero, then the decomposition is usc.

(b) Let $X = Z/\sim$ be the quotient space. Show (cf. XIII.9.5.11.) that $dim(X) \leq dim(Z)$. Conclude from XIII.26.2.27. that X is an AR. X is "Z with all the Y_n collapsed to points."

(c) More generally, let f_n be a map of Z_n onto an AR Y_n for each n. Decompose Z into the preimages of points of the Y_n and singletons (show that this is usc), and let X be the quotient space. Show that X is an AR. [This does not follow from XIII.26.2.27., but an argument based on XIII.26.2.23. can be used.] By () we have $dim(X) \leq \max(dim(Z), \sup_n dim(Y_n))$.

The construction can also be done if the f_n are not surjective, but is more delicate since then the X is not a quotient of Z; it is tricky to even give the proper definition of X.

XIII.26.6.5. (a) Show that XIII.26.2.16. is equivalent to:

PROPOSITION. Let X be an AR. Then there is a horizontal strong deformation retraction of $X \times X$ onto the diagonal ["horizontal" means horizontal slices remain globally invariant].

(b) Similarly show that if κ is any cardinal, then there is a horizontal strong deformation retraction of X^κ to the diagonal.

(c) Does the property in (a) or (b) characterize ARs among all compacta? All metrizable compacta? (See [Bor69] for related results.)

(d) Discuss similar results and problems for ANRs.

XIII.26.6.6. Here is a nice application of XIII.26.6.4.(c) (see [Bor67,] for a nearly identical example). If $k \geq 3$, say a *k-star* consists of k copies of $[0,1]$ joined at 0. Fix $d \geq 2$, and let Z be the closed d-ball B^d or the d-cube $[0,1]^d$. Let S be an infinite subset of $\mathbb{N}$, and $\boldsymbol{k} = (k_1, k_2, \dots)$ a sequence listing the elements of S, with each $k \in S$ occurring infinitely often. Let (Z_n) be a sequence of pairwise disjoint subsets of the interior of Z whose diameters go to 0, where Z_n is a k_n-star, with $\cup_n Z_n$ dense in Z.

(a) Show that a k-star is an AR for any k.

(b) Let f_n be the obvious map of Z_n onto $[0,1]$ identifying together all the spikes. Use XIII.26.6.4.(c) to show that the quotient space $X_{\boldsymbol{k}}$ is an AR. (The notation is a little ambiguous, since $X_{\boldsymbol{k}}$ depends not only on $\boldsymbol{k}$ but also on how the Y_n are chosen.) $X_{\boldsymbol{k}}$ consists of Z with each star "puckered together."

(c) If S and S' are infinite subsets of $\mathbb{N}$ with $S \triangle S'$ infinite, and $\boldsymbol{k}$ and $\boldsymbol{k}'$ corresponding sequences, show that no open set in $X_{\boldsymbol{k}}$ can be homeomorphic to an open subset of $X_{\boldsymbol{k}'}$.

(d) Show that $dim(X_{\boldsymbol{k}}) = d$ for any $\boldsymbol{k}$.

(e) Conclude that there are $2^{\aleph_0}$ pairwise incomparable metrizable ARs of dimension d.

(f) Fix $\boldsymbol{k}$. Let C be the boundary sphere of Z. Then C embeds naturally in $X_{\boldsymbol{k}}$ since all the Y_n are disjoint from C. A space like $X_{\boldsymbol{k}}$ is described in [Bor67] as "a $(d-1)$-sphere with a membrane attached." Since $X_{\boldsymbol{k}}$ is contractible, C can be contracted to a point in $X_{\boldsymbol{k}}$. Show that C cannot be contracted to a point in any proper subcontinuum of $X_{\boldsymbol{k}}$ containing it.

(g) If $\boldsymbol{k}$ has no repeated terms, the identical construction gives a two-dimensional AR $X_{\boldsymbol{k}}$ which is totally inhomogeneous. There are $2^{\aleph_0}$ pairwise nonhomeomorphic such $X_{\boldsymbol{k}}$ (although they are not all pairwise incomparable). Compare with XIII.19.6.11..

XIII.26.6.7. [**?**,] Let $T = [0,1]^\kappa$, where κ is uncountable. Denote by $\mathbf{0}$ the point of T with all coordinates 0. Let $Z = [0,1] \times T$; then $Z \cong T$ is an AR. Set $X_1 = [0,1] \times \{\mathbf{0}\}$ and $X_2 = \{0\} \times T$. Then $X_1 \cong [0,1]$ and $X_2 \cong T$ are ARs, and $X_1 \cap X_2 = \{(0,\mathbf{0})\}$ is also an AR. If $X = X_1 \cup X_2$, show that there is no retraction of Z onto X, so X is not an AR; in fact there is no retraction of any neighborhood of X in Z to X, so X is not an ANR. [$X_1 \setminus \{(0,\mathbf{0})\}$ and $X_2 \setminus \{(0,\mathbf{0})\}$ do not have disjoint neighborhoods in $Z \setminus \{(0,\mathbf{0})\}$.] Contrast with XIII.26.2.20.(i).

XIII.26.6.8. Let X be a metrizable ANR. Prove the metric statement in XIII.26.3.16., as follows:

(a) Show that if the statement is true for one metric, it is true (with possibly different δ) for any other metric. [By compactness, any two metrics are uniformly equivalent.]

(b) Let $\epsilon > 0$. Let $\mathcal{U}$ be the covering of X by balls of radius $\frac{\epsilon}{2}$. Define $\tilde{\mathcal{U}}$, $\tilde{\mathcal{V}}$, $\mathcal{V}$ as in the proof.

(c) Let $\delta > 0$ be a Lebesgue number for the cover $\mathcal{V}$. Show that this δ works.

XIII.26.6.9. Let X be a compact subset of $\mathbb{R}^n$. If there is a retraction of X onto ∂X, show that $X = \partial X$, i.e. the interior X° of X is empty. [Embed X in a closed ball B, and let S be the boundary sphere of B. Show there is a retraction of B onto $B \setminus X^\circ$. If $X^\circ \neq \emptyset$, compose this retraction with a retraction of $B \setminus X^\circ$ to S, contradicting the Brouwer Fixed-Point Theorem.] In particular, if U is a bounded open set in $\mathbb{R}^n$, there is no retraction of $\bar{U}$ onto ∂U.

XIII.26.6.10. Let X be a subset of $\mathbb{R}^n$ which is an ANR. Show that $\mathbb{R}^n \setminus X$ has only finitely many components, as follows.

(a) Let U be a bounded open neighborhood of X in $\mathbb{R}^n$ with a retraction r from U to X. Show that every component C of $\mathbb{R}^n \setminus X$ meets $\mathbb{R}^n \setminus U$. [If C is a component of $\mathbb{R}^n \setminus X$ contained in U, then $\partial C \subseteq X$ and r defines a retraction of $\bar{C}$ to ∂C, contradicting XIII.26.6.9..]

(b) Let B be a closed ball containing U, and $\mathcal{C} = \{C_i : i \in I\}$ the set of all components of $\mathbb{R}^n \setminus X$. Show that $\mathcal{C}$ is an open cover of $B \setminus U$. So $\mathcal{C}$ has a finite subcover. Since the sets in $\mathcal{C}$ are pairwise disjoint, conclude that I is finite.

(c) If X is an AR and $n > 1$, show that $\mathbb{R}^n \setminus X$ is connected.

XIII.26.6.11. Prove or disprove: if X is an AR, and (X_n) a decreasing sequence of sub-ARs of X, then $\cap_n X_n$ is also an AR. [Consider the pseudo-arc.]

The corresponding statement for ANRs is false: the Cantor set is a counterexample.

XIII.26.6.12. Is every compact convex set in a locally convex topological vector space homeomorphic to a cube? Is such a set at least an AR? This is true in the metrizable case by [Kle55].

Chapter XIV

Differential Geometry and Topology

In this chapter, we describe the basics of the theory of smooth manifolds and smooth maps between them, although it is beyond the scope of this book to give anything close to a complete treatment of this entire large branch of mathematics.

XIV.1. Smooth Manifolds

Smooth manifolds are a generalization of smooth curves and surfaces. They are topological manifolds in the sense of XIII.20.1.7. with extra structure allowing notions of calculus and geometry to be sensibly used. They are the basic objects of study in differential geometry and topology.

There are at least two reasonable approaches to the concept of smooth manifolds. We can define them to be

(1) Nicely behaved subsets of $\mathbb{R}^n$ (*concrete (smooth) manifolds*)

or

(2) Abstract spaces with a collection of local charts covering the space smoothly compatible on the intersections (*abstract (smooth) manifolds*)

The second approach is apparently more general, although it motivates the definition of "nicely behaved" in the first approach. But the Whitney Embedding Theorem says the two approaches are equivalent: every abstract (smooth) n-manifold is diffeomorphic to a concrete (smooth) manifold, in fact a (smooth) submanifold of $\mathbb{R}^{2n+1}$.

Each approach has its advantages and disadvantages. Concrete manifolds have a conceptual advantage since they are, well, *concrete*: they can generally be pictured as geometric objects. There are also considerable technical advantages in working with concrete manifolds. On the other hand, there are some disadvantages:

(i) Any concrete manifold sits in a specific way in a Euclidean space. Two concrete manifolds can be diffeomorphic () but sit differently in Euclidean space, even in spaces of different dimensions. It can be difficult to separate intrinsic (up to diffeomorphism) properties of a manifold from properties depending on how it is embedded in Euclidean space.

(ii) Some important manifolds (e.g. Lie groups) do not have a natural embedding in Euclidean space, and the various ways they can be embedded can be complicated. For example, it is a large and important industry to describe how Lie groups can be represented as (closed) subgroups of $GL_n(\mathbb{R})$ for various n ().

(iii) Abstract manifolds arising in applications frequently have additional structure, e.g. a Riemannian metric, which play a crucial role. As a simple example, a Lie group has a volume form from Haar measure (). These additional structures do not necessarily come from Euclidean space via an embedding.

For ease of exposition, we will mostly take the first approach, but will describe the second also. Thus for us, "manifold" will generally mean "concrete manifold" unless otherwise specified.

There are various flavors of smooth manifolds. We will mostly consider only $\mathcal{C}^\infty$-manifolds, which are generally the most important; there are also $\mathcal{C}^r$-manifolds for any $r \in \mathbb{N}$, and (real and complex) analytic manifolds. (The study of $\mathcal{C}^r$-manifolds is important in some contexts, but can be largely reduced to the $\mathcal{C}^\infty$ case since every $\mathcal{C}^r$-manifold has an essentially unique compatible structure as a $\mathcal{C}^\infty$-manifold. Analytic manifolds are another story, cf. ().) To avoid unnecessary verbosity, in this chapter "smooth" will always mean $\mathcal{C}^\infty$, and "manifold" will mean "smooth manifold" (i.e. $\mathcal{C}^\infty$-manifold) unless specifically qualified otherwise.

We will change some of our notational conventions from vector calculus. In particular, we will not use boldface to denote points in Euclidean space or functions between Euclidean spaces. Some indexing conventions will also be different.

XIV.1.1. Concrete Manifolds

XIV.1.1.1. Definition. Let $m.p \in \mathbb{N}$, $m \leq p$, $r \in \mathbb{N} \cup \{0\} \cup \{\infty\}$. A *concrete $\mathcal{C}^r$ m-submanifold of $\mathbb{R}^p$* is a nonempty subset M of $\mathbb{R}^p$ with the property that for every $x \in M$, there is an open neighborhood U of x (in $\mathbb{R}^p$) and a $\mathcal{C}^r$-diffeomorphism f from U onto an open ball B around 0 in $\mathbb{R}^p$ such that $f(U \cap M) = B \cap \mathbb{R}^m$, where $\mathbb{R}^m$ is regarded as the subspace of $\mathbb{R}^p$ whose last $p - m$ coordinates are 0. The preimages of the first m coordinate axes under f are a *set of local coordinates* for M around x. Such an f is a *local straightening function* around x, and $f^{-1}|_{\mathbb{R}^m}$ a *local coordinate function* or *local parametrization* around x.
A *(concrete $\mathcal{C}^r$) m-manifold* is a concrete $\mathcal{C}^r$ m-submanifold of $\mathbb{R}^p$ for some $p \geq m$.
(Recall that a $\mathcal{C}^r$ diffeomorphism is a bijection such that both f and f^{-1} are $\mathcal{C}^r$.)

XIV.1.1.2. We will almost exclusively consider the case $r = \infty$, and in this section "manifold" without additional qualification will mean "$\mathcal{C}^\infty$-manifold." Most of what we do generalizes to the $\mathcal{C}^r$ case for $r \geq 1$ ($\mathcal{C}^0$-manifolds are essentially just topological manifolds).

XIV.1.1.3. A $\mathcal{C}^r$ m-manifold M in $\mathbb{R}^p$ has a natural topology, the relative topology from $\mathbb{R}^p$ (we will always use this topology), which is locally Euclidean, and which is obviously metrizable, hence M is a topological m-manifold in the sense of XIII.20.1.7.. Thus all results from that section apply.

XIV.1.1.4. By this definition, the local dimension m of an m-manifold is constant. We could more generally let the local dimension vary, but this is of little benefit and it is cleaner to work with manifolds of constant dimension. (The local dimension would always be at least locally constant, hence constant on connected components; we primarily consider connected manifolds anyway.) It is convenient notation to denote the dimension of M by m.

XIV.1.1.5. The ambient space $\mathbb{R}^p$ is generally unimportant, at least for this section; it is sort of a geometric version of a set-theoretic universe which just facilitates picturing and working with manifolds. However, a concrete submanifold M of $\mathbb{R}^p$ inherits additional structure from the ambient space not part of its manifold structure: for example, m-dimensional Hausdorff measure on $\mathbb{R}^p$ (XVII.6.4.3.) defines a σ-finite measure (m-volume) on M which depends greatly on how M is embedded.

XIV.1.1.6. An m-submanifold M is not generally closed in $\mathbb{R}^p$, although every point of M has a neighborhood (in the relative topology) which is compact and thus closed in $\mathbb{R}^p$. For example, if $m = p$, any m-submanifold of $\mathbb{R}^m$ is an open set by Invariance of Domain () (or by the Inverse Function Theorem); conversely, any nonempty open set in $\mathbb{R}^m$ is an m-submanifold: the local coordinate function f could be just taken to be a translation.

XIV.1.1.7. We can extend the definition to the case $m = 0$: a 0-submanifold (automatically smooth) in $\mathbb{R}^p$ is just a set of points which is discrete in the relative topology. (We could also let $p = 0$, taking $\mathbb{R}^0 = \{0\}$, not a very interesting case.)

XIV.1.1.8. Being an m-submanifold of $\mathbb{R}^p$ is strictly a local property: if $M \subseteq \mathbb{R}^p$ has the property that for each $x \in M$, there is an open neighborhood V of x in $\mathbb{R}^p$ such that $M \cap V$ is an m-submanifold, then M is also an m-submanifold. (Specifying m is necessary here since we are requiring constant local dimension.) The converse is also trivially true; in fact if M is an m-submanifold, then so is $M \cap V$ for every open set V in $\mathbb{R}^p$ for which $M \cap V \neq \emptyset$. (We do not consider $\emptyset$ to be a manifold: what would its dimension be?)

Local Coordinate Functions

XIV.1.1.9. There is, of course, nothing unique about a local straightening function or local coordinates around a point x of a manifold. We could require that $f(x) = 0$, but it is sometimes convenient not to do this. We could also relax or modify the definition of a local straightening function to allow the target set B to be any open set in $\mathbb{R}^p$, since such a function could be restricted to have an open ball as image. We could also require that B be a different kind of open set such as a rectangle centered at 0

$$\{(x_1, \ldots, x_p) : -a_j < x_j < a_j\}$$

for $a_j > 0$, $1 \leq j \leq p$, or an open cube

$$\{(x_1, \ldots, x_p) : -a < x_j < a\}$$

for $a > 0$, since such a set contains an open ball around 0, and conversely. By scaling, we could require the target set to be the open unit ball or open cube with $a = 1$. We could even require the target to be the entire space $\mathbb{R}^p$, although this is rarely useful.

Examples

XIV.1.1.10. As noted above, $\mathbb{R}^m$ itself, or any nonempty open set in $\mathbb{R}^m$, is an m-submanifold of $\mathbb{R}^m$.

XIV.1.1.11. Any m-dimensional affine subspace of $\mathbb{R}^p$ is an m-submanifold. Local straightening functions can be taken to be affine functions.

XIV.1.1.12. Let U be an open set in $\mathbb{R}^m$, and g a smooth function from U to $\mathbb{R}^n$. Then the graph $\Gamma = \{(x, g(x)) : x \in U\}$ of g is an m-submanifold of $\mathbb{R}^m \times \mathbb{R}^n = \mathbb{R}^{m+n}$: a local (even global) straightening function f can be defined by $f(x, y) = (x, y - g(x))$. Thus, for example, S^m is an m-submanifold of $\mathbb{R}^{m+1}$ for any m since it is locally the graph of a smooth function (cf. XIV.1.1.8.). More generally, many subsets of Euclidean space arising as in VIII.5.1.18. are submanifolds by the Implicit Function Theorem.

XIV.1.1.13. If M is an m-submanifold of $\mathbb{R}^p$ and N an n-submanifold of $\mathbb{R}^q$, then $M \times N$ is a $(m+n)$-submanifold of $\mathbb{R}^p \times \mathbb{R}^q = \mathbb{R}^{p+q}$. The product of two local straightening functions is a local straightening function (cf. XIV.1.1.9.).

XIV.1.2. Smooth Maps and Diffeomorphisms

If U is an open set in $\mathbb{R}^q$, there is a well-defined notion of a smooth ($\mathcal{C}^\infty$) map from U to $\mathbb{R}^p$ for any p (). Note that this is strictly a local condition. We extend this to arbitrary subsets of $\mathbb{R}^q$:

XIV.1.2.1. Definition. Let $X \subseteq \mathbb{R}^q$, $Y \subseteq \mathbb{R}^p$, and $f : X \to Y$ be a function. Then f is a *smooth function* if for each $x \in X$, there is an open neighborhood U of x in $\mathbb{R}^q$ and a smooth function $h : U \to \mathbb{R}^p$ such that $h|_{U \cap X} = f|_{U \cap X}$, i.e. f locally extends to a smooth function on a neighborhood of each point. The function f is a *diffeomorphism* if f is a smooth bijection and f^{-1} is also smooth, and is a *local diffeomorphism* if for each $x \in X$ there is a relatively open set V containing x such that $f(V)$ is relatively open in Y and $f|_V$ is a diffeomorphism of V to $f(V)$.

It is obvious (from the corresponding result on open sets, i.e. the Chain Rule) that a composition of smooth maps is smooth.

XIV.1.2.2. Note that a diffeomorphism from X to Y does not necessarily extend even locally to a diffeomorphism of point neighborhoods. In fact, X and Y could be subsets of Euclidean spaces of different dimensions. Even if the dimensions of the ambient spaces are the same, it is not obvious that a diffeomorphism extends locally to diffeomorphisms of neighborhoods (it does for submanifolds, cf.()).

The next result gives an important alternate characterization of manifolds:

XIV.1.2.3. Theorem. Let M be a nonempty subset of $\mathbb{R}^p$. Then M is an m-submanifold of $\mathbb{R}^p$ if and only if M is locally diffeomorphic to $\mathbb{R}^m$, i.e. for every $x \in M$, there is a relatively open neighborhood V of x which is diffeomorphic to $\mathbb{R}^m$ (or to an open ball or open cube in $\mathbb{R}^m$). The preimage of the coordinate axes under such a local diffeomorphism is a set of local coordinates for M around x.

Proof: If M is an m-submanifold of $\mathbb{R}^p$ and $x \in M$, a local straightening function from an open neighborhood U of x in $\mathbb{R}^p$ to an open ball B in $\mathbb{R}^q$ gives a local diffeomorphism of $V = U \cap M$ onto the open ball $B \cap \mathbb{R}^m$. Conversely, suppose for each $x \in M$ there is a neighborhood V of x in M and a diffeomorphism h from V to $\mathbb{R}^m$. Fix x, V, and h. Since h is smooth, there is an open neighborhood U of x in $\mathbb{R}^p$ with $U \cap M \subseteq V$ and a smooth map $g : U \to \mathbb{R}^m$ extending h. The map g must have rank m at all points of a (possibly smaller) neighborhood of X. Thus, by the Domain Straightening Theorem (), there is a neighborhood W of x in $\mathbb{R}^p$ contained in U and a change of coordinates in W such that $g|_W$ consists of projection onto the first m coordinates followed by a diffeomorphism, giving a local straightening function around x. Thus M is an m-submanifold of $\mathbb{R}^p$.

We will often use this alternate characterization.

XIV.1.2.4. Corollary. Let $M \subseteq \mathbb{R}^p$ and $N \subseteq \mathbb{R}^q$. If M and N are diffeomorphic, and M is an m-submanifold of $\mathbb{R}^p$, then N is an m-submanifold of $\mathbb{R}^q$.

XIV.1.2.5. We often regard diffeomorphic manifolds as being the "same," and the main goal of differential topology is to understand manifolds up to diffeomorphism.

Submanifolds

XIV.1.2.6. Definition. Let M be a submanifold of $\mathbb{R}^p$. A *submanifold* of M is a subset of M which is also a submanifold of $\mathbb{R}^p$.

XIV.1.2.7. This definition is consistent with the previous use of "submanifold": a submanifold of $\mathbb{R}^p$ is a submanifold of the submanifold $\mathbb{R}^p$ of $\mathbb{R}^p$. Any open subset of an m-submanifold M of $\mathbb{R}^p$ is a submanifold. See XIV.1.2.8.(ii) for a converse, a version of Invariance of Domain.

XIV.1.2.8. Theorem. Let M be an m-submanifold of $\mathbb{R}^p$, and N a submanifold of M which is an n-submanifold of $\mathbb{R}^p$. Then

(i) $n \leq m$.

(ii) If $n = m$, then N is an open subset of M.

(iii) For any $x \in N$, there is a straightening function f for M around x which is also a straightening function for N around x. Equivalently, there is a set of local coordinates for M around x such that the first n of the coordinates are a set of local coordinates for N around x.

Proof: Let $x_0 \in N$, and let g be a straightening function for M around x_0, i.e. $g : U \to B$ a diffeomorphism with $g(U \cap M) = B \cap \mathbb{R}^m$ and U an open neighborhood of x_0 in $\mathbb{R}^p$. Then $g(U \cap N)$ is a subset of $\mathbb{R}^m$ diffeomorphic to $U \cap N$, which is a manifold; hence $N' = g(U \cap N)$ is a submanifold of $\mathbb{R}^m$ (thus $n \leq m$, proving (i)). Let $y_0 = g(x_0)$, and let h be a straightening function for $N' \subseteq \mathbb{R}^m$ around y_0, i.e. $h : V \to A$ a diffeomorphism with $h(V \cap N') = A \cap \mathbb{R}^n$, A an open ball in $\mathbb{R}^m$, and V an open neighborhood of y_0 in $\mathbb{R}^m$.

If $n = m$, N' fills an entire neighborhood of y_0 in $\mathbb{R}^m$, and hence N fills an entire neighborhood of x_0 in M. Since x_0 is an arbitrary point of N, N is open in M, proving (ii).

Suppose $n < m$. There is an open neighborhood W of y_0 in $\mathbb{R}^m$ contained in V such that $W \times C \subseteq B$, where C is an open ball around 0 in $\mathbb{R}^{m-n}$. The diffeomorphism h extends to a diffeomorphism ϕ on $W \times C$ defined by $\phi(y, z) = (h(y), z)$. The diffeomorphism $f = \phi \circ g$ defined on a smaller neighborhood of x_0 is a simultaneous straightening function for M and N around x_0. 👍

XIV.1.2.9. This can be iterated: any (finite) chain of submanifolds can be simultaneously straightened.

XIV.1.3. Tangent Space and Tangent Bundle

A curve in $\mathbb{R}^2$ has a tangent line at each point, and a surface in $\mathbb{R}^3$ has a tangent plane at each point (). If the curve or surface is the graph of a function, the tangent line or plane can be described simply. Every submanifold of $\mathbb{R}^p$ has a tangent space at each point generalizing these.

XIV.1.3.1. Let M be an m-submanifold of $\mathbb{R}^p$, and $x \in M$. To define the tangent space to M at x, let f be a straightening function for M around x, a diffeomorphism from an open set in $\mathbb{R}^p$ containing x onto an open ball B in $\mathbb{R}^p$ around 0, with $f(U \cap M) = B \cap \mathbb{R}^m$ and $f(x) = 0$. The inverse map f^{-1} has invertible total differential $d(f^{-1})(0)$, an invertible linear transformation from $\mathbb{R}^p$ to $\mathbb{R}^p$. This linear transformation sends $\mathbb{R}^m$ onto an m-dimensional vector subspace of $\mathbb{R}^p$.

XIV.1.3.2. Definition. The image of $\mathbb{R}^m$ under $d(f^{-1})(0)$ is the *tangent space* to M at x, denoted T_xM.

XIV.1.3.3. To see that this subspace is independent of the straightening function, let g be another straightening function around x. By cutting down neighborhoods, $f \circ g^{-1}$ is a diffeomorphism of a neighborhood of 0 to another neighborhood of 0 leaving the intersection with $\mathbb{R}^m$ invariant, so $d(f \circ g^{-1})(0)$ is an invertible linear map on $\mathbb{R}^p$ sending $\mathbb{R}^m$ onto $\mathbb{R}^m$. Thus $d(g^{-1})(0)$, which equals $d(f^{-1})(0) \circ d(f \circ g^{-1})(0)$ by the Chain Rule, sends the subspace $\mathbb{R}^m$ to the same subspace as $d(f^{-1})(0)$.

XIV.1.3.4. Here is a better way to think of the tangent space at x. Let ϕ be a local diffeomorphism from M to $\mathbb{R}^m$ around x with $\phi(x) = 0$. Then ϕ^{-1} is a smooth function from $\mathbb{R}^m$ to $\mathbb{R}^p$ with $\phi^{-1}(0) = x$, so $d(\phi^{-1})(0)$ is a linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^p$. The tangent space T_xM is the range of $d(\phi^{-1})(0)$, as is easily seen from the proof of (). The trouble with defining the tangent space this way is that it is (slightly) harder to see directly that it is an m-dimensional subspace, i.e. that $d(\phi^{-1})(0)$ has rank m.

XIV.1.3.5. The tangent space T_xM to M at x is a vector subspace of $\mathbb{R}^p$, and hence goes through the origin and (usually) not x. But it can be translated by x to give the affine space tangent to M at x, thinking of a vector $v \in T_xM$ as a vector "beginning at x" to $x + v$, as is easily seen in the curve and surface cases where it can be pictured.

The Tangent Bundle

The tangent spaces to M can all be tied together into one space:

XIV.1.3.6. Definition. Let M be an m-submanifold of $\mathbb{R}^p$. The *tangent bundle* of M, denoted TM, is

$$TM = \{(x, v) : x \in M, v \in T_xM\} \subseteq \mathbb{R}^p \times \mathbb{R}^p \cong \mathbb{R}^{2p} .$$

XIV.1.3.7. Proposition. If M is an m-manifold in $\mathbb{R}^p$, then TM is a $2m$-manifold in $\mathbb{R}^{2p}$.

XIV.1.3.8. TM is also a vector bundle over M in the sense of (). The local triviality can be seen by straightening, since a straightening around a point x is also a straightening around y for any y sufficiently close to x.

XIV.1.3.9. EXAMPLE. Even if M is the graph of a function from $\mathbb{R}$ to $\mathbb{R}$, the simplest kind of 1-manifold in $\mathbb{R}^2$, the tangent bundle is a 2-manifold in $\mathbb{R}^2 \times \mathbb{R}^2 \cong \mathbb{R}^4$ and thus is difficult to visualize. Suppose g is a smooth function from $\mathbb{R}$ to $\mathbb{R}$ and Γ is its graph. If $x_0 \in \mathbb{R}$, a smoothing for Γ around $(x_0, g(x_0))$ is given by $f(x, y) = (x - x_0, y - g(x))$, so $f^{-1}(x, y) = x + x_0, y + g(x + x_0))$ and we have

$$D(f^{-1})(0,0) = \begin{bmatrix} 1 & 0 \\ g'(x_0) & 1 \end{bmatrix}$$

so $T_{(x_0,g(x_0))}\Gamma$ is the subspace of $\mathbb{R}^2$ spanned by the image of (1,0):

$$T_{(x_0,g(x_0))}\Gamma = \{(t, g'(x_0)t) : t \in \mathbb{R}\}$$

which is the line through the origin with slope $g'(x_0)$, as it should be.

This can be seen alternately by taking the local homeomorphism $\phi : \Gamma \to \mathbb{R}^1$ given by $\phi(x, g(x)) = x - x_0$. Then $\phi^{-1}(x) = (x + x_0, g(x + x_0))$, so

$$D(\phi^{-1})(0) = \begin{bmatrix} 1 \\ g'(x_0) \end{bmatrix}$$

and $T_{(x_0,g(x_0))}\Gamma$ is the range of the linear transformation with this matrix.

The tangent bundle is thus

$$T\Gamma = \{(x, y, z, w) : y = g(x), w = g'(x)z)\} = \{(x, g(x), z, g'(x)z) : x, z \in \mathbb{R}\} \subseteq \mathbb{R}^4 \ .$$

This can be pictured by ignoring the second coordinate (projecting into (x, z, w)-space), i.e. "straightening" the graph. We obtain the surface

$$S = \{(x, z, g'(x)z) : x, z \in \mathbb{R}\}$$

which is the graph of the function $h(x, z) = g'(x)z$. The section of this surface by the plane parallel to the zw-plane with x-coordinate x_0 is the straight line through the origin with slope $g'(x_0)$.

XIV.1.4. Derivatives

XIV.1.5. Immersions, Submersions, and Embeddings

XIV.1.6. Transversality

XIV.2. Abstract Smooth Manifolds

XIV.2.1. Charts and Atlases

Chapter XV

Algebraic Topology

Algebraic topology is one of the most important parts of the vast subject of topology. In a nutshell, it is the process of assigning algebraic invariants to topological spaces and using algebraic techniques to obtain results about the topology of spaces. One of the most important uses is to distinguish topological spaces by showing that an algebraic invariant differs.

Algebraic topology is an outgrowth of "combinatorial topology" (Emmy Noether was one of the most influential people in the transition in the 1920s and 1930s), and much of algebraic topology has a distinct combinatorial flavor. Indeed, homological algebra (which can be thought of as "algebraic topology with the topology abstracted out") is heavily combinatorial.

Topology is often thought of as dividing into two rather separate parts: "point-set topology" and "algebraic topology" ("point-set topology" is the subject matter of Chapter ()). Almost all topology texts are strictly limited to one or the other, although algebraic topology texts must include a certain amount, generally small, of point-set topology. Algebraic topology has developed into a huge discipline in its own right. But an arguably better point of view is that algebraic topology is a fundamental tool within the overall subject of topology, where "point-set topology" is the foundational material. A good, but imperfect, analogy is to compare algebraic topology within topology to functional analysis as a part of analysis; in this analogy point-set topology might be analogous to calculus. There are other subdisciplines in topology such as "low-dimensional topology," just as there are other subdisciplines in analysis such as differential equations; low-dimensional topology heavily uses algebraic topology as a tool, just as the subject of differential equations uses functional analysis.

The Limitations of Algebraic Topology

Although algebraic topology is extremely powerful within its domain of application, it has at least two important limitations:

(i) Most algebraic topology techniques only apply or are useful within a tiny corner of the world of topological spaces, although this corner consists of the most nicely-behaved and arguably most interesting and important topological spaces (manifolds, cell complexes, etc.)

(ii) Even within this well-behaved corner, algebraic topology invariants are homotopy invariants, and hence cannot distinguish between homotopy-equivalent spaces. Pure algebraic topology cannot even distinguish the real line from a single point! Some elementary point-set topology is needed. A little more point-set topology is needed to distinguish $\mathbb{R}^1$ from $\mathbb{R}^2$, or $\mathbb{R}^n$ and $\mathbb{R}^m$ for $n \neq m$ (the one-point compactification operation

plus some algebraic topology is the easiest way). Algebraic topology cannot even always distinguish between spaces which are not homotopy equivalent, such as the real line and the long line (), where some elementary point-set topology does the trick.

The moral is that algebraic and point-set techniques are intertwined and need to be used together.

Our treatment of algebraic topology will be rudimentary, partly because much of it is somewhat off the subject of analysis. A far better and more complete online treatment is [Hat02]. There are also many books on the subject. Actually algebraic topology is such a large discipline that no book can treat the entire subject, so interested readers should explore widely.

XV.1. Homotopy and the Fundamental Group

XV.1.1. Homotopy

Homotopy is the careful mathematical formulation of the notion of "continuous deformation," one of the most important ideas in topology and the basis for the entire subject of algebraic topology. There are several slight variations which we will discuss. If f and g are continuous functions between topological spaces X and Y, we want to describe how f can be "continuously deformed" to g; the intuitive idea is that there should be a "continuous path" (f_t) (say, for $0 \le t \le 1$) of continuous functions with $f_0 = f$, $f_1 = g$. Experience shows that the following definition is the appropriate one:

XV.1.1.1. DEFINITION. Let X and Y be topological spaces, and $f_0, f_1 : X \to Y$ continuous functions. A *homotopy* from f_0 to f_1 is a continuous function $H : X \times [0,1] \to Y$ with $H(x,0) = f_0(x)$ and $H(x,1) = f_1(x)$ for all $x \in X$. Two continuous functions f and g from X to Y are *homotopic*, written $f \simeq g$, if there exists a homotopy from f to g.

XV.1.1.2. We can then define, for fixed $t \in [0,1]$, a continuous function $f_t : X \to Y$ by $f_t(x) = H(x,t)$ for $x \in X$. The functions f_t "vary continuously" from f_0 to f_1 as t varies from 0 to 1. The f_t are "jointly continuous" in x and t. We often think of the homotopy H as being the family (f_t) of deformations of f_0 to f_1.

XV.1.1.3. The choice of $[0,1]$ for the parametrizing interval is arbitrary, as is the specification that H goes from $X \times [0,1]$ to Y instead of from $[0,1] \times X$ to Y (the opposite convention is also common in references). To add to the potential notational confusion, in many applications the space X will also be $[0,1]$.

XV.1.1.4. In considering homotopies, the target space Y must be carefully specified or understood. If Y is a subspace of a larger space Z, and $f, g : X \to Y$, then f and g can well be homotopic as maps from X to Z but not as maps from X to Y, i.e. a homotopy might necessarily pass through points of Z outside Y. Thus we should really write $f \simeq g$ in Y (although we usually don't).

XV.1.1.5. If $f \simeq g$ via a homotopy H, then for any $x \in X$ the function $\gamma_x : [0,1] \to Y$ defined by $\gamma_x(t) = H(x,t)$ is a continuous path in Y from $f(x)$ to $g(x)$. In particular, if $f \simeq g$, then for every $x \in X$, $f(x)$ and $g(x)$ must be in the same path component of Y.

XV.1.1.6. EXAMPLES. (i) Two constant functions from X to Y with values y_0 and y_1 are homotopic if and only if y_0 and y_1 are in the same path component of Y. Necessity comes from XV.1.1.5.. Conversely, if $\gamma : [0,1] \to Y$ is a path from y_0 to y_1, set $H(x,t) = \gamma(t)$; then H is a homotopy (via constant functions). In particular, if Y is path-connected, any two constant functions from any X to Y are homotopic.

(ii) A very special but important case is when the space Y is a convex subset of $\mathbb{R}^n$ or, more generally, of a topological vector space. If X is any topological space, then any two continuous functions f and g from X to Y are homotopic by a homotopy of the form

$$H(x,t) = tg(x) + (1-t)f(x) \ .$$

Such a homotopy is called a *linear homotopy*, and is defined only when Y has additional structure from a topological vector space.

The first fundamental property of homotopy is:

XV.1.1.7. PROPOSITION. Let X and Y be topological spaces. Then homotopy is an equivalence relation on the set of maps (continuous functions) from X to Y.

PROOF: Reflexivity and symmetry are very simple, but transitivity, although it is relatively simple also, involves an important notion of concatenation for homotopies.
If $f : X \to Y$, define $H(x,t) = f(x)$ for all x, t. Then H is a homotopy from f to f, so homotopy is reflexive.
If H is a homotopy from f_0 to f_1, define $\tilde{H} : X \times [0,1] \to Y$ by $\tilde{H}(x,t) = H(x, 1-t)$. Then $\tilde{H}$ is a homotopy from f_1 to f_0. Thus homotopy is symmetric.
Let H_1 be a homotopy from f_0 to f_1, and H_2 a homotopy from f_1 to f_2. Define $H : X \times [0,1] \to Y$ by

$$H(x,t) = \begin{cases} H_1(x,2t) & \text{if} \quad 0 \le t \le \frac{1}{2} \\ H_2(x,2t-1) & \text{if} \quad \frac{1}{2} < t \le 1 \end{cases} .$$

Then H is continuous since $H_1(x,1) = H_2(x,0) = f_1(x)$ for all x, and is a homotopy from f_0 to f_2. 👍

Homotopy respects compositions:

XV.1.1.8. PROPOSITION. Let X, Y, Z be topological spaces, and let $f, f_0, f_1 : X \to Y$ and $g, g_0, g_1 : Y \to Z$ with $f_0 \simeq f_1$ and $g_0 \simeq g_1$. Then $g_0 \circ f_0 \simeq g_1 \circ f_1$. In particular, $g \circ f_0 \simeq g \circ f_1$ and $g_0 \circ f \simeq g_1 \circ f$.

PROOF: We need only prove the last statement, since then $g_0 \circ f_0 \simeq g_0 \circ f_1$ and $g_0 \circ f_1 \simeq g_1 \circ f_1$, so $g_0 \circ f_0 \simeq g_1 \circ f_1$ by XV.1.1.7..
If H is a homotopy from f_0 to f_1, define $\tilde{H}(x,t) = g(H(x,t))$; then $\tilde{H}$ is a homotopy from $g \circ f_0$ to $g \circ f_1$.
If H is a homotopy from g_0 to g_1, define $\tilde{H}(x,t) = H(f(x),t)$; then $\tilde{H}$ is a homotopy from $g_0 \circ f$ to $g_1 \circ f$.
👍

Contractible Spaces

XV.1.1.9. DEFINITION. If X and Y are topological spaces, and $f : X \to Y$ is a map, then f is *nullhomotopic* if it is homotopic to a constant function.
A topological space X is *contractible* if the identity map on X is nullhomotopic, i.e. homotopic to a constant function from X to X.

XV.1.1.10. If X is contractible and H is a homotopy from the identity map ι_X to a constant function with value p, then the functions f_t "contract" the space down to the point p, and we say X is *contractible to p*. But a contractible space is path-connected since every point of X can be connected to p by a path, so if X is contractible to p it is also contractible to q for any other point q.

XV.1.1.11. EXAMPLES. (i) If X is a convex subset of $\mathbb{R}^n$, or more generally of a topological vector space, then X is contractible to any point vis a linear homotopy (XV.1.1.6.(ii)). In particular, $[0,1]$ is contractible, as is any ball in $\mathbb{R}^n$.

(ii) Let X be the infinite comb (). Then X is contractible by first moving all points down to the x-axis and then moving them to $\mathbf{0}$.

(iii) It is a fundamental theorem of topology, equivalent to the Brouwer fixed-point theorem, that S^n is not contractible for any n (). Note that S^n is path-connected for $n > 0$.

The next result is a special case of XV.1.1.8..

XV.1.1.12. PROPOSITION. Let X and Y be topological spaces. If either X or Y is contractible, then every map from X to Y is nullhomotopic. If Y is path-connected, any two maps from X to Y are homotopic.

Homotopy Relative to a Subspace

An important variation is homotopy relative to a subspace:

XV.1.1.13. DEFINITION. Let X and Y be topological spaces, and $A \subseteq X$, $B \subseteq Y$. Let f_0, f_1 be maps from X to Y for which $f_0(A), f_1(A) \subseteq B$. Then a *homotopy from f_0 to f_1 relative to* (A, B) is a continuous function $H : X \times [0,1] \to Y$ such that $H(x,0) = f_0(x)$ and $H(x,1) = f_1(x)$ for all $x \in X$ and $H(x,t) \in B$ for all $x \in A$ and all t. If there is such a homotopy, write $f_0 \simeq f_1$ rel (A, B).

XV.1.1.14. A homotopy relative to a subspace is a homotopy. But the notion is much more restrictive, since the image of the set A must remain within B under the entire homotopy. In many applications, A and B are sets with only one or two points, and the restriction that f take A into B amounts to specifying the value of f at the point(s) of A.

XV.1.1.15. An important example of relative homotopy is when $X = [0,1]$ and $A = \{0,1\}$. A map γ from $[0,1]$ to Y is a path in Y. If γ_0 and γ_1 are paths in Y with the same initial point p and final point q, then γ_0 and γ_1 are always homotopic since they are both homotopic to the constant map with value p; but if $B = \{p, q\}$ a homotopy rel (A, B) is one in which all intermediate paths also go from p to q (as long as p and q have disjoint neighborhoods in Y), so γ_0 and γ_1 may not be homotopic rel (A, B). For example, if Y is the unit circle in $\mathbb{R}^2$ and γ_0 and γ_1 are paths in Y from $p = (-1, 0)$ to $q = (1, 0)$ via the upper and lower semicircles respectively, then γ_0 and γ_1 are not homotopic rel $(\{0,1\}, \{p, q\})$ (this seems intuitively obvious, but requires a nontrivial proof, and is equivalent to the Brouwer fixed-point theorem in the plane). In this case we often notationally suppress the B (if γ_0 and γ_1 have the same endpoints) and simply write $\gamma_0 \simeq \gamma_1$ rel $\{0,1\}$. Even the A is often understood in this context, but a reader must carefully understand whether it is assumed (as it most commonly is) that homotopies of paths fix endpoints.

XV.1.1.16. An especially important special case of XV.1.1.15. comes when $p = q$, i.e. B is a singleton p. We thus consider *loops with basepoint p*, maps $\gamma : [0,1] \to Y$ with $\gamma(0) = \gamma(1) = p$. Two such loops are *homotopic rel p* if they are homotopic rel $(\{0,1\}, \{p\})$, i.e. homotopic via loops γ_t with basepoint p for all t. These will be discussed in detail in ().

XV.1.1.17. Here is another interesting special case. If X is a topological space and $p \in X$, then X is *contractible to p relative to p* (rel p) if the identity map is homotopic to the constant map with value p relative to $(\{p\}, \{p\})$, i.e. if the space can be contracted to p in a way leaving p fixed. A convex set X in a topological vector space is contractible to p rel p for any $p \in X$, but if X is the infinite comb and p is a point on the y-axis other than $\mathbf{0}$, then X is not contractible to p rel p since all points must move through $\mathbf{0}$ in any contraction to p (XV.1.5.4.).

A slight variation of the proofs of XV.1.1.7. and XV.1.1.8. (left to the reader) show:

XV.1.1.18. PROPOSITION. Let X, Y, Z be topological spaces, and $A \subseteq X$, $B \subseteq Y$, $C \subseteq Z$. Then

(i) Homotopy rel (A, B) is an equivalence relation on the set of maps from X to Y.

(ii) If $f_0, f_1 : X \to Y$, $f_0 \simeq f_1$ rel (A, B), and $g_0, g_1 : Y \to Z$, $g_0 \simeq g_1$ rel (B, C), then $g_1 \circ f_1 \simeq g_0 \circ f_0$ rel (A, C).

Strong and Local Contractibility

XV.1.1.19. DEFINITION. Let X be a topological space, and $p \in X$. Then X is *strongly contractible to p* if it is contractible to p rel $\{p\}$. X is *locally [strongly] contractible to p* if for every open neighborhood U of p, there is an open neighborhood V of p, contained in U, such that V is [strongly] contractible to p within U.

XV.1.1.20. These notions are all distinct. A circle is locally strongly contractible to any point, but not contractible. The infinite comb () is contractible to any point but not strongly contractible or locally contractible to points on the y-axis (this also shows that a contractible space can be strongly contractible to some points but not others). The example of () is locally contractible to any point but not locally strongly contractible or strongly contractible to some points.

XV.1.1.21. PROPOSITION. Let X be a topological space and $p \in X$. If X is strongly contractible to p, then X is locally strongly contractible to p.

PROOF: Let $h : X \times [0,1] \to X$ be a strong contraction to p, i.e. $h(x, 0) = x$ and $h(x, 1) = p$ for all $x \in X$, and $h(p, t) = p$ for all t. Let U be an open neighborhood of p. Let $W = h^{-1}(U)$; then W is an open neighborhood of $\{p\} \times [0,1]$ in $X \times [0,1]$. Since $[0,1]$ is compact, there is an open neighborhood V of p such that $V \times [0,1] \subseteq W$ (). Then h strongly contracts V to p within U. 👍

XV.1.1.22. This result is false in general if "strongly" is removed, as the infinite comb shows.

Homotopy Equivalence

XV.1.1.23. DEFINITION. Let X and Y be topological spaces. Then X and Y are *homotopy equivalent* if there are maps $f : X \to Y$ and $g : Y \to X$ with $g \circ f \simeq \iota_X$ and $f \circ g \simeq \iota_Y$.
X *homotopically dominates* Y if there are maps $f : X \to Y$ and $g : Y \to X$ with $f \circ g \simeq \iota_Y$.

XV.1.1.24. It is an easy consequence of XV.1.1.8. that homotopy dominance is transitive and homotopy equivalence is an equivalence relation. Homotopy dominance is a sort of "homotopy retraction" ().

XV.1.1.25. Examples. (i) Every topological space homotopically dominates any contractible space. A topological space is homotopy equivalent to a contractible space if and only if it is contractible. Any two contractible spaces are homotopy equivalent.

(ii) A circle in $\mathbb{R}^2$ is not homotopy equivalent to $\mathbb{R}^2$ since $\mathbb{R}^2$ is contractible. But a circle is homotopy equivalent to $\mathbb{R}^2$ with a point removed.

(iii) The Greek letter θ and a figure 8 are homotopy equivalent.

(iv) A cylinder and a Möbius strip are homotopy equivalent since each is homotopy equivalent to a circle.

(v) If X and Y are homotopy equivalent, they each homotopy dominate the other. But the converse is false in general:

XV.1.1.26. Homotopy equivalence is *much* weaker than homeomorphism in general: for example, every contractible space is homotopy equivalent to a one-point space. (Homeomorphism can be concluded from homotopy equivalence in some special cases.) Most methods and results of algebraic topology do not distinguish between homotopy equivalent spaces, i.e. algebraic topology describes and studies topological spaces up to homotopy equivalence.

XV.1.1.27. We can also define relative homotopy equivalence. If $A \subseteq X$ and $B \subseteq Y$, a homotopy equivalence between X and Y rel (A, B) is a pair of maps $f : X \to Y$ and $g : Y \to X$ with $f(A) \subseteq B$, $g(B) \subseteq A$, and $g \circ f \simeq \iota_X$ rel (A, A) and $f \circ g \simeq \iota_Y$ rel (B, B). Relative homotopy dominance can be defined similarly.

Smooth Homotopy

XV.1.1.28. In situations where there is a notion of differentiability, one also wants to consider smooth homotopies. For example, if X and Y are $\mathcal{C}^r$-manifolds and f, g are $\mathcal{C}^r$ maps from X to Y, we say f and g are $\mathcal{C}^r$-homotopic if there is a homotopy H which is a $\mathcal{C}^r$ function from $X \times [0, 1]$ to Y. The relation between ordinary homotopies and smooth homotopies in this context is somewhat subtle, but in good cases an ordinary homotopy can be "smoothed" to a smooth one ().

Isotopy

In some applications in topology, such as knot theory, it is important not to allow the image to "pass through itself" during the deformation. This is accomplished by the notion of isotopy.

XV.1.1.29. Definition. Let X and Y be topological spaces, and f_0, f_1 topological embeddings of X into Y (homeomorphisms onto their images). An *isotopy* from f_0 to f_1 is a homotopy $H = (f_t)$ from f_0 to f_1 in which each f_t $(0 \leq t \leq 1)$ is a topological embedding.

Isotopic maps are obviously homotopic, but the converse is false:

XV.1.1.30. EXAMPLES. (i) Let $X = [0,1]$, $Y = [0,1]$ (or $Y = \mathbb{R}$). Let $f_0(t) = t$, and $f_1(t) = 1-t$. Then $f_0 \simeq f_1$ (any two maps from X to Y are homotopic). But f_0 and f_1 are not isotopic: if (f_t) is a homotopy from f_0 to f_1, there is a t such that $f_t(0) = f_t(1)$, so f_t is not injective [set $\phi(t) = f_t(1) - f_t(0)$ and apply the Intermediate Value theorem].

(ii) Any two knots () are homotopic as maps from S^1 to $\mathbb{R}^3$. But a nontrivial knot like the trefoil () is not isotopic to an unknot. In fact, two knots are isotopic if and only if they are equivalent as knots.

(iii) The Fox-Artin wild arc () is isotopic to an ordinary arc as maps from $[0,1]$ to $\mathbb{R}^3$ [pull in the ends a little]. But the more complicated embedding of $[0,1]$ into $\mathbb{R}^3$ outlined at the end of () which is wild at every point is not isotopic to an ordinary arc.

Ambient Isotopy

XV.1.2. The Fundamental Group

XV.1.3. Covering Spaces

XV.1.4. Other Homotopy Groups

n-Homotopy

XV.1.4.1. If $n \in \mathbb{N}$, and $f_0, \ldots, f_n$ are $n+1$ maps from X to Y, an *n-homotopy* of $f_0, \ldots, f_n$ is a map h from $X \times \Delta_n$ to Y, where Δ_n is the n-simplex (), with $h(x, e_k) = f_k(x)$ for $0 \leq k \leq n$, where e_k is the k'th vertex of Δ_n. A 1-homotopy is just a homotopy. Possible variations include n-homotopies relative to a subspace of X or relative to specified 1-homotopies between the f_k.

An n-homotopy of $f_0, \ldots, f_n$ gives an m-homotopy of any $m+1$ of the functions, for $1 \leq m \leq n$, but is more than that: it gives a coherent family of such homotopies.

This notion is only interesting when smaller m-homotopies are specified:

XV.1.4.2. PROPOSITION. Let $f_0, \ldots, f_n$ be maps from X to Y. If the f_k are pairwise homotopic, then there is an n-homotopy of $f_0, \ldots, f_n$.

PROOF: We write the proof for $n = 2$; the general case is similar but notationally more complicated. Let $h_1 : X \times [0,1] \to Y$ be a homotopy from f_0 to f_1 and h_2 a homotopy from f_0 to f_2. Let L be the line segment in Δ_2 from e_0 to the midpoint of the edge between e_1 and e_2. Define $h : X \times \Delta_2 \to Y$ as follows. If $p \in L$, for $x \in X$ set $h(x,p) = f_0(x)$. If $p = te_1 + (1-t)q$ for $q \in L$, for $x \in X$ set $h(x,p) = h_1(x,t)$, and if $p = te_2 + (1-t)q$ for $q \in L$ set $h(x,p) = h_2(x,t)$. Then h is a 2-homotopy of f_0, f_1, f_2. 👍

XV.1.4.3. EXAMPLE. If 1-homotopies between the f_k are specified, a compatible 2-homotopy need not exist. Let $X = \{p\}$ be a singleton and Y a circle, and y_0, y_1, y_2 vertices of an equilateral triangle in Y. Let f_0, f_1, f_2 be the functions sending p to y_0, y_1, y_2 respectively. Specify the homotopies to move the points cyclically along the minor arcs between points. Then there is no 2-homotopy, which would effectively be a retraction of a disk onto a circle.

This example shows that existence of n-homotopies with specified boundary behavior is closely related to higher-order connectedness.

XV.1.5. Exercises

XV.1.5.4. Let X be the infinite comb (), and p a point of X on the y-axis with positive y-coordinate. Then X is contractible to p. Show that X is not contractible to p rel p:

(a) If q is a point of X with positive x-coordinate, then any path in X from q to p must pass through $\mathbf{0}$.

(b) Let (p_n) be a sequence in X of points with positive x-coordinate with $p_n \to p$. If H is a contraction of X to p, then for each n there must be a t_n with $H(p_n, t_n) = \mathbf{0}$.

(c) If t_0 is a limit point of the sequence (t_n), then $H(p, t_0) = \mathbf{0}$. Thus we cannot have that H is a contraction of X to p rel p.

XV.2. Homology

XV.2.1. Singular Homology

XV.2.2. Simplicial Homology

XV.3. Homological Algebra

XV.3.1. Preliminaries

Diagram Chasing

A common proof technique in homological algebra is diagram chasing. We illustrate the technique with a basic result called the *Five Lemma*. The film *It's My Turn* has a scene where Jill Clayburgh presents a diagram chase proof of the Snake Lemma (XV.3.1.10.).

XV.3.1.1. Theorem. [Basic Five Lemma] Let R be a ring, and let

$$\begin{array}{ccccccccc} A_{11} & \xrightarrow{\alpha_1} & A_{12} & \xrightarrow{\alpha_2} & A_{13} & \xrightarrow{\alpha_3} & A_{14} \\ \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} & & \downarrow{\scriptstyle \phi_4} \\ A_{21} & \xrightarrow{\beta_1} & A_{22} & \xrightarrow{\beta_2} & A_{23} & \xrightarrow{\beta_3} & A_{24} \end{array}$$

be a commutative diagram of R-modules, with exact rows. If ϕ_2 and ϕ_4 are injective and ϕ_1 is surjective, then ϕ_3 is injective.

The result applies for either left or right modules. Since abelian groups are the same thing as $\mathbb{Z}$-modules, it also applies to abelian groups. In fact, it is valid in any abelian category ().

The result can be rephrased:

XV.3.1.2. Theorem. [Basic Five Lemma] Let R be a ring, and let

$$\begin{array}{ccccccccc} & & & & 0 & & & & \\ & & & & \downarrow & & & & \\ & & 0 & & B & & 0 & & \\ & & \downarrow & & \downarrow & & \downarrow & & \\ A_{11} & \xrightarrow{\alpha_1} & A_{12} & \xrightarrow{\alpha_2} & A_{13} & \xrightarrow{\alpha_3} & A_{14} \\ \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} & & \downarrow{\scriptstyle \phi_4} \\ A_{21} & \xrightarrow{\beta_1} & A_{22} & \xrightarrow{\beta_2} & A_{23} & \xrightarrow{\beta_3} & A_{24} \\ \downarrow & & & & & & \\ 0 & & & & & & \end{array}$$

be a commutative diagram of R-modules, with exact rows and columns. Then $B = 0$.

Proof: We will use the following notation. If $1 \leq j, k \leq 4$, write x_{jk} for a generic element of A_{jk}. If x_{jk} has been specified, and $j \leq m \leq 4$, $k \leq n \leq 4$, write x_{mn} for the image of x_{jk} in A_{mn} (this image is unambiguous since the diagram commutes). Also since the map from B into A_{13} is injective, we may regard B as a subset

of A_{13}. So let $x_{13} \in B \subseteq A_{13}$; we want to show $x_{13} = 0$. The image x_{23} of x_{13} in A_{23} is 0 by exactness; thus $x_{24} = 0$. Since ϕ_4 is injective, $x_{14} = 0$. Thus x_{13} has a preimage $x_{12} \in A_{12}$ by exactness. The image $x_{22} \in A_{22}$ goes to $x_{23} = 0$, and thus by exactness x_{22} has a preimage $x_{21} \in A_{21}$. Since ϕ_1 is surjective, x_{21} has a preimage $x_{11} \in A_{11}$. If y is the image of x_{11} in A_{12}, y and x_{12} have the same image x_{22} in A_{22}, and thus $y = x_{12}$ since ϕ_2 is injective. Thus x_{12} is in the image of A_{11}, and so the image x_{13} in A_{13} is 0 by exactness. 👍

We have named the elements in each module to make the argument clearer and more complete; but in serious diagram chasing it is considered unsporting to give the elements names – one just starts with an element somewhere and moves ("chases") it around the diagram.

There is a dual version:

XV.3.1.3. THEOREM. [BASIC FIVE LEMMA, DUAL VERSION] Let R be a ring, and let

$$\begin{array}{ccccccc} A_{11} & \xrightarrow{\alpha_1} & A_{12} & \xrightarrow{\alpha_2} & A_{13} & \xrightarrow{\alpha_3} & A_{14} \\ \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} & & \downarrow{\scriptstyle \phi_4} \\ A_{21} & \xrightarrow{\beta_1} & A_{22} & \xrightarrow{\beta_2} & A_{23} & \xrightarrow{\beta_3} & A_{24} \end{array}$$

be a commutative diagram of R-modules, with exact rows. If ϕ_1 and ϕ_3 are surjective and ϕ_4 is injective, then ϕ_2 is surjective.

The result again applies for either left or right modules, and even in any abelian category ().

The result can be rephrased:

XV.3.1.4. THEOREM. [BASIC FIVE LEMMA, DUAL VERSION] Let R be a ring, and let

$$\begin{array}{ccccccc} & & & & & & 0 \\ & & & & & & \downarrow \\ A_{11} & \xrightarrow{\alpha_1} & A_{12} & \xrightarrow{\alpha_2} & A_{13} & \xrightarrow{\alpha_3} & A_{14} \\ \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} & & \downarrow{\scriptstyle \phi_4} \\ A_{21} & \xrightarrow{\beta_1} & A_{22} & \xrightarrow{\beta_2} & A_{23} & \xrightarrow{\beta_3} & A_{24} \\ \downarrow & & \downarrow & & \downarrow & & \\ 0 & & B & & 0 & & \\ & & \downarrow & & & & \\ & & 0 & & & & \end{array}$$

be a commutative diagram of R-modules, with exact rows and columns. Then $B = 0$.

This can be proved either by dualizing, or by a direct diagram chase. The direct proof is left as an exercise for the reader (), and is good practice in diagram chasing. It is slightly more complex than the proof of XV.3.1.2., but similar.

Putting the two versions together, we obtain the commonly stated version, which is the origin of the name *Five Lemma*:

XV.3.1.5. Corollary. [Five Lemma, Common Version] Let R be a ring, and let

$$\begin{array}{ccccccccc} A_{11} & \xrightarrow{\alpha_1} & A_{12} & \xrightarrow{\alpha_2} & A_{13} & \xrightarrow{\alpha_3} & A_{14} & \xrightarrow{\alpha_4} & A_{15} \\ \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} & & \downarrow{\scriptstyle \phi_4} & & \downarrow{\scriptstyle \phi_5} \\ A_{21} & \xrightarrow{\beta_1} & A_{22} & \xrightarrow{\beta_2} & A_{23} & \xrightarrow{\beta_3} & A_{24} & \xrightarrow{\beta_4} & A_{25} \end{array}$$

be a commutative diagram of R-modules, with exact rows. If ϕ_1, ϕ_2, ϕ_4, and ϕ_5 are isomorphisms, then ϕ_3 is also an isomorphism.

The special case of short exact sequences is worth stating:

XV.3.1.6. Corollary. [Five Lemma, Short Version] Let R be a ring, and let

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A_{12} & \xrightarrow{\alpha_2} & A_{13} & \xrightarrow{\alpha_3} & A_{14} & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} & & \downarrow{\scriptstyle \phi_4} & & \\ 0 & \longrightarrow & A_{22} & \xrightarrow{\beta_2} & A_{23} & \xrightarrow{\beta_3} & A_{24} & \longrightarrow & 0 \end{array}$$

be a commutative diagram of R-modules, with short exact rows. If ϕ_2 and ϕ_4 are injective [resp. surjective, isomorphisms], then ϕ_3 is also injective [resp. surjective, an isomorphism].

Connecting Homomorphisms and the Snake Lemma

XV.3.1.7. Consider left R-modules over a fixed ring R (this is not important; we could work in any abelian category). If $\phi : A \to B$ is a module homomorphism, as usual define $ker(\phi) = \{x \in A : \phi(x) = 0\}$. This is a submodule of A. Also define $coker(\phi) = B/\phi(A)$, a quotient module. There is then an exact sequence

$$0 \longrightarrow ker(\phi) \xrightarrow{\iota} A \xrightarrow{\phi} B \xrightarrow{\pi} coker(\phi) \longrightarrow 0$$

where ι is the inclusion map and π the quotient map. Conversely, if

$$0 \longrightarrow K \longrightarrow A \xrightarrow{\phi} B \longrightarrow C \longrightarrow 0$$

is an exact sequence including ϕ, then K and C are naturally isomorphic to $ker(\phi)$ and $coker(\phi)$ respectively, so the exact sequence is essentially unique.

XV.3.1.8. If

$$\begin{array}{ccccc} A_1 & \xrightarrow{\alpha_1} & A_2 & \xrightarrow{\alpha_2} & A_3 \\ \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} \\ B_1 & \xrightarrow{\beta_1} & B_2 & \xrightarrow{\beta_2} & B_3 \end{array}$$

is a commutative diagram, then α_1 maps $ker(\phi_1)$ into $\ker(\phi_2)$ and α_2 maps $ker(\phi_2)$ into $\ker(\phi_3)$ by commutativity. Similarly, β_1 maps $\phi_1(A_1)$ into $\phi_2(A_2)$ and β_2 maps $\phi_2(A_2)$ into $\phi_3(A_3)$ by commutativity; thus β_1 drops to a homomorphism $\bar{\beta}_1$ from $coker(\phi_1)$ to $coker(\phi_2)$, and β_2 to a homomorphism $\bar{\beta}_2 : coker(\phi_2) \to coker(\phi_3)$, and we get a commutative diagram

$$\begin{array}{ccccc} 0 & & 0 & & 0 \\ \downarrow & & \downarrow & & \downarrow \\ ker(\phi_1) & \xrightarrow{\tilde{\alpha}_1} & ker(\phi_2) & \xrightarrow{\tilde{\alpha}_2} & ker(\phi_3) \\ \downarrow & & \downarrow & & \downarrow \\ A_1 & \xrightarrow{\alpha_1} & A_2 & \xrightarrow{\alpha_2} & A_3 \\ \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} \\ B_1 & \xrightarrow{\beta_1} & B_2 & \xrightarrow{\beta_2} & B_3 \\ \downarrow & & \downarrow & & \downarrow \\ coker(\phi_1) & \xrightarrow{\bar{\beta}_1} & coker(\phi_2) & \xrightarrow{\bar{\beta}_2} & coker(\phi_3) \\ \downarrow & & \downarrow & & \downarrow \\ 0 & & 0 & & 0 \end{array}$$

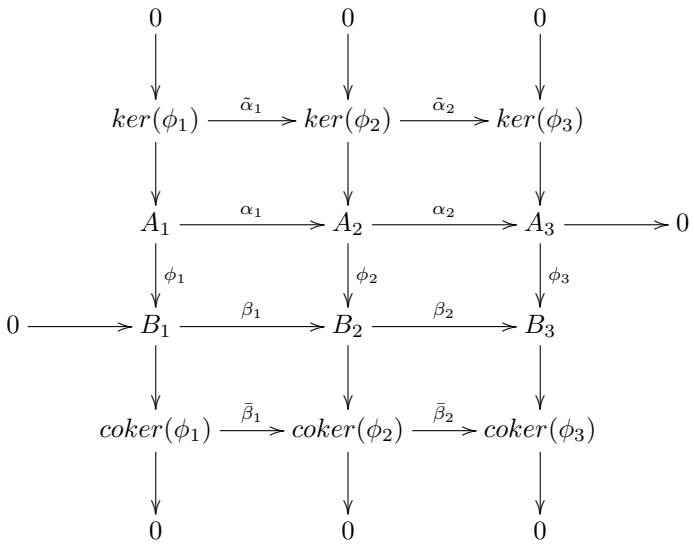
with exact columns, where $\tilde{\alpha}_k$ is the restriction of α_k.

If the original diagram is a little bigger with exact rows, the expanded diagram has exact rows too:

XV.3.1.9. Proposition. Let

$$\begin{array}{ccccccc} & A_1 & \xrightarrow{\alpha_1} & A_2 & \xrightarrow{\alpha_2} & A_3 & \longrightarrow 0 \\ & \downarrow{\scriptstyle \phi_1} & & \downarrow{\scriptstyle \phi_2} & & \downarrow{\scriptstyle \phi_3} & \\ 0 \longrightarrow & B_1 & \xrightarrow{\beta_1} & B_2 & \xrightarrow{\beta_2} & B_3 & \end{array}$$

be a commutative diagram of left R-modules with exact rows. Then the expanded commutative diagram



has exact rows (and columns).

PROOF: Consider the first row. The map $\tilde{\alpha}_2 \circ \tilde{\alpha}_1$ is 0 since $\alpha_2 \circ \alpha_1 = 0$. Conversely, let $x \in ker(\phi_2)$ with $\tilde{\alpha}_2(x) = 0$. Then $\alpha_2(x) = 0$, so there is a $y \in A_1$ with $\alpha_1(y) = x$. Since $x \in ker(\phi_2)$, we have $\beta_1 \circ \phi_1(y) = 0$, and hence $\phi_1(y) = 0$ since β_1 is injective. Thus $y \in ker(\phi_1)$.

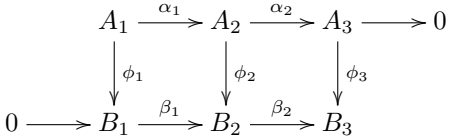
Now consider the last row. We have $\bar{\beta}_2 \circ \bar{\beta}_1 = 0$ since $\beta_2 \circ \beta_1 = 0$. Conversely, let $x \in B_2$ with $[x] \in coker(\phi_2)$ satisfying $\bar{\beta}_2([x]) = 0$. This exactly means $\beta_2(x) \in \phi_3(A_3)$, i.e. there is $y \in A_3$ with $\phi_3(y) = \beta_2(x)$. Since α_2 is surjective, there is $z \in A_2$ with $\alpha_2(z) = y$. We have $\beta_2(x - \phi_2(z)) = 0$, so by exactness there is $w \in B_1$ with $\beta_1(w) = x - \phi_2(z)$. Since $\phi_2(z) \in \phi_2(A_2)$, we have $[x - \phi_2(z)] = [x]$ in $coker(\phi_2)$, and

$$\bar{\beta}_1([w]) = [x - \phi_2(z)] = [x] \ .$$



The following result is called the *Snake Lemma* or the *Connecting Homomorphism Theorem*. The existence of connecting homomorphisms of this type is one of the most important features of homological algebra, particularly for applications in topology.

XV.3.1.10. THEOREM. [SNAKE LEMMA] Let

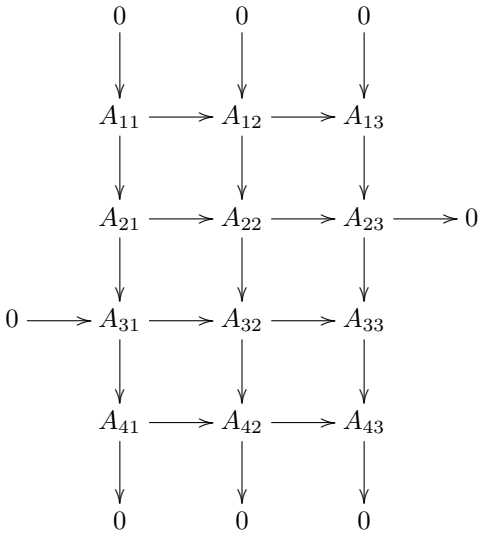


be a commutative diagram of R-modules (or in any abelian category), with exact rows. Then there is a natural connecting homomorphism $d : ker(\phi_3) \to coker(\phi_1)$ making the following sequence exact:

$$ker(\phi_1) \xrightarrow{\tilde{\alpha}_1} ker(\phi_2) \xrightarrow{\tilde{\alpha}_2} ker(\phi_3) \xrightarrow{d} coker(\phi_1) \xrightarrow{\bar{\beta}_1} coker(\phi_2) \xrightarrow{\bar{\beta}_2} coker(\phi_3)$$

To construct the connecting homomorphism and prove the theorem, in light of Proposition XV.3.1.9. it suffices to prove the following, where we change notation to make a diagram chase proof more efficient, and suppress the names of the maps:

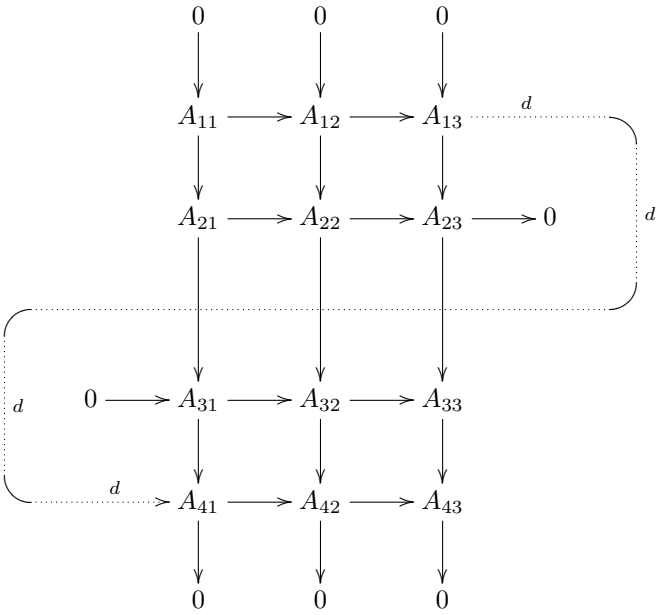
XV.3.1.11. THEOREM. [SNAKE LEMMA] Let



be a commutative diagram of R-modules (or in any abelian category), with exact rows and columns. Then there is a natural connecting homomorphism $d : A_{13} \to A_{41}$ making the following sequence exact:

$$A_{11} \longrightarrow A_{12} \longrightarrow A_{13} \xrightarrow{d} A_{41} \longrightarrow A_{42} \longrightarrow A_{43}$$

To see why this is called the Snake Lemma, the diagram can be drawn like this:



PROOF: The proof will consist of several steps:

(i) Definition of d.

(ii) Showing that d is well defined.

(iii) Showing that d is a homomorphism.

(iv) Showing the sequence is exact at A_{13}.

(v) Showing the sequence is exact at A_{41}.

(Note that by hypothesis the sequence is exact at A_{12} and A_{42}; cf. XV.3.1.9..) We use the notational conventions of the proof of XV.3.1.2., and regard $A_{1k} \subseteq A_{2k}$ for $1 \leq k \leq 3$.

(i): We first define d. Let $x_{13} \in A_{13}$. Then $x_{23} = x_{13} \in A_{23}$, so by exactness of the second row there is an $x_{22} \in A_{22}$ whose image is x_{23}. We have $x_{33} = 0$ by exactness of the third column, so by exactness of the third row x_{32} has a unique preimage $x_{31} \in A_{31}$ (i.e. $x_{31} = x_{32}$ if we regard $A_{31} \subseteq A_{32}$). Let x_{41} be the image of x_{31} in A_{41}, and set $d(x_{13}) = x_{41}$.

(ii) To show that d is well defined, we must show it does not depend on the choice of x_{22}. If y_{22} is another preimage of x_{23} in A_{22}, then $y_{22} - x_{22}$ has a preimage z_{21} in A_{21} by exactness. The elements x_{31} and y_{31} differ in A_{31} by the image z_{31} of z_{21}, i.e. by an element in the image of A_{21}; hence their images in A_{41} agree. Thus d is well defined.

(iii) Once d is known to be well defined, it is routine to check that it is a homomorphism, and the argument is left to the reader (it is just a matter of choosing the x_{22} properly).

(iv): To show that the sequence is exact at A_{13}, first let $x_{12} \in A_{12}$. Then $x_{32} = x_{31} = 0$ by exactness of the second column, so $x_{41} = 0$. Conversely, if $x_{13} \in A_{13}$ with $d(x_{13}) = 0$, there is a preimage y_{21} of $x_{31} = x_{32}$ in

A_{21}. Since $y_{32} = x_{32}$, $z_{12} = x_{22} - y_{22} \in A_{12}$ by exactness of the second column. We have $z_{23} = x_{23}$ since $y_{23} = 0$ by exactness of the third row, so $z_{13} = z_{23} = x_{23} = x_{13}$ and x_{13} is in the image of A_{12}.

(v): To show the sequence is exact at A_{41}, let $x_{13} \in A_{13}$. Then x_{32} is the image of x_{22}, hence in the image of A_{22} in A_{32} and $x_{42} = 0$. Conversely, suppose x_{41} satisfies $x_{42} = 0$. If x_{31} is a preimage of x_{41}, this means precisely that x_{32} has a preimage $x_{22} \in A_{22}$. We have $x_{33} = 0$ by exactness of the third row, so x_{23} has a preimage $x_{13} \in A_{13}$ by exactness of the third column, and $d(x_{13}) = x_{41}$. 👍

XV.3.1.12. There is one part of the theorem not yet proved: naturality. This means: suppose we also have another diagram

$$\begin{array}{ccccccccc}
 & & 0 & & 0 & & 0 & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & B_{11} & \longrightarrow & B_{12} & \longrightarrow & B_{13} & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & B_{21} & \longrightarrow & B_{22} & \longrightarrow & B_{23} & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & B_{31} & \longrightarrow & B_{32} & \longrightarrow & B_{33} & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & B_{41} & \longrightarrow & B_{42} & \longrightarrow & B_{43} & & \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

and $\gamma_{jk} : A_{jk} \to B_{jk}$ for $1 \leq j, k \leq 4$ making a three-dimensional commutative diagram, and d and d' are the connecting maps from the two diagrams. Then

$$\begin{array}{ccccccccccc}
A_{11} & \longrightarrow & A_{12} & \longrightarrow & A_{13} & \overset{d}{\longrightarrow} & A_{41} & \longrightarrow & A_{42} & \longrightarrow & A_{43} \\
\downarrow{\scriptstyle \gamma_{11}} & & \downarrow{\scriptstyle \gamma_{12}} & & \downarrow{\scriptstyle \gamma_{13}} & & \downarrow{\scriptstyle \gamma_{41}} & & \downarrow{\scriptstyle \gamma_{42}} & & \downarrow{\scriptstyle \gamma_{43}} \\
B_{11} & \longrightarrow & B_{12} & \longrightarrow & B_{13} & \overset{d'}{\longrightarrow} & B_{41} & \longrightarrow & B_{42} & \longrightarrow & B_{43}
\end{array}$$

commutes (i.e. the center square commutes). The proof is easy and similar to the proof of (iii) above, and is left to the reader.

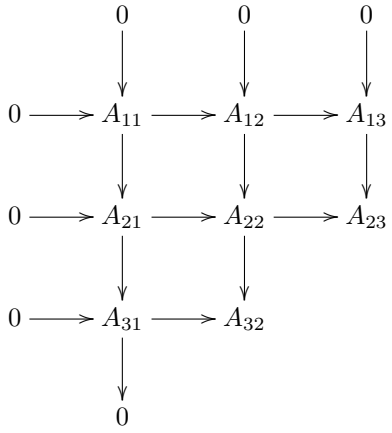
XV.3.2. Chain Complexes

XV.3.3. Exercises

XV.3.3.1. Use a diagram chase to prove the dual Basic Five Lemma XV.3.1.3..

XV.3.3.2. (a) Use a diagram chase to prove the following:

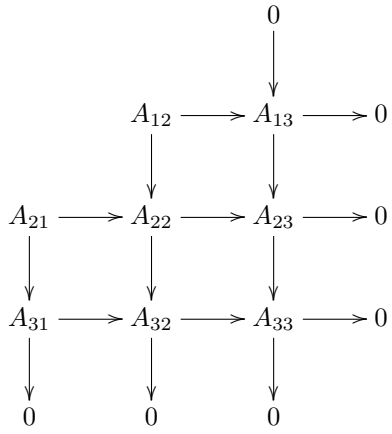
THEOREM. [BASIC NINE LEMMA] Let R be a ring, and let



be a commutative diagram of R-modules, with exact columns and exact middle row. Then the first row is exact if and only if the third row is exact.

(b) Prove the dual version either by duality or by a similar but somewhat trickier diagram chase:

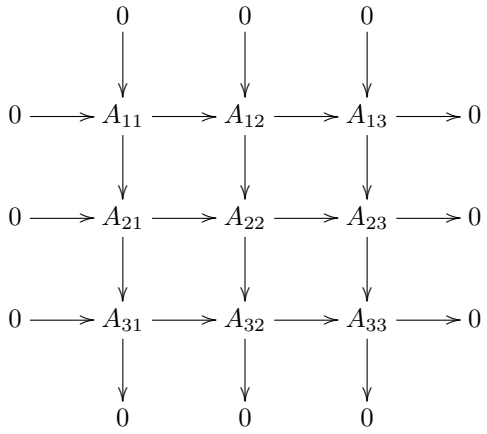
THEOREM. [BASIC NINE LEMMA, DUAL VERSION] Let R be a ring, and let



be a commutative diagram of R-modules, with exact columns and exact middle row. Then the first row is exact if and only if the third row is exact.

(c) Put the two together to obtain the common version of the Nine Lemma ([Fre64] suggests that the *Three-by-three Lemma* would be a more appropriate name):

Corollary. [Common Nine Lemma] Let R be a ring, and let



be a commutative diagram of R-modules, with exact columns and exact middle row. Then the first row is exact if and only if the third row is exact.

XV.3.3.3. Show that all the versions of the Five Lemma, the Nine Lemma, and the Snake Lemma are valid in an arbitrary abelian category (), with "injective" and "surjective" replaced by "monomorphism" and "epimorphism" respectively.

XV.3.3.4. The following exercise appears in [Lan02]:

Exercise. Take any book on homological algebra, and prove all the theorems without looking at the proofs given in that book.

XV.4. Cohomology

XV.4.1. General Cohomology

XV.4.2. Cup and Cap Products

XV.4.3. Čech Cohomology

XV.4.4. De Rham Cohomology

XV.4.5. Poincaré Duality

XV.4.6. K-Theory

Chapter XVI

Measurable Spaces and Subset Structures

In this chapter, we will set up the structures on which measures can be naturally defined. While these structures are of interest in their own right, the reader should not lose sight of the primary goal of this work: developing a good notion of measure and integration. In this context, this chapter should be regarded as a set of elaborate and technical preliminary work necessary to obtain a good theory of measure. Strictly speaking, not all of the material of this chapter will be necessary for basic measure theory, but it makes sense to give a reasonably complete study of these topics all at once.

The set structures discussed in this chapter should be compared to topologies, another very important kind of set structure discussed in Chapter (). Although the various kinds of set structures have properties making them useful in different contexts, all are of the same general nature and have many features in common.

XVI.1. Algebras and σ-Algebras of Sets

XVI.1.1. Set Operations

In this subsection, we review standard operations on subsets of a set X. Throughout the section, X will be a fixed set which we will regard as our "universe."

XVI.1.1.1. We first recall basic set operations on subsets of X. If $A, B \subseteq X$, then

The *union* $A \cup B$ is $\{x \in X : x \in A \text{ or } x \in B\}$.

The *intersection* $A \cap B$ is $\{x \in X : x \in A \text{ and } x \in B\}$.

The *complement* of A is $A^c = \{x \in X : x \notin A\}$.

The *difference* $A \setminus B$ is $A \cap B^c = \{x \in A : x \notin B\}$.

The *symmetric difference* $A \triangle B$ is $\{x \in X : x \in A \text{ or } x \in B \text{ but not both}\} = (A \setminus B) \cup (B \setminus A)$.

More generally, if $\{A_i : i \in I\}$ is an indexed collection of subsets of X, then

$$\cup_{i \in I} A_i = \{x \in X : x \in A_i \text{ for some } i \in I\}$$

$$\cap_{i \in I} A_i = \{x \in X : x \in A_i \text{ for all } i \in I\}$$

Limits of Sequences of Sets

XVI.1.1.2. Let $(A_1, A_2, \dots)$ be a sequence of subsets of X. Define

$$\limsup_{n \to \infty} A_n = \bigcap_n \bigcup_{k \geq n} A_k$$

$$\liminf_{n \to \infty} A_n = \bigcup_n \bigcap_{k \geq n} A_k$$

(the $n \to \infty$ is often notationally omitted).

Note that a set-theoretic lim sup or lim inf has, at least on the surface, nothing to do with the notion of the lim sup or lim inf of a sequence of numbers (); the terminology is just an analogy. (It is actually somewhat more than an analogy, though – see XVI.1.1.9.(iv).)

XVI.1.1.3. Proposition. Let $(A_1, A_2, \dots)$ be a sequence of subsets of X.

(i) $\limsup A_n$ is the set of all $x \in X$ which are in infinitely many A_n (we say such an x is *frequently in* A_n).

(ii) $\liminf A_n$ is the set of all $x \in X$ which are in all but (at most) finitely many A_n, i.e. which are in A_n for all sufficiently large n (we say such an x is *eventually in* A_n).

In particular, $\liminf A_n \subseteq \limsup A_n$.

The proof is a good exercise left to the reader.

XVI.1.1.4. De Morgan's Laws (or direct inspections) show:

$$(\limsup A_n)^c = \liminf(A_n^c)$$

$$(\liminf A_n)^c = \limsup(A_n^c)$$

XVI.1.1.5. Simple examples show that for a general sequence (A_n) of subsets of a set X, $\liminf A_n$ is "usually" a proper subset of $\limsup A_n$. For example, if $A, B \subseteq X$, and $A_n = A$ for n odd and $A_n = B$ for n even, then

$$\limsup A_n = A \cup B, \quad \liminf(A_n) = A \cap B \ .$$

But it can happen that the sets are equal; if so, we say the sequence (A_n) of subsets *converges*, and we write

$$\lim_{n\to\infty} A_n = \limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n \ .$$

The sequence (A_n) converges if and only if every $x \in X$ is either eventually in A_n or eventually in A_n^c.

There are situations where sequences automatically converge:

XVI.1.1.6. PROPOSITION. Let (A_n) be a sequence of subsets of X.
(i) If (A_n) is increasing, i.e. $A_n \subseteq A_{n+1}$ for all n, then (A_n) converges, and

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n \ .$$

(ii) If (A_n) is decreasing, i.e. $A_n \supseteq A_{n+1}$ for all n, then (A_n) converges, and

$$\lim_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} A_n \ .$$

(iii) If the A_n are pairwise disjoint, then (A_n) converges, and $\lim A_n = \emptyset$.
(iv) If (A_n) converges, then (A_n^c) converges, and $\lim(A_n^c) = (\lim A_n)^c$.

Indicator Functions

XVI.1.1.7. DEFINITION. Let A be a subset of X. The *indicator function* of A is the function $\chi_A : X \to [0,1]$ defined by

$$\chi_A(x) = \begin{cases} 1 & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases} \ .$$

The term *indicator function* is standard in probability theory; the name *characteristic function* is more standard in set theory and analysis. But the term "characteristic function" is also used to mean something else (), so we will standardize on "indicator function."

XVI.1.1.8. Conversely, any function f from X to $\{0,1\}$ is an indicator function: $f = \chi_A$, where $A = \{x \in X : f(x) = 1\}$. Thus $\mathcal{P}(X)$ can be identified with the set of functions from X to $\{0,1\}$, often called 2^X, cf. (). (More generally, a function $f : X \to [0,1]$ can be regarded as a "fuzzy subset" of X: if $x \in X$, $f(x)$ is the probability that x belongs to the corresponding subset.)

Set operations correspond to algebraic operations on the indicator functions:

XVI.1.1.9. Proposition. (i) Let A and B be subsets of X. Then

$$\chi_{A\cup B} = \chi_A \vee \chi_B = \max(\chi_A, \chi_B) \ .$$

$$\chi_{A\cap B} = \chi_A\chi_B = \min(\chi_A, \chi_B) \ .$$

(ii) Let $\{A_i : i \in I\}$ be a collection of subsets of X. Then

$$\chi_{\cup_i A_i} = \max_i \chi_{A_i} \ .$$

$$\chi_{\cap_i A_i} = \min_i \chi_{A_i} \ .$$

(iii) If $A \subseteq B \subseteq X$, then $\chi_{B\setminus A} = \chi_B - \chi_A$. In particular, $\chi_{A^c} = 1 - \chi_A$.
(iv) If (A_n) is a sequence of subsets of X, then

$$\chi_{\limsup A_n} = \limsup \chi_{A_n} \ .$$

$$\chi_{\liminf A_n} = \liminf \chi_{A_n} \ .$$

In particular, if (A_n) converges, then (χ_{A_n}) converges pointwise to $\chi_{\lim A_n}$. Conversely, if (χ_{A_n}) converges pointwise, then (A_n) converges.

XVI.1.2. Rings, Semirings, and Semialgebras

In this subsection, we will discuss various types of collections of subsets of a set X which are important to a greater or lesser extent in measure theory. Throughout the section, X will be a fixed set which we will regard as our "universe."

XVI.1.2.1. Definition. A nonempty collection $\mathcal{S}$ of subsets of X is a *semiring* if

(i) $\mathcal{S}$ is closed under finite intersections.

(ii) If $A, B \in \mathcal{S}$, then $A \setminus B$ is a finite disjoint union of sets in $\mathcal{S}$.

A semiring $\mathcal{S}$ for which $X \in \mathcal{S}$ is a *semialgebra*. A semiring $\mathcal{S}$ with the property that X is a countable union of sets in $\mathcal{S}$ is *of countable type*. A semialgebra is a semiring of countable type.

Taking $B = A$ in (ii), it follows that $\emptyset$ is in any semiring.

The definition of a semiring may look strange at first, but many important collections of subsets which naturally arise in analysis turn out to be semirings, and on the other hand semirings have enough structure that important constructions can be done with them.

XVI.1.2.2. EXAMPLES. (i) Here is perhaps the most important example. Let $\mathcal{I}$ be the set of intervals in $\mathbb{R}$, either open, closed, or half-open; include in $\mathcal{I}$ the "degenerate" intervals $\emptyset$ and singleton sets $\{x\}$ (intervals of length 0). $\mathcal{I}$ is closed under finite intersections (in fact under arbitrary intersections), and the complement of one interval in another is either an interval or the disjoint union of two intervals, so $\mathcal{I}$ is a semiring. In fact, $\mathcal{I}$ is a semialgebra.
(ii) As a variation of (i), let $\mathcal{I}_f$ be the set of bounded intervals in $\mathbb{R}$, including degenerate intervals. By the same argument, $\mathcal{I}_f$ is a semiring. It is not a semialgebra, but it is a semiring of countable type.
(iii) Similarly, if $\delta > 0$, let $\mathcal{I}_\delta$ be the set of intervals in $\mathbb{R}$ of length $\leq \delta$, again including degenerate intervals. By the same argument, using the Archimedean property of $\mathbb{R}$, $\mathcal{I}_\delta$ is a semiring of countable type.
(iv) If $\mathcal{C}$ is any semiring (e.g. an algebra or σ-algebra) on X, and $\{A_i : i \in I\}$ is a nonempty collection of sets in $\mathcal{C}$, then $\mathcal{S} = \{S \in \mathcal{C} : S \subseteq A_i \text{ for some } i \in I\}$ is a semiring. More generally, if P is a property of sets inherited by all subsets, then the set of $C \in \mathcal{C}$ with property P form a semiring.

XVI.1.2.3. The intersection of two semirings is not necessarily a semiring, and hence if $\mathcal{C}$ is a collection of subsets of X, there is not necessarily a smallest semiring containing $\mathcal{C}$. But there is a standard construction of a semiring containing $\mathcal{C}$, which is particularly important in probability:

XVI.1.2.4. PROPOSITION. Let $\mathcal{C}$ be any nonempty collection of subsets of X, and let

$$\mathcal{D} = \{A \setminus B : A, B \in \mathcal{C}\} .$$

Let $\mathcal{S}$ be the set of finite intersections of sets in $\mathcal{C} \cup \mathcal{D}$. Then $\mathcal{S}$ is a semiring containing $\mathcal{C}$.

Note that $\emptyset \in \mathcal{D}$ since it is $A \setminus A$ for any $A \in \mathcal{C}$; if $\emptyset \in \mathcal{C}$, then $\mathcal{C} \subseteq \mathcal{D}$ since $A = A \setminus \emptyset$ for any A.
PROOF: *(Outline)* This is completely routine, but tedious to write out carefully; see XVI.1.7.1.. The proof consists of repeated applications of the following facts for subsets A, B, C of X:

$(A \cap B) \setminus C = (A \setminus C) \cap (B \setminus C)$.

$(A \setminus B) \setminus C = (A \setminus B) \cap (A \setminus C)$.

$A \setminus (B \setminus C)$ is the disjoint union of $A \setminus B$ and $A \cap B \cap C$.

$A \setminus (B \cap C)$ is the disjoint union of $A \setminus B$ and $(A \cap B) \setminus C$.

The details of the proof are left to the reader, who should at least verify the above facts. 👍

A similar proof shows the following:

XVI.1.2.5. PROPOSITION. Let $\mathcal{C}$ be a nonempty collection of subsets of X with the property that $A \in \mathcal{C}$ implies $A^c \in \mathcal{C}$. Let $\mathcal{S}$ be the collection of finite intersections of sets in $\mathcal{C}$. Then $\mathcal{S}$ is a semialgebra.

Collections with more structure behave more nicely and are of central importance in measure theory:

XVI.1.2.6. DEFINITION. A *ring* of subsets of a set X is a nonempty collection of subsets of X which is closed under finite unions and relative complements.

XVI.1.2.7. PROPOSITION. A ring is closed under symmetric difference and finite intersections; in particular, a ring is a semiring.

PROOF: If $A, B \subseteq X$, then $A \triangle B = (A \setminus B) \cup (B \setminus A)$ and $A \cap B = (A \cup B) \setminus (A \triangle B)$. 👍

XVI.1.2.8. The converse also holds: a nonempty collection of subsets of X closed under finite intersections and symmetric differences is a ring, because of the relations

$$A \cup B = (A \triangle B) \triangle (A \cap B), \quad A \setminus B = A \triangle (A \cap B)$$

for subsets of X. This explains the use of the term "ring": the operations $A + B = A \triangle B$ and $AB = A \cap B$ make $\mathcal{P}(X)$ into a ring in the algebraic sense of III.3.6.3. (in fact, a special kind of ring called a *Boolean algebra*); a ring of subsets is precisely a subring of $\mathcal{P}(X)$ with these operations. This use of *ring* (of subsets) is apparently due to HALMOS [Hal50].

In many references, especially older ones, $A \cap B$ is written AB, and in many of these $A \cup B$ is written $A + B$ and called the *sum* of A and B. But $\mathcal{P}(X)$ is not a ring under these algebraic operations. Also, for sets of numbers or subsets of a topological group, these notations are commonly used to denote something quite different (). Thus we will not use the notation AB or $A + B$ for set-theoretic operations. ✋

XVI.1.2.9. Any intersection of rings is a ring (the intersection is nonempty since $\emptyset$ is in every ring). Thus if $\mathcal{C}$ is any collection of subsets of X, there is a smallest ring of subsets containing $\mathcal{C}$, called the *ring generated by* $\mathcal{C}$ (there is at least one ring containing $\mathcal{C}$, namely $\mathcal{P}(X)$). The ring generated by $\emptyset$ is $\{\emptyset\}$.

XVI.1.2.10. PROPOSITION. Let $\mathcal{S}$ be a semiring. The ring generated by $\mathcal{S}$ is precisely the set of finite disjoint unions of sets in $\mathcal{S}$.

PROOF: Let $\mathcal{A}$ be the set of finite disjoint unions of sets in $\mathcal{S}$. We show that $\mathcal{A}$ is closed under finite intersections, then finite unions and relative complements.

For finite intersections, let A be the disjoint union of $A_1, \ldots, A_n \in \mathcal{S}$ and B the disjoint union of $B_1, \ldots, B_m \in \mathcal{S}$. Then $A \cap B$ is the disjoint union of

$$\{A_j \cap B_k : 1 \leq j \leq n, 1 \leq k \leq m\} .$$

For finite unions, it suffices to show that if $A_1, \ldots, A_n, B \in \mathcal{S}$ with $A_1, \ldots, A_n$ pairwise disjoint, then $A_1 \cup \cdots \cup A_n \cup B$ can be written as a finite disjoint union of sets in $\mathcal{S}$. For each j, $B \setminus A_j$ can be written as a finite disjoint union of sets in $\mathcal{S}$. Thus by the previous argument

$$B \setminus (A_1 \cup \cdots A_n) = (B \setminus A_1) \cap \cdots \cap (B \setminus A_n)$$

can be written as a finite disjoint union of sets in $\mathcal{S}$; the union of these sets and the A_j is $A_1 \cup \cdots \cup A_n \cup B$.

Now let $A, B \in \mathcal{A}$, with A the disjoint union of $A_1, \ldots, A_n \in \mathcal{S}$ and B the disjoint union of $B_1, \ldots, B_m \in \mathcal{S}$. We have

$$A \setminus B = \bigcup_{j=1}^{n} \bigcap_{k=1}^{m} (A_j \setminus B_k)$$

and each $A_j \setminus B_k$ is a finite disjoint union of sets in $\mathcal{S}$, hence in $\mathcal{A}$. So $A \setminus B$ is in $\mathcal{A}$ by the previous arguments. 👍

XVI.1.2.11. COROLLARY. A semiring which is closed under finite disjoint unions is a ring.

XVI.1.2.12. COROLLARY. Let $\mathcal{S}$ be a semiring, A, A_1, ..., A_n sets in $\mathcal{S}$ with $A_j \subseteq A$ for all j and the A_j pairwise disjoint. Then there are finitely many sets $B_1, \ldots, B_m \in \mathcal{S}$ such that $\{A_1, \ldots, A_n, B_1, \ldots, B_m\}$ are pairwise disjoint and

$$A = \left(\bigcup_{j=1}^{n} A_j \right) \cup \left(\bigcup_{k=1}^{m} B_k \right) .$$

PROOF: Let $\mathcal{A}$ be the ring generated by $\mathcal{S}$. If $D = \cup_{j=1}^{n} A_j$, then $D \in \mathcal{A}$, so $(A \setminus D) \in \mathcal{A}$, i.e. $A \setminus D$ can be written as a finite disjoint union of sets in $\mathcal{S}$. 👍

XVI.1.2.13. PROPOSITION. Let $\mathcal{S}$ be a semiring of subsets of a set X, and let A, A_1, A_2, ... be sets in $\mathcal{S}$. Suppose $A \subseteq \cup_{n=1}^{\infty} A_n$. Then there is a sequence (B_k) of pairwise disjoint sets in $\mathcal{S}$ with each B_k contained in a unique A_n, only finitely many B_k in each A_n, such that $A = \cup_k B_k$.

PROOF: Set $B_1 = A \cap A_1 \in \mathcal{S}$. Write $A \setminus A_1 = C_{11} \cup \cdots \cup C_{1m_1}$ with the C_{1j} pairwise disjoint sets in $\mathcal{S}$. Set $B_2 = A_2 \cap C_{11}$, ..., $B_{m_1+1} = A_2 \cap C_{1m_1}$. Then the B_k $(1 \leq k \leq m_1 + 1)$ are in $\mathcal{S}$ and are pairwise disjoint, $B_1 \subseteq A_1$, $B_k \subseteq A_2 \setminus A_1$ for $2 \leq k \leq m_1 + 1$, and

$$\bigcup_{k=1}^{m_1+1} B_k = A \cap (A_1 \cup A_2) .$$

Write $A \setminus A_2 = C_{21} \cup \cdots \cup C_{2m_2}$ with the C_{2j} pairwise disjoint sets in $\mathcal{S}$. The next set of B_k's (for $m_1 + 2 \leq k \leq m_1 + m_1 m_2 + 1$) are of the form $A_3 \cap C_{1i} \cap C_{2j}$ for $1 \leq i \leq m_1$ and $1 \leq j \leq m_2$. The new B_k are pairwise disjoint and contained in $A_3 \setminus (A_1 \cup A_2)$, and

$$\bigcup_{k=1}^{m_1+m_1m_2+1} B_k = A \cap (A_1 \cup A_2 \cup A_3) .$$

The process can be continued inductively, although the notation becomes increasingly complicated. At the end we have

$$\bigcup_{k=1}^{\infty} B_k = A \cap \left(\bigcup_{n=1}^{\infty} A_n \right) = A .$$

👍

XVI.1.3. Algebras and σ-Algebras

For measure theory purposes, the most important set structures are algebras and especially σ-algebras.

XVI.1.3.1. DEFINITION. Let X be a set. An *algebra* of subsets of X is a set $\mathcal{A}$ of subsets of X satisfying:

(i) $\emptyset, X \in \mathcal{A}$.

(ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

(iii) $\mathcal{A}$ is closed under finite unions.

$\mathcal{A}$ is a *σ-algebra* of subsets of X if it is an algebra and closed under countable unions. Algebras and σ-algebras are also sometimes called *fields* and *σ-fields* respectively, especially in probability.

Note that the set X and the phrase "of subsets of X" is crucial to specify (e.g. if $Y \subseteq X$, then $\mathcal{P}(Y)$ is a σ-algebra of subsets of Y but not of X); thus a statement like "Let $\mathcal{A}$ be a σ-algebra" is technically incorrect or at least incomplete unless the X is understood (as it often is).

XVI.1.3.2. Since $\emptyset$ is in any algebra or σ-algebra, the requirement that a σ-algebra be closed under finite unions as well as countable unions is redundant, i.e. a collection $\mathcal{A} \subseteq \mathcal{P}(X)$ is a σ-algebra if and only if

(i) $\emptyset, X \in \mathcal{A}$.

(ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

(iii) $\mathcal{A}$ is closed under countable unions.

By de Morgan's laws, a collection $\mathcal{A} \subseteq \mathcal{P}(X)$ is an algebra [resp. σ-algebra] if and only if

(i) $\emptyset, X \in \mathcal{A}$.

(ii) If $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$.

(iii) $\mathcal{A}$ is closed under finite [resp. countable] intersections.

In particular, an algebra is closed under finite intersections, and a σ-algebra under countable intersections (including finite intersections).

Requirement (i) can be relaxed (in the presence of (ii) and (iii)) to only require that $\mathcal{A}$ be nonempty, since if $A \in \mathcal{A}$, then $A \cup A^c = X \in \mathcal{A}$ and $A \cap A^c = \emptyset \in \mathcal{A}$.

XVI.1.3.3. An algebra of subsets of X is a ring of subsets of X; a ring $\mathcal{R}$ of subsets of X is an algebra if and only if $X \in \mathcal{R}$ (see XVI.1.4.7. for a precise version).

XVI.1.3.4. EXAMPLES. (i) $\mathcal{P}(X)$ and $\{\emptyset, X\}$ are σ-algebras of subsets of X, and are the largest and smallest σ-algebras on X.

(ii) Let X be an infinite set. The collection of all finite subsets of X and their complements (*cofinite sets*) is an algebra of subsets of X which is not a σ-algebra.

(iii) Let X be an uncountable set. The collection of all countable subsets of X and their complements (*cocountable sets*) is a σ-algebra of subsets of X.

There are many much more interesting examples which will take some effort to define.

Generated Algebras and σ-Algebras

XVI.1.3.5. The intersection of any family of algebras [resp. σ-algebras] of subsets of a set X is an algebra [resp. σ-algebra] of subsets of X. So if $\mathcal{S}$ is any collection of subsets of a set X, there is a smallest algebra [resp. σ-algebra] containing $\mathcal{S}$, the intersection of all algebras [resp. σ-algebras] containing $\mathcal{S}$ (there is at least one, $\mathcal{P}(X)$), called the *algebra [resp. σ-algebra] generated by* $\mathcal{S}$.

If $\mathcal{R}$ is a ring of subsets of X, there is a simple description of the algebra generated by $\mathcal{R}$:

XVI.1.3.6. Proposition. Let $\mathcal{R}$ be a ring of subsets of a set X, and $\mathcal{A}$ the algebra of subsets of X generated by $\mathcal{R}$. Then $\mathcal{A}$ consists precisely of the sets in $\mathcal{R}$ and all sets of the form $A \cup B^c$ for $A, B \in \mathcal{R}$.

Proof: Let $\mathcal{B}$ be this collection of subsets of X. Then $\mathcal{R} \subseteq \mathcal{B} \subseteq \mathcal{A}$, so we need only show that $\mathcal{B}$ is an algebra. Since $\emptyset \in \mathcal{R}$, $\emptyset \cup \emptyset^c = X \in \mathcal{B}$. If $A \in \mathcal{R}$, then $\emptyset \cup A^c = A^c \in \mathcal{B}$, and if $A, B \in \mathcal{R}$, then

$$(A \cup B^c)^c = A^c \cap B = B \setminus A \in \mathcal{R} \subseteq \mathcal{B}$$

so $\mathcal{B}$ is closed under complements. Since $\mathcal{R}$ is closed under finite unions and intersections, if $A, B, C \in \mathcal{R}$, then $A \cup (B \cup C^c) = (A \cup B) \cup C^c \in \mathcal{B}$, and if $A, B, C, D \in \mathcal{R}$, then

$$(A \cup B^c) \cup (C \cup D^c) = (A \cup C) \cup (B^c \cup D^c) = (A \cup C) \cup (B \cap D)^c \in \mathcal{B}$$

do $\mathcal{B}$ is closed under finite unions. 👍

See XVI.1.4.7. for a consequence, and XVI.1.3.11. for a generalization.

XVI.1.3.7. The σ-algebra generated by a collection of subsets, even a ring, is much more complicated to describe. To do this, if $\mathcal{S} \subseteq \mathcal{P}(X)$, use the following notation: $\tilde{S} = \{A^c : A \in \mathcal{S}\}$; $\mathcal{S}_\sigma$ denotes the set of countable unions of sets in $\mathcal{S}$; and $\mathcal{S}_\delta$ the set of countable intersections of sets in $\mathcal{S}$. (Note that $\mathcal{S} \subseteq \mathcal{S}_\sigma$ and $\mathcal{S} \subseteq \mathcal{S}_\delta$ by unions or intersections of constant sequences.) ✋

Fix $\mathcal{S} \subseteq \mathcal{P}(X)$. For countable ordinals α, define $\mathcal{S}^\sigma_\alpha$ recursively as follows. Set $\mathcal{S}^\sigma_0 = \mathcal{S} \cup \{\emptyset\}$ (it is only necessary for the construction to include $\emptyset$ if $\mathcal{S}$ is empty). Suppose $\mathcal{S}^\sigma_\alpha$ has been defined. Set

$$\mathcal{S}^\sigma_{\alpha+1} = (S^\sigma_\alpha \cup \widetilde{\mathcal{S}^\sigma_\alpha})_\sigma \ .$$

If α is a limit ordinal, set $\mathcal{S}^\sigma_\alpha = \cup_{\beta<\alpha}\mathcal{S}^\sigma_\beta$. Similarly, set $\mathcal{S}^\delta_0 = \mathcal{S} \cup \{\emptyset\}$. Suppose $\mathcal{S}^\delta_\alpha$ has been defined. Set

$$\mathcal{S}^\delta_{\alpha+1} = (S^\delta_\alpha \cup \widetilde{\mathcal{S}^\delta_\alpha})_\delta \ .$$

If α is a limit ordinal, set $\mathcal{S}^\delta_\alpha = \cup_{\beta<\alpha}\mathcal{S}^\delta_\beta$. We clearly have $\mathcal{S}^\sigma_\beta \subseteq \mathcal{S}^\sigma_\alpha$ and $\mathcal{S}^\delta_\beta \subseteq \mathcal{S}^\delta_\alpha$ if $\beta < \alpha$, so these sets are nested.

XVI.1.3.8. Theorem. Let $\mathcal{S}$ be a collection of subsets of a set X, and $\mathcal{A}$ the σ-algebra of subsets of X generated by $\mathcal{S}$. Then

$$\mathcal{A} = \bigcup_{\alpha<\omega_1} \mathcal{S}^\sigma_\alpha = \bigcup_{\alpha<\omega_1} \mathcal{S}^\delta_\alpha$$

(where ω_1 is the first uncountable ordinal, i.e. the union is over all countable ordinals).

PROOF: Let $\mathcal{B} = \bigcup_{\alpha<\omega_1} \mathcal{S}^\sigma_\alpha$. Then $\mathcal{S} \subseteq \mathcal{S}^\sigma_0 \subseteq \mathcal{B}$, and it is easily seen recursively that any σ-algebra containing $\mathcal{S}$ must contain $\mathcal{S}^\sigma_\alpha$ for any α; thus $\mathcal{B} \subseteq \mathcal{A}$. Thus it suffices to show that $\mathcal{B}$ is a σ-algebra. If $A \in \mathcal{S}^\sigma_\alpha$, then $A^c \in \mathcal{S}^\sigma_{\alpha+1}$, so $\mathcal{B}$ is closed under complements, and since $\emptyset \in \mathcal{S}^\sigma_0 \subseteq \mathcal{B}$, $X \in \mathcal{S}^\sigma_1 \subseteq \mathcal{B}$. If (A_n) is a sequence of sets in $\mathcal{B}$, say $A_n \in \mathcal{S}^\sigma_{\alpha_n}$, there is a countable α such that $\alpha_n \leq \alpha$ for all n; thus $A_n \in \mathcal{S}^\sigma_\alpha$ for all n, and so $\cup_n A_n \in \mathcal{S}^\sigma_{\alpha+1} \subseteq \mathcal{B}$. Thus $\mathcal{B}$ is closed under countable unions.

A similar proof shows that $\mathcal{A} = \bigcup_{\alpha<\omega_1} \mathcal{S}^\delta_\alpha$, or just note that by de Morgan, $\mathcal{S}^\delta_\alpha \subseteq \mathcal{S}^\sigma_{\alpha+1}$ and $\mathcal{S}^\sigma_\alpha \subseteq \mathcal{S}^\delta_{\alpha+1}$ for all α. 👍

XVI.1.3.9. In some situations, the transfinite sequences $\mathcal{S}^\sigma_\alpha$ and $\mathcal{S}^\delta_\alpha$ stabilize at some $\alpha < \omega_1$, even occasionally for some finite α, but in general, and even in the most important cases (cf. ()), they do not stabilize and the construction must be carried out uncountably.

XVI.1.3.10. There is a similar but finite process to explicitly construct the algebra generated by a collection $\mathcal{S}$ of subsets of a set X, involving both unions and intersections. Set $\mathcal{S}_0 = \mathcal{S} \cup \tilde{\mathcal{S}} \cup \{\emptyset, X\}$, let $\mathcal{S}_1$ be the set of finite intersections of sets in $\mathcal{S}_0$, and $\mathcal{S}_2$ the set of finite unions of sets in $\mathcal{S}_1$.

XVI.1.3.11. PROPOSITION. $\mathcal{S}_2$ is the algebra $\mathcal{A}$ of subsets of X generated by $\mathcal{S}$.

PROOF: We have $\mathcal{S} \subseteq \mathcal{S}_0 \subseteq \mathcal{S}_1 \subseteq \mathcal{S}_2 \subseteq \mathcal{A}$ and $\emptyset, X \in \mathcal{S}_0 \subseteq \mathcal{S}_2$. $\mathcal{S}_1$ is closed under finite intersections and $\mathcal{S}_2$ closed under finite unions. $\mathcal{S}_2$ is closed under finite intersections too: if $A, B \in \mathcal{S}_2$, $A = \cup_{j=1}^n A_j$, $B = \cup_{k=1}^m B_k$ with $A_j, B_k \in \mathcal{S}_1$, then

$$A \cap B = \bigcup_{j=1}^{n} \bigcup_{k=1}^{m} (A_j \cap B_k) \in \mathcal{S}_2 \ .$$

We now show $\mathcal{S}_2$ is closed under complements. Let $A \in \mathcal{S}_2$. Then $A = \cup_{j=1}^n A_j$, where $A_j \in \mathcal{S}_1$; $A_j = \cap_{k=1}^{m_j} A_{jk}$ for $A_{jk} \in \mathcal{S}_0$. We have

$$A^c = \bigcap_{j=1}^{n} \bigcup_{k=1}^{m_j} A^c_{jk} \ .$$

Since $\mathcal{S}_0$ is closed under complements, $A^c_{jk} \in \mathcal{S}_0 \subseteq \mathcal{S}_2$ for all j, k; since $\mathcal{S}_2$ is closed under finite unions and intersections, $A^c \in \mathcal{S}_2$. Thus $\mathcal{S}_2$ is an algebra, so $\mathcal{A} \subseteq \mathcal{S}_2$. 👍

XVI.1.4. σ-Rings and δ-Rings

It is useful for some purposes to have two other types of structures intermediate between rings and σ-algebras.

XVI.1.4.1. DEFINITION. Let X be a set. A ring $\mathcal{D}$ of subsets of X is a *σ-ring* if it is closed under countable unions. $\mathcal{D}$ is a *δ-ring* if it is closed under countable intersections.

Note that a σ-algebra is a σ-ring, and a σ-ring is a σ-algebra if and only if it contains X. In fact, a σ-ring $\mathcal{D}$ of subsets of X can only fail to be a σ-algebra because there are subsets of X which are "too large" to be in $\mathcal{D}$. In fact, if $\mathcal{D}$ is a σ-ring or δ-ring of subsets of a set X, and $Y \in \mathcal{D}$, then $\mathcal{D}_Y = \{A \in \mathcal{D} : A \subseteq Y\}$ is a σ-algebra of subsets of Y.

XVI.1.4.2. PROPOSITION. Let X be a set. A ring $\mathcal{D}$ of subsets of X is a δ-ring if and only if whenever $\{A_n : n \in \mathbb{N}\}$ is a sequence in $\mathcal{D}$, and there is a $B \in \mathcal{D}$ with $A_n \subseteq B$ for all n, then $\cup_{n=1}^{\infty} A_n \in \mathcal{D}$.

PROOF: Suppose $\mathcal{D}$ is a δ-ring, and the A_n and B are as in the statement. For each n, set $B_n = B \setminus A_n$. Then $B_n \in \mathcal{D}$, so $\cap_{n=1}^{\infty} B_n \in \mathcal{D}$. Then

$$\bigcup_{n=1}^{\infty} A_n = B \setminus \left(\bigcap_{n=1}^{\infty} B_n \right) \in \mathcal{D} \ .$$

Conversely, if $\mathcal{D}$ has the property in the statement, and (B_n) is a sequence in $\mathcal{D}$, set $A_n = B_1 \setminus B_n$ for each n. We have $A_n \in \mathcal{D}$ and $A_n \subseteq B_1$ for each n, so $\cup_{n=1}^{\infty} A_n \in \mathcal{D}$. Then

$$\bigcap_{n=1}^{\infty} B_n = B_1 \setminus \left(\bigcup_{n=1}^{\infty} A_n \right) \in \mathcal{D} \ .$$

XVI.1.4.3. COROLLARY. A σ-ring is a δ-ring.

XVI.1.4.4. EXAMPLES. (i) Let X be an uncountable set. The collection $\mathcal{D}$ of countable subsets of X is a σ-ring.

(ii) Let X be a set, and $Y \subseteq X$. Any σ-ring (e.g. any σ-algebra) of subsets of Y is a σ-ring of subsets of X.

(iii) The collection of all bounded subsets of $\mathbb{R}$ is a δ-ring of subsets of $\mathbb{R}$ which is not a σ-ring. This, or slight modifications, is the prototype example.

(iv) Another (not too interesting) example of a δ-ring which is not a σ-ring: let X be an infinite set, and $\mathcal{D}$ the collection of finite subsets of X.

(v) If X is a locally compact Hausdorff space which is not σ-compact, the collection of subsets of X whose closure is σ-compact is a σ-ring.

(vi) If X is any set, then $\{\emptyset\}$ is a (not very interesting) σ-ring.

XVI.1.4.5. Any intersection of σ-rings [resp. δ-rings] of subsets of a set X is a σ-ring [resp. δ-ring] of subsets of X. So if $\mathcal{S}$ is any collection of subsets of X, there is a smallest σ-ring [resp. δ-ring] containing $\mathcal{S}$, called the *σ-ring [resp. δ-ring] generated by $\mathcal{S}$*; it is the intersection of all σ-rings [resp. δ-ring] containing $\mathcal{S}$ (there is at least one, $\mathcal{P}(X)$).

Hereditary Subcollections and a Characterization

XVI.1.4.6. Let X be a set, and $\mathcal{S}$ a collection of subsets of X (in applications $\mathcal{S}$ will usually be at least a ring, but this is not necessary for the definitions). A nonempty subcollection $\mathcal{H}$ of $\mathcal{S}$ is called *hereditary* if $A \in \mathcal{H}$, $B \in \mathcal{S}$, $B \subseteq A$ implies $B \in \mathcal{H}$. If $\mathcal{R}$ is another nonempty collection of subsets of X (in applications $\mathcal{R}$ will usually be a ring contained in $\mathcal{S}$, but this is not necessary for the definition), write

$$\mathcal{H}_{\mathcal{R}}(\mathcal{S}) = \{B \in \mathcal{S} : B \subseteq A \text{ for some } A \in \mathcal{R}\} .$$

Then it is easily checked that $\mathcal{H}_{\mathcal{R}}(\mathcal{S})$, if nonempty (automatic if $\mathcal{R} \subseteq \mathcal{S}$), is a hereditary subset of $\mathcal{S}$, and $\mathcal{R} \subseteq \mathcal{H}$ if $\mathcal{R} \subseteq \mathcal{S}$. (Every hereditary subcollection $\mathcal{H}$ of $\mathcal{S}$ is of this form: we have $\mathcal{H} = \mathcal{H}_{\mathcal{H}}(\mathcal{S})$.)

If $\mathcal{S}$ is a ring, and $\mathcal{R}$ is closed under finite unions (e.g. if $\mathcal{R}$ is also a ring), then $\mathcal{H}_{\mathcal{R}}(\mathcal{S})$ is also a ring: since $\emptyset \in \mathcal{S}$ and $\mathcal{R} \neq \emptyset$, $\emptyset \in \mathcal{H}_{\mathcal{R}}(\mathcal{S})$, so $\mathcal{H}_{\mathcal{R}}(\mathcal{S}) \neq \emptyset$; if $A, B \in \mathcal{H}_{\mathcal{R}}(\mathcal{S})$, then $A \subseteq D$ for some $D \in \mathcal{R}$, so $A \setminus B \subseteq A \subseteq D$ and $A \setminus B \in \mathcal{S}$, so $A \setminus B \in \mathcal{H}_{\mathcal{R}}(\mathcal{S})$; and if $A, B \in \mathcal{H}_{\mathcal{R}}(\mathcal{S})$, then $A \subseteq C$ and $B \subseteq D$ for some $C, D \in \mathcal{R}$, and therefore $A \cup B \in \mathcal{S}$ and $A \cup B \subseteq C \cup D \in \mathcal{R}$, so $A \cup B \in \mathcal{H}_{\mathcal{R}}(\mathcal{S})$, and thus $\mathcal{H}_{\mathcal{R}}(\mathcal{S})$ is closed under finite unions. If $\mathcal{S}$ is a δ-ring, so is $\mathcal{H}_{\mathcal{R}}(\mathcal{S})$. Similarly, if $\mathcal{R}$ is closed under countable unions and $\mathcal{S}$ is a σ-ring, then $\mathcal{H}_{\mathcal{R}}(\mathcal{S})$ is also a σ-ring.

If $\mathcal{R}$ is a ring, there is a nice connection with the algebra it generates:

XVI.1.4.7. Proposition. Let X be a set, $\mathcal{R}$ a ring of subsets of X, and $\mathcal{A}$ the algebra of subsets of X generated by $\mathcal{R}$. Then $\mathcal{R} = \mathcal{H}_{\mathcal{R}}(\mathcal{A})$.

Proof: We have $\mathcal{R} \subseteq \mathcal{H}_{\mathcal{R}}(\mathcal{A})$ since $\mathcal{R} \subseteq \mathcal{A}$. Conversely, we use the characterization in XVI.1.3.6.. Suppose $A \in \mathcal{H}_{\mathcal{R}}(\mathcal{A})$. Then either $A \in \mathcal{R}$, in which case there is nothing to prove, or $A = B \cup C^c$ for $B, C \in \mathcal{R}$, and there is a $D \in \mathcal{R}$ with $A \subseteq D$. Then $D \setminus C \in \mathcal{R}$, and $C^c \subseteq D$, so $C^c = D \setminus C \in \mathcal{R}$, and thus $A = B \cup (D \setminus C) \in \mathcal{R}$. Thus $\mathcal{H}_{\mathcal{R}}(\mathcal{A}) \subseteq \mathcal{R}$. 👍

There is a similar characterization of δ-rings (and, in particular, σ-rings):

XVI.1.4.8. Theorem. et X be a set, $\mathcal{D}$ a δ-ring of subsets of X, and $\mathcal{A}$ the σ-algebra of subsets of X generated by $\mathcal{D}$. Then $\mathcal{D} = \mathcal{H}_{\mathcal{D}}(\mathcal{A})$.

Proof: We have $\mathcal{D} \subseteq \mathcal{H}_{\mathcal{D}}(\mathcal{A})$ since $\mathcal{D} \subseteq \mathcal{A}$. Conversely, we use the construction in XVI.1.3.8.. It suffices to show by transfinite induction that (using the notation of XVI.1.3.8.) $\mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\alpha}) \subseteq \mathcal{D}$ for every $\alpha < \omega_1$. We have $\mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{0}) \subseteq \mathcal{D}^{\delta}_{0} = \mathcal{D}$. If α is a limit ordinal and $\mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\beta}) \subseteq \mathcal{D}$ for all $\beta < \alpha$, then clearly $\mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\alpha}) \subseteq \mathcal{D}$. Now suppose $\mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\alpha}) \subseteq \mathcal{D}$. Suppose $A \in \mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\alpha} \cup \widetilde{\mathcal{D}^{\delta}_{\alpha}})$. Then there is a $B \in \mathcal{D}$ with $A \subseteq B$, and $A \in \mathcal{D}^{\delta}_{\alpha} \cup \widetilde{\mathcal{D}^{\delta}_{\alpha}}$. If $A \in \mathcal{D}^{\delta}_{\alpha}$, then $A \in \mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\alpha}) \subseteq \mathcal{D}$. If $A \in \widetilde{\mathcal{D}^{\delta}_{\alpha}}$, then $A^c \in \mathcal{D}^{\delta}_{\alpha} \subseteq \mathcal{D}$. So $A = B \setminus A^c \in \mathcal{D}$. Finally, if $A \in \mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\alpha+1}) = \mathcal{H}_{\mathcal{D}}([\mathcal{D}^{\delta}_{\alpha} \cup \widetilde{\mathcal{D}^{\delta}_{\alpha}}]_{\delta})$, then there is a $B \in \mathcal{D}$ with $A \subseteq B$, and a sequence (A_n) in $\mathcal{D}^{\delta}_{\alpha} \cup \widetilde{\mathcal{D}^{\delta}_{\alpha}}$ with $A = \cap_n A_n$. Fix n. If $A_n \in \mathcal{D}^{\delta}_{\alpha}$, then $B \cap A_n \in \mathcal{D}^{\delta}_{\alpha}$ since $\mathcal{D} \subseteq \mathcal{D}^{\delta}_{\alpha}$ and $\mathcal{D}^{\delta}_{\alpha}$ is closed under finite (even countable) intersections. Thus $B \cap A_n \in \mathcal{H}_{\mathcal{D}}(\mathcal{D}^{\delta}_{\alpha}) \subseteq \mathcal{D}$. And if $A_n \in \widetilde{\mathcal{D}^{\delta}_{\alpha}}$, then $A_n^c \in \mathcal{D}^{\delta}_{\alpha}$, so $B \cap A_n^c \in \mathcal{D}$ by the same argument, and thus $B \cap A_n = B \setminus (B \cap A_n^c) \in \mathcal{D}$. Therefore $B \cap A_n \in \mathcal{D}$ for all n, so $A = \cap_n (B \cap A_n) \in \mathcal{D}$. This completes the inductive argument. 👍

XVI.1.4.9. **Caution:** Putting together XVI.1.4.8. and XVI.1.4.7., if $\mathcal{D}$ is a δ-ring, $\mathcal{A}$ the σ-algebra generated by $\mathcal{D}$, and $\mathcal{B}$ the algebra generated by $\mathcal{D}$, we have $\mathcal{H}_{\mathcal{D}}(\mathcal{A}) = \mathcal{H}_{\mathcal{D}}(\mathcal{B}) = \mathcal{D}$. However, this does *not* mean that $\mathcal{A} = \mathcal{B}$! ($\mathcal{B}$ will not be closed under countable unions in general.) But if $\mathcal{D}$ is a σ-ring, it is true that $\mathcal{A} = \mathcal{B}$: an argument nearly identical to the one in XVI.1.3.6. shows that $\mathcal{B}$ is closed under countable unions and is thus a σ-algebra (this argument gives a simpler proof of XVI.1.4.8. if $\mathcal{D}$ is a σ-ring).

Locally Measurable Sets

In measure theory, the sets in a specified collection of subsets of a fixed set, typically at least a δ-ring and normally a σ-algebra, are called "measurable" sets since a numerical size or measure is assigned to them. For general collections of subsets, there is a more general notion of "measurability" which is useful:

XVI.1.4.10. Definition. Let $\mathcal{S}$ be a collection of subsets of a set X. A subset E of X is *locally $\mathcal{S}$-measurable* if $E \cap A \in \mathcal{S}$ for every $A \in \mathcal{S}$.

XVI.1.4.11. If $\mathcal{S}$ is understood, we just say "locally measurable" instead of "locally $\mathcal{S}$-measurable." If $\mathcal{S}$ is closed under finite intersections, then every set in $\mathcal{S}$ is locally measurable; the converse holds if $X \in \mathcal{S}$. Thus if $\mathcal{S}$ is an algebra, then the locally $\mathcal{S}$-measurable sets are precisely the sets in $\mathcal{S}$. Local measurability is most important if $\mathcal{S}$ is a δ-ring which is not a σ-algebra.

XVI.1.4.12. If $\mathcal{S}$ is closed under finite intersections, so is the collection $\mathcal{M}$ of locally $\mathcal{S}$-measurable sets since $(E \cap F) \cap A = (E \cap A) \cap (F \cap A)$, and $\mathcal{S}$ is a hereditary subset of $\mathcal{M}$, motivating the term "locally measurable." If $\mathcal{R}$ is a ring, then the collection $\mathcal{M}$ of locally $\mathcal{R}$-measurable sets is an algebra, since if $E \in \mathcal{M}$ and $A \in \mathcal{R}$, then $E^c \cap A = A \setminus (A \cap E) \in \mathcal{R}$. If $\mathcal{D}$ is a δ-ring, then $\mathcal{M}$ is a σ-algebra and $\mathcal{D}$ is a hereditary subset of $\mathcal{M}$ (). If $\mathcal{A}$ is the σ-algebra generated by the δ-ring $\mathcal{D}$, then $\mathcal{A} \subseteq \mathcal{M}$ by (). The containment is proper in general unless $\mathcal{D}$ is σ-finite: for example, if X is an uncountable set and $\mathcal{D}$ is the σ-ring of countable subsets of X, then $\mathcal{A}$ is the collection of countable and cocountable sets, but every subset of X is locally measurable.

XVI.1.5. π-Systems, λ-Systems, and Monotone Classes

We will occasionally have use for some collections of sets with other combinations of closure properties, and it will be convenient to have names for them.

XVI.1.5.1. Definition. Let X be a set.

A nonempty collection $\mathcal{Q}$ of subsets of X is a *π-system* if it is closed under finite intersections.

A collection $\mathcal{L}$ of subsets of X is a *λ-system* if it contains X and is closed under complements and countable disjoint unions.

A nonempty collection $\mathcal{N}$ of subsets of X is a *monotone class* if it is closed under countable monotone unions and intersections: if (A_n) is an increasing sequence in $\mathcal{N}$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{N}$, and if (B_n) is a decreasing sequence in $\mathcal{N}$, then $\cap_{n=1}^{\infty} B_n \in \mathcal{N}$.

XVI.1.5.2. EXAMPLES. (i) A semiring is a π-system. A σ-algebra is both a λ-system and a monotone class. A λ-system contains $\emptyset$, and is thus also closed under finite disjoint unions.

(ii) A π-system need not be a semiring. For example, a topology on X is a π-system but usually not a semiring.

(iii) A λ-system need not be closed under even finite unions. For example, let X be a finite set with an even number of elements (> 2), and $\mathcal{L}$ the collection of all subsets with an even number of elements. Then $\mathcal{L}$ is a λ-system, but is not closed under finite unions. $\mathcal{L}$ is also a monotone class.

(iv) A monotone class need not be closed under either finite unions or finite intersections, or even finite disjoint unions. In fact, any finite collection of subsets of a set X is a monotone class.

(v) Any intersection of π-systems (on a fixed set X) is a π-system, and similarly for λ-systems and monotone classes. So if $\mathcal{S}$ is any collection of subsets of a set X, there is a smallest π-system containing $\mathcal{S}$, denoted $\pi(\mathcal{S})$; similarly there is a smallest λ-system $\lambda(\mathcal{S})$ and a smallest monotone class $\upsilon(\mathcal{S})$ containing $\mathcal{S}$. $\pi(\mathcal{S})$ just consists of all finite intersections of sets in $\mathcal{S}$, but $\lambda(\mathcal{S})$ and $\upsilon(\mathcal{S})$ do not have simple descriptions in general.

XVI.1.5.3. PROPOSITION. (i) A λ-system of subsets of a set X is closed under relative monotone complements: if $A, B \in \mathcal{L}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{L}$.

(ii) Let $\mathcal{L}$ be a collection of subsets of a set X containing X and closed under relative monotone complements and finite and countable disjoint unions. Then $\mathcal{L}$ is a λ-system.

PROOF: (i): $B \setminus A = [A \cup B^c]^c$, and $A \cap B^c = \emptyset$.
(ii): If $A \in \mathcal{L}$, then $A^c = X \setminus A \in \mathcal{L}$ since $X \in \mathcal{L}$. 👍

XVI.1.5.4. PROPOSITION. A collection $\mathcal{A}$ of subsets of a set X which is both a π-system and a λ-system is a σ-algebra.

PROOF: $\mathcal{A}$ is closed under complements and finite intersections, hence an algebra. Thus $\mathcal{A}$ is closed under relative complements. If (A_n) is a sequence of sets in A, let $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n > 1$, then (B_n) is a pairwise disjoint sequence of sets in $\mathcal{A}$, so $\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n \in \mathcal{A}$, so $\mathcal{A}$ is closed under countable unions. 👍

The following variation of this result will be important:

XVI.1.5.5. THEOREM. [DYNKIN'S $(\pi - \lambda)$-THEOREM] Let X be a set, $\mathcal{Q}$ a π-system in X, and $\mathcal{L}$ a λ-system in X. If $\mathcal{Q} \subseteq \mathcal{L}$, then $\sigma(\mathcal{Q}) \subseteq \mathcal{L}$.

PROOF: By XVI.1.5.4., it suffices to show that $\lambda(\mathcal{Q})$ is a π-system, since $\lambda(\mathcal{Q})$ is a λ-system contained in $\mathcal{L}$ by minimality.

If $A \subseteq X$, let

$$\mathcal{L}_A = \{B \subseteq X : A \cap B \in \lambda(\mathcal{Q})\} .$$

If $A \in \lambda(\mathcal{Q})$, then we show that $\mathcal{L}_A$ is a λ-system. X is trivially in $\mathcal{L}_A$. If $B, C \in \mathcal{L}_A$, $C \subseteq B$, then $A \cap B$ and $A \cap C$ are in $\lambda(\mathcal{Q})$, so $A \cap (B \setminus C) = (A \cap C) \setminus (A \cap B) \in \lambda(\mathcal{Q})$ by XVI.1.5.3.(i), since $\lambda(\mathcal{Q})$ is a λ-system. Thus $B \setminus C \in \mathcal{L}_A$ and $\mathcal{L}_A$ is closed under relative monotone complements. If (B_n) is a pairwise disjoint sequence in $\mathcal{L}_A$, then $(A \cap B_n)$ is a pairwise disjoint sequence in $\lambda(\mathcal{Q})$, hence $A \cap [\cup_{n=1}^{\infty} B_n] = \cup_{n=1}^{\infty}(A \cap B_n) \in \lambda(\mathcal{Q})$, i.e. $\cup_{n=1}^{\infty} B_n \in \mathcal{L}_A$, and $\mathcal{L}_A$ is closed under countable disjoint unions. So $\mathcal{L}_A$ is a λ-system if $A \in \lambda(\mathcal{Q})$, and in particular if $A \in \mathcal{Q}$.

Now suppose $A \in \mathcal{Q}$. If $B \in \mathcal{Q}$, then $A \cap B \in \mathcal{Q}$ since $\mathcal{Q}$ is a π-system, and thus $B \in \mathcal{L}_A$. So if $A \in \mathcal{Q}$, then $\mathcal{Q} \subseteq \mathcal{L}_A$, and thus $\lambda(\mathcal{Q}) \subseteq \mathcal{L}_A$ by minimality of $\lambda(\mathcal{Q})$.

Note that if $A, B \subseteq X$, then $B \in \mathcal{L}_A$ if and only if $A \in \mathcal{L}_B$. If $A \in \mathcal{Q}$ and $B \in \lambda(\mathcal{Q})$, then $B \in \mathcal{L}_A$ since $\lambda(\mathcal{Q}) \subseteq \mathcal{L}_A$, and thus $A \in \mathcal{L}_B$. Thus, for $B \in \lambda(\mathcal{Q})$, $\mathcal{Q} \subseteq \mathcal{L}_B$, and hence $\lambda(\mathcal{Q}) \subseteq \mathcal{L}_B$ by minimality of $\lambda(\mathcal{Q})$. If $B, C \in \lambda(\mathcal{Q})$, we have $C \in \mathcal{L}_B$, i.e. $B \cap C \in \lambda(\mathcal{Q})$, so $\lambda(\mathcal{Q})$ is a π-system. 👍

We have similar results for monotone classes, due to Halmos, which will be useful:

XVI.1.5.6. Proposition. An algebra of subsets of a set X which is also a monotone class is a σ-algebra.

Proof: Let $\mathcal{A}$ be an algebra of subsets of X, and suppose $\mathcal{A}$ is also a monotone class. Let (A_n) be a sequence of sets in $\mathcal{A}$. For each n, set $B_n = \cap_{k=1}^{n} A_k$. Then (B_n) is a decreasing sequence of sets in $\mathcal{A}$ since $\mathcal{A}$ is an algebra, and hence $\cap_{n=1}^{\infty} B_n = \cap_{n=1}^{\infty} A_n \in \mathcal{A}$ since $\mathcal{A}$ is a monotone class. 👍

XVI.1.5.7. Theorem. [Monotone Class Theorem] Let X be a set, $\mathcal{A}$ an algebra of subsets of a set X, and $\mathcal{N}$ a monotone class of subsets of X. If $\mathcal{A} \subset \mathcal{N}$, then $\sigma(\mathcal{A}) \subseteq \mathcal{N}$.

XVI.1.6. Measurable Spaces and Measurable Functions

XVI.1.6.1. Definition. A *measurable space* $(X, \mathcal{A})$ is a set X with a specified σ-algebra $\mathcal{A}$ of subsets.

Measurable spaces are the natural setting for measure theory, which will be developed in (). Note the analogy with the definition of a topological space. Measurable spaces are often called *Borel spaces*, especially (but not necessarily) if they come from a topological space as in ().

As usual with mathematical theories, it is important not only to have sets with a structure, but to have a class of functions which preserve the structure to serve as the morphisms of the theory. Experience shows that the appropriate morphisms between measurable spaces are the *measurable functions*:

XVI.1.6.2. Definition. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and $f : X \to Y$ a function. Then f is $(\mathcal{A}, \mathcal{B})$-*measurable* (or just *measurable* if the $\mathcal{A}$ and $\mathcal{B}$ are understood) if $f^{-1}(B) \in \mathcal{A}$ for every $B \in \mathcal{B}$.

Note the close analogy with the definition of a continuous function between topological spaces () (although there is no useful notion of "measurability at a point.") When measurable spaces are called Borel spaces, measurable functions between them are often called *Borel functions*, although this terminology can sometimes be slightly ambiguous.

XVI.1.6.3. PROPOSITION. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces, and $f : X \to Y$ and $g : Y \to Z$ functions. If f is $(\mathcal{A}, \mathcal{B})$-measurable and g is $(\mathcal{B}, \mathcal{C})$-measurable, then $g \circ f$ is $(\mathcal{A}, \mathcal{C})$-measurable.

The proof is virtually identical to the proof that a composition of continuous functions is continuous.

XVI.1.6.4. Thus the measurable spaces and measurable functions form a category. An isomorphism in this category, i.e. a bijection f such that both f and f^{-1} are measurable, is often called a *Borel isomorphism*. An isomorphism (Borel isomorphism) between measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ is a bijection $f : X \to Y$ such that, for $A \subseteq X$, $f(A) \in \mathcal{B}$ if and only if $A \in \mathcal{A}$.

XVI.1.6.5. Just as with continuity, the larger $\mathcal{A}$ is, or the smaller $\mathcal{B}$ is, the easier it is for a function to be measurable. Given $\mathcal{A}$, there is a largest $\mathcal{B}$ making f measurable, and similarly, given $\mathcal{B}$, there is a smallest corresponding $\mathcal{A}$. These are given by the following construction:

XVI.1.6.6. DEFINITION. Let X and Y be sets, and $f : X \to Y$ a function. If $\mathcal{A}$ is a σ-algebra on X, set

$$f^{\rightarrow}(\mathcal{A}) = \{B \subset Y : f^{-1}(B) \in \mathcal{A}\}.$$

If $\mathcal{B}$ is a σ-algebra on Y, set

$$f^{\leftarrow}(\mathcal{B}) = \{f^{-1}(B) : B \in \mathcal{B}\}.$$

XVI.1.6.7. PROPOSITION. Let X and Y are measurable spaces and $f : X \to Y$ a function.

(i) If $\mathcal{A}$ is a σ-algebra on X, then $f^{\rightarrow}(\mathcal{A})$ is a σ-algebra on Y.

(ii) If $\mathcal{B}$ is a σ-algebra on Y, then $f^{\leftarrow}(\mathcal{B})$ is a σ-algebra on X.

PROOF: Both parts follow easily from the fact that $f^{-1}(\emptyset) = \emptyset$, $f^{-1}(Y) = X$, $f^{-1}(B^c) = [f^{-1}(B)]^c$, and $f^{-1}(\cup_{n=1}^{\infty} B_n) = \cup_{n=1}^{\infty} f^{-1}(B_n)$ for any $B, B_n \subseteq Y$ (). 👍

XVI.1.6.8. In fact, if $(X, \mathcal{A})$ is a measurable space and Y is a set, and $f : X \to Y$ is a function, then $f^{\rightarrow}(\mathcal{A})$ is the largest σ-algebra on Y making f measurable. Similarly, if $f : X \to (Y, \mathcal{B})$, then $f^{\leftarrow}(\mathcal{B})$ is the smallest σ-algebra on X making f measurable.

Note, however, that $f^{\rightarrow}(\mathcal{A})$ and $f^{\leftarrow}(\mathcal{B})$ can be quite degenerate: for example, if f is a constant function, then $f^{\rightarrow}(\mathcal{A}) = \mathcal{P}(Y)$ and $f^{\leftarrow}(\mathcal{B}) = \{\emptyset, X\}$ no matter what $\mathcal{A}$ and $\mathcal{B}$ are. There are examples where $f^{\rightarrow}(\mathcal{A}) = \{\emptyset, Y\}$ even for nondegenerate f and $\mathcal{A}$ (); $f^{\leftarrow}(\mathcal{B})$ is less degenerate in general.

Although it would appear that $f^{\rightarrow}$ and $f^{\leftarrow}$ are "inverse" to each other, this is not the case in general: we always have that $f^{\leftarrow}(f^{\rightarrow}(\mathcal{A})) \subseteq \mathcal{A}$ and $f^{\rightarrow}(f^{\leftarrow}(\mathcal{B})) \supseteq \mathcal{B}$, but equality does not hold in general ().

The next theorem is one of the most useful facts about mappings of σ-algebras.

XVI.1.6.9. THEOREM. Let X and Y be sets, $f : X \to Y$ a function, and $\mathcal{B}$ a σ-algebra on Y. If $\mathcal{C} \subseteq \mathcal{B}$ is a subset which generates $\mathcal{B}$ as a σ-algebra, then $f^{\leftarrow}(\mathcal{C}) = \{f^{-1}(C) : C \in \mathcal{C}\}$ generates $f^{\leftarrow}(\mathcal{B})$ as a σ-algebra.

PROOF: Let $\mathcal{A}$ be the σ-algebra on X generated by $f^{\leftarrow}(\mathcal{C})$. Then $\mathcal{A} \subseteq f^{\leftarrow}(\mathcal{B})$ since by XVI.1.6.7. $f^{\leftarrow}(\mathcal{B})$ is a σ-algebra containing $f^{\leftarrow}(\mathcal{C})$. Let

$$\mathcal{B}_0 = f^{\rightarrow}(\mathcal{A}) \cap \mathcal{B} = \{B \in \mathcal{B} : f^{-1}(B) \in \mathcal{A}\}.$$

By XVI.1.6.7., $\mathcal{B}_0$ is the intersection of two σ-algebras, hence is a σ-algebra; and $\mathcal{B}_0 \subseteq \mathcal{B}$. Also, $\mathcal{C} \subseteq \mathcal{B}_0$. But $\mathcal{C}$ generates $\mathcal{B}$ as a σ-algebra, so $\mathcal{B}_0 = \mathcal{B}$, and hence $f^{-1}(B) \in \mathcal{A}$ for any $B \in \mathcal{B}$, i.e. $f^{\leftarrow}(\mathcal{B}) \subseteq \mathcal{A}$, so $\mathcal{A} = f^{\leftarrow}(\mathcal{B})$. 👍

XVI.1.6.10. COROLLARY. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, $f : X \to Y$ a function, and $\mathcal{C}$ a subset of $\mathcal{B}$ which generates $\mathcal{B}$ as a σ-algebra. Then f is measurable if and only if $f^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{C}$.

PROOF: It is obvious that if f is measurable, then $f^{-1}(C) \in \mathcal{A}$ for every $C \in \mathcal{C}$. The converse follows immediately from XVI.1.6.9., since by assumption the σ-algebra on X generated by $f^{\leftarrow}(\mathcal{C})$, which by the theorem is $f^{\leftarrow}(\mathcal{B})$, is contained in $\mathcal{A}$. 👍

The next simple observation is extraordinarily useful in defining measurable functions piecewise:

XVI.1.6.11. PROPOSITION. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and let (E_n) be a sequence in $\mathcal{A}$ with $\cup_{n=1}^{\infty} E_n = X$. (Since $\emptyset \in \mathcal{A}$, (E_n) could also be a finite sequence.) Let $f : X \to Y$ be a function, and set $f_n = f|_{E_n} : E_n \to Y$. If each f_n is measurable, then f is measurable.

PROOF: Let $B \in \mathcal{B}$. We need to show that $f^{-1}(B) \in \mathcal{A}$. By assumption, $f_n^{-1}(B) = \{x \in E_n : f_n(x) = f(x) \in B\}$ is in $\mathcal{A}$. But $f^{-1}(B) = \cup_{n=1}^{\infty} f_n^{-1}(B)$. 👍

A slight rephrasing is usually used in piecewise definition of functions:

XVI.1.6.12. COROLLARY. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and let (E_n) be a sequence in $\mathcal{A}$ which partitions X, i.e. the E_n are pairwise disjoint and $\cup_{n=1}^{\infty} E_n = X$. (Since $\emptyset \in \mathcal{A}$, (E_n) could also be a finite sequence.) For each n, let $f_n : E_n \to Y$ be a function, and define $f : X \to Y$ by $f(x) = f_n(x)$ if $x \in E_n$. If each f_n is measurable, then f is measurable.

XVI.1.6.13. We can also rephrase the last two results in another way. If $(X, \mathcal{A})$ is a measurable space, and $E \in \mathcal{A}$, the *restriction* of $\mathcal{A}$ to E is

$$\mathcal{A}_E = \{A \cap E : A \in \mathcal{A}\} = \{B : B \in \mathcal{A}, B \subseteq E\}$$

which can be easily verified to be a σ-algebra of subsets of E (). The measurable space $(E, \mathcal{A}_E)$ is called a *measurable subspace*, or *Borel subspace*, of $(X, \mathcal{A})$.

XVI.1.6.14. DEFINITION. (i) If $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are measurable spaces, the *sum* or *disjoint union* is the measurable space $(Z, \mathcal{C})$, where Z is the disjoint union $X \sqcup Y$ () and

$$\mathcal{C} = \{A \sqcup B : A \in \mathcal{A}, B \in \mathcal{B}\}$$

(it is easy to check that $\mathcal{C}$ is a σ-algebra of subsets of $X \sqcup Y$). The sum is denoted $(X, \mathcal{A}) \oplus (Y, \mathcal{B})$.
(ii) Let $(X_n, \mathcal{A}_n)$ be a finite or countable set of measurable spaces. The *sum* or *disjoint union*, denoted $\bigoplus_n (X_n, \mathcal{A}_n)$, is the measurable space $(Z, \mathcal{C})$, where $Z = \sqcup_n X_n$ and

$$\mathcal{C} = \{\sqcup_n A_n : A_n \in \mathcal{A}_n \text{ for all } n\}$$

(it is again easy to check that $\mathcal{C}$ is a σ-algebra of subsets of Z (Exercise ()).)

Then XVI.1.6.11. can be rephrased:

XVI.1.6.15. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space, and let (E_n) be a sequence in $\mathcal{A}$ which partitions X, i.e. the E_n are pairwise disjoint and $\cup_{n=1}^{\infty} E_n = X$. (Since $\emptyset \in \mathcal{A}$, (E_n) could also be a finite sequence.) Let f_n be the inclusion map of E_n into X, and $f : \sqcup_n E_n \to X$ the induced map. Then f is a Borel isomorphism from $\bigoplus_n (E_n, \mathcal{A}_{E_n})$ onto $(X, \mathcal{A})$. In particular, if $E \in \mathcal{A}$, then $(X, \mathcal{A})$ is Borel isomorphic to $(E, \mathcal{A}_E) \oplus (X \setminus E, \mathcal{A}_{X \setminus E})$.

The Schröder-Bernstein Theorem for Measurable Spaces

XVI.1.6.16. THEOREM. If $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are measurable spaces, $(X, \mathcal{A})$ is Borel isomorphic to a measurable subspace of $(Y, \mathcal{B})$, and $(Y, \mathcal{B})$ is Borel isomorphic to a measurable subspace of $(X, \mathcal{A})$, then $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are Borel isomorphic.

The proof is a simple adaptation of the proof of the Schröder-Bernstein Theorem of set theory. See Exercise XVI.1.7.3..

XVI.1.7. Exercises

XVI.1.7.1. This exercise should convince the reader of (1) the truth of the statement of XVI.1.2.4. and (2) the author's wisdom in not attempting to write out a full proof. Adopt the situation and notation of XVI.1.2.4..

(a) Verify the four identities in the proof outline.

(b) Let A_i, B_i, C_i, D_i $(i = 1, 2)$ be sets in $\mathcal{C}$. Show that

$$[(A_1 \setminus B_1) \cap (A_2 \setminus B_2)] \setminus [(C_1 \setminus D_1) \cap (C_2 \setminus D_2)]$$

is a finite disjoint union of finite intersections of sets in $\mathcal{C} \cup \mathcal{D}$, using the four identities. [The first step is to use the fourth identity to rewrite the set as

$$[(A_1 \setminus B_1) \cap (A_2 \setminus B_2)] \setminus (C_1 \setminus D_1) \sqcup \{[(A_1 \setminus B_1) \cap (A_2 \setminus B_2)] \cap (C_1 \setminus D_1)\} \setminus (C_2 \setminus D_2)$$

where $\sqcup$ denotes disjoint union.]

XVI.1.7.2. Let $\mathcal{E}$ be a collection of subsets of a set X. Set $\tilde{\mathcal{E}} = \{X \setminus E : E \in \mathcal{E}\}$.

(a) Show that if $\mathcal{E}$ is closed under finite unions and finite intersections, and $\mathcal{E} \cap \tilde{\mathcal{E}}$ is nonempty, then $\mathcal{E} \cap \tilde{\mathcal{E}}$ is an algebra of subsets of X and is the largest algebra contained in $\mathcal{E}$.

(b) Show that if $\mathcal{E}$ is closed under countable unions and countable intersections, and $\mathcal{E} \cap \tilde{\mathcal{E}}$ is nonempty, then $\mathcal{E} \cap \tilde{\mathcal{E}}$ is a σ-algebra of subsets of X and is the largest σ-algebra contained in $\mathcal{E}$.

(c) Even if $\mathcal{E}$ is infinite and closed under countable unions and countable intersections, $\mathcal{E} \cap \tilde{\mathcal{E}}$ can be empty. [Let $\mathcal{E}$ be the collection of countable subsets of an uncountable set.]

XVI.1.7.3. Prove the Schröder-Bernstein Theorem for measurable spaces. [Mimic the proof of the Schröder-Bernstein Theorem of set theory (II.5.6.11.).]

XVI.1.7.4. Modify the construction in XVI.1.3.11., replacing finite unions and intersections with countable unions and intersections in the construction of $\mathcal{S}_1$ and $\mathcal{S}_2$. Is $\mathcal{S}_2$ then the σ-algebra generated by $\mathcal{S}$? Where does the proof break down?

XVI.1.7.5. Let X be an infinite set, and $\mathcal{A}$ an algebra of subsets of X containing a sequence of distinct subsets (this will hold if $\mathcal{A}$ is infinite if the Countable AC is assumed).

(a) Show by recursion (without any form of Choice) that $\mathcal{A}$ contains a sequence of pairwise disjoint nonempty subsets of X. [If (A_n) is the sequence (we may assume no A_n is X or $\emptyset$), then either $\{A_1 \cap A_n : n > 1\}$ or $\{A_1^c \cap A_n : n > 1\}$ (or both) contains infinitely many distinct sets. If the first holds, set $B_1 = A_1^c$ and if not set $B_1 = A_1$. Repeat replacing X by A_1 in the first case and A_1^c in the second, and A_n by $A_1 \cap A_n$ or $A_1^c \cap A_n$ (reduced to a subsequence of distinct nonempty proper subsets of A_1 or A_1^c respectively).]

(b) Show that if $\mathcal{A}$ is a σ-algebra, then $\mathcal{A}$ has cardinality at least $2^{\aleph_0}$. In fact, $\mathcal{A}$ contains a chain of (distinct) sets order-isomorphic to $\mathbb{R}$, i.e. a collection $\{C_r : r \in \mathbb{R}\}$ with $C_r \subseteq C_s$ if and only if $r \leq s$. [Index the B's by $\mathbb{Q}$ and let $C_r = \cup_{q<r} B_q$; cf. II.9.5.21..]

(c) If $\mathcal{A}$ is a σ-algebra, the sets in (a) can be chosen to have union all of X. This is not possible in general if $\mathcal{A}$ is just an algebra. [Consider the algebra of finite/cofinite subsets of an uncountable set.]

XVI.1.8. Filters and Ultrafilters

Filters are another important type of subspace structure. Filters are useful in a wide variety of mathematical applications, especially in topology.

XVI.1.8.1. DEFINITION. Let X be a set. A *filter* on X is a nonempty collection $\mathcal{F}$ of subsets of X such that

(i) $\mathcal{F}$ is closed under finite intersections.

(ii) $\mathcal{F}$ has the *superset property*: if $A \in \mathcal{F}$ and $A \subseteq B \subseteq X$, then $B \in \mathcal{F}$.

(iii) $\emptyset \notin \mathcal{F}$.

A filter on X which is not properly contained in any other filter on X is called an *ultrafilter* on X.

XVI.1.8.2. It follows from the superset property that a filter $\mathcal{F}$ is closed under arbitrary unions, and, since $\mathcal{F}$ is nonempty, that $X \in \mathcal{F}$. Thus a filter is somewhat similar to a topology. But property (iii) gives a crucial difference between a filter and a topology (a topology also does not have the superset property in general).

In fact, if $\mathcal{F}$ is a filter on X and $A \subseteq X$, we cannot have both A and A^c in $\mathcal{F}$, since then we would have $A \cap A^c = \emptyset \in \mathcal{F}$. It can happen, though, that neither A nor A^c is in $\mathcal{F}$. (This cannot happen in an ultrafilter, and in fact gives a characterization of ultrafilters (XVI.1.8.12.).)

XVI.1.8.3. A useful point of view, which is helpful in determining abstract properties of filters and appropriate for some, but not all, applications, is to regard a filter (especially an ultrafilter) as giving a notion of "size" for subsets of X: sets in the filter are regarded as "large" subsets and those not in the filter as "small" subsets.

XVI.1.8.4. As with many other subset structures, the intersection of any collection $\{\mathcal{F}_i : i \in I\}$ of filters on X is a filter on X (note that the intersection is a nonempty collection of subsets since $X \in \mathcal{F}_i$ for all i). However, the union of two filters is not a filter in general (in fact, see XVI.1.8.8.); but the union of a *chain* (under inclusion) of filters is a filter (Exercise ()).

XVI.1.8.5. If $\mathcal{F}$ is a filter on X, the set $\mathcal{I} = \{A^c : A \in \mathcal{F}\}$ is closed under finite unions and subsets, and does not contain the whole set X. Such a collection is called an *order ideal*. Order ideals and hence filters can be defined in general partially ordered sets ().

XVI.1.8.6. EXAMPLES. (i) Let A be a fixed nonempty subset of X, and let

$$\mathcal{F}_A = \{B \subseteq X : A \subseteq B\} .$$

Then $\mathcal{F}_A$ is a filter on X, which is an ultrafilter if and only if A is a singleton (note that $\mathcal{F}_A \subseteq \mathcal{F}_B$ if and only if $A \supseteq B$). If $x \in X$, write $\mathcal{F}_x$ to mean $\mathcal{F}_{\{x\}}$. Thus $\mathcal{F}_x = \{B \subseteq X : x \in B\}$.

(ii) Let X be an infinite set, and $\mathcal{F}$ the collection of all subsets of X with finite complement. Then $\mathcal{F}$ is a filter on X. This filter is never an ultrafilter (cf. XVI.1.8.12.). More generally, if κ is an infinite cardinal less

than or equal to $card(X)$, the set $\mathcal{F}$ of subsets A of X for which $card(A^c) < \kappa$ is a filter on X; in particular, if X is uncountable, the set of cocountable subsets of X is a filter on X.

(iii) Let $(X, \mathcal{T})$ be a topological space, and $x \in X$. Let $\mathcal{O}_x$ be the collection of all $\mathcal{T}$-neighborhoods () of x. Then $\mathcal{O}_x$ is a filter on X which is contained in $\mathcal{F}_x$; they are not the same in general since $\{x\} \in \mathcal{O}_x$ if and only if x is an isolated point in X (i.e. $\{x\} \in \mathcal{T}$). Thus $\mathcal{O}_x$ is usually not an ultrafilter on X. More generally, x can be replaced by any nonempty subset of X. (The notation $\mathcal{O}_x$ should really include the topology $\mathcal{T}$ since this filter depends on $\mathcal{T}$ as well as x.)

The intersection of all the sets in a filter may be empty (XVI.1.8.6.(ii)) or nonempty (XVI.1.8.6.(i)). We make the following definition:

XVI.1.8.7. Definition. Let $\mathcal{F}$ be a filter on a set X. $\mathcal{F}$ is a *fixed filter* with fixed set A if the intersection of all the sets in $\mathcal{F}$ is the nonempty set A. If the intersection of all sets in $\mathcal{F}$ is empty, $\mathcal{F}$ is a *free filter* on X.

An ultrafilter on a set X is either of the form $\mathcal{F}_x$ (XVI.1.8.6.(i)) for some $x \in X$, or is free. The ultrafilter $\mathcal{F}_x$ is often called the *principal ultrafilter* at the point x.

XVI.1.8.8. If $\mathcal{F}$ and $\mathcal{G}$ are filters on X, then there may not be any filter on X at all containing $\mathcal{F} \cup \mathcal{G}$: this will fail if there is a subset A of X with $A \in \mathcal{F}$ and $A^c \in \mathcal{G}$ (XVI.1.8.2.), e.g. if $\mathcal{F}$ and $\mathcal{G}$ are $\mathcal{F}_x$ and $\mathcal{F}_y$ as in XVI.1.8.6.(i) for distinct x and y. If there is such a set A, $\mathcal{F}$ and $\mathcal{G}$ are said to be *separated* (). Thus an arbitrary collection of nonempty subsets of X need not be contained in a filter on X. We will characterize when a collection of subsets is contained in a filter.

XVI.1.8.9. Definition. A collection $\mathcal{S}$ of subsets of a set X has the *finite intersection property* if the intersection of any finite number of sets in $\mathcal{S}$ is nonempty.

Any subset of a filter on a set X has the finite intersection property. The converse is also true:

XVI.1.8.10. Proposition. Let $\mathcal{S}$ be a collection of subsets of a set X, with the finite intersection property. Then there is a filter on X containing $\mathcal{S}$.

Proof: Let $\mathcal{F}$ be the collection of all subsets of X which contain a finite intersection of sets in $\mathcal{S}$. It is easy to verify that $\mathcal{F}$ is a filter containing $\mathcal{S}$, and is the smallest such filter. 👍

We now aim towards the question of whether every filter is contained in an ultrafilter, and along the way we will obtain an important explicit characterization of ultrafilters. We first obtain a fairly simple, but important, one-step extension theorem.

XVI.1.8.11. Theorem. [Filter Extension Theorem] Let $\mathcal{F}$ be a filter on a set X, and $A \subseteq X$. If $A^c \notin \mathcal{F}$, then there is a filter $\mathcal{G}$ on X containing $\mathcal{F}$ and A, i.e. with $\mathcal{F} \subseteq \mathcal{G}$ and $A \in \mathcal{G}$.

Note that the condition $A^c \notin \mathcal{F}$ is necessary for the existence of $\mathcal{G}$, since otherwise both A and A^c would be in $\mathcal{G}$.

PROOF: Let

$$\mathcal{G} = \{C \subseteq X : C \supseteq (B \cap A) \text{ for some } B \in \mathcal{F}\} .$$

If $B \in \mathcal{F}$, then $B \supseteq (B \cap A)$, so $B \in \mathcal{G}$, and thus $\mathcal{F} \subseteq \mathcal{G}$. We also have $A \in \mathcal{G}$ since $A \supseteq (X \cap A)$.

We show that $\mathcal{G}$ is a filter on X. $\mathcal{G}$ obviously has the superset property, and is nonempty since $\mathcal{F} \subseteq \mathcal{G}$. If $C_1, C_2 \in \mathcal{G}$, with $C_1 \supseteq (B_1 \cap A)$ and $C_2 \supseteq (B_2 \cap A)$ for $B_1, B_2 \in \mathcal{F}$, then

$$C_1 \cap C_2 \supseteq (B_1 \cap A) \cap (B_2 \cap A) = (B_1 \cap B_2) \cap A$$

and $(B_1 \cap B_2) \in \mathcal{F}$, so $(C_1 \cap C_2) \in \mathcal{G}$. Hence $\mathcal{G}$ is closed under finite intersections. Finally, we show $\emptyset \notin \mathcal{G}$. If $\emptyset \in \mathcal{G}$, then $\emptyset \supseteq (B \cap A)$ for some $B \in \mathcal{F}$; for this B, we have $B \cap A = \emptyset$, so $B \subseteq A^c$ and $A^c \in \mathcal{F}$ by the superset property, contrary to hypothesis. Thus $\mathcal{G}$ is the desired filter. 👍

As an immediate consequence, we get a characterization of ultrafilters:

XVI.1.8.12. THEOREM. [ULTRAFILTER CHARACTERIZATION THEOREM] Let $\mathcal{F}$ be a filter on a set X. Then $\mathcal{F}$ is an ultrafilter if and only if, for every subset A of $\mathcal{F}$, exactly one of A and A^c is in $\mathcal{F}$.

PROOF: We cannot have both A and A^c in $\mathcal{F}$ (XVI.1.8.2.). If neither A nor A^c is in $\mathcal{F}$ for some subset A of X, then the filter $\mathcal{G}$ of XVI.1.8.11. properly contains $\mathcal{F}$, so $\mathcal{F}$ is not an ultrafilter. Conversely, if $\mathcal{F}$ is not an ultrafilter, and $\mathcal{G}$ is a filter on X properly containing $\mathcal{F}$, let $A \in \mathcal{G} \setminus \mathcal{F}$. Then $A^c \notin \mathcal{G}$ by XVI.1.8.2., so neither A nor A^c is in $\mathcal{F}$. 👍

Here is another useful observation which is a simple corollary:

XVI.1.8.13. COROLLARY. Let $\mathcal{F}$ be a filter on a set X. Then $\mathcal{F}$ is an ultrafilter if and only if, whenever $A, B \subseteq X$ and $A \cup B \in \mathcal{F}$, we have either $A \in \mathcal{F}$ or $B \in \mathcal{F}$.

PROOF: A filter with this property is an ultrafilter by setting $B = A^c$. Conversely, suppose $\mathcal{F}$ is an ultrafilter and $A \cup B \in \mathcal{F}$. If $A \notin \mathcal{F}$ and $B \notin \mathcal{F}$, we have $A^c \in \mathcal{F}$ and $B^c \in \mathcal{F}$, so

$$A^c \cap B^c = (A \cup B)^c \in \mathcal{F}$$

which contradicts XVI.1.8.2.. 👍

The property in the statement of the corollary is called the *prime property*, and a filter with this property is called a *prime filter*. Thus prime filters and maximal filters (ultrafilters) are the same thing in this context. The notions are not the same in the more general setting of filters on partially ordered sets (), however.

Passing to the set of complements, an order ideal $\mathcal{I}$ is *prime* if $A, B \subseteq X$, $A \cap B \in \mathcal{I}$ implies that $A \in \mathcal{I}$ or $B \in \mathcal{I}$. This looks more like the definition of a prime ideal in a ring (), and is the origin of the term *prime* in this context.

XVI.1.8.14. Every filter which is not an ultrafilter can be extended to a larger filter; by XVI.1.8.11., this can even be done in a very explicit way. The question then becomes whether every filter can be extended to an ultrafilter. As may be expected, the answer depends, like the answer to many similar questions (e.g. (), ...), on the Axiom of Choice:

XVI.1.8.15. THEOREM. (AC) Every filter on a set X is contained in an ultrafilter on X.

The proof of this theorem is a standard Zorn's Lemma argument, and is left to the reader (Exercise ()).

XVI.1.8.16. This theorem cannot be proved without some form of the Axiom of Choice. But the full AC is not needed: the statement of this theorem is equivalent to the Boolean Prime Ideal Theorem (II.6.3.11.), which is strictly weaker than AC.

We cannot even prove without some form of AC that there is a free ultrafilter on $\mathbb{N}$ (this statement actually follows from the Measure Extension Theorem (), which is strictly weaker than the Boolean Prime Ideal Theorem). With AC (or BPI), this is an immediate application of XVI.1.8.15. to the filter of cofinite subsets of $\mathbb{N}$ (XVI.1.8.6.(ii)).

It can be shown even without AC that many sets have a large number of pairwise separated filters. To give the best statement possible without AC, we first discuss two key property of at least some cardinals:

XVI.1.8.17. If κ is a cardinal, and X is a set of cardinality κ, then we denote by κ^2 the cardinality of $X \times X$. We always have $\kappa^2 \geq \kappa$. If $1 < \kappa < \aleph_0$, then $\kappa^2 > \kappa$. If κ is infinite, then it is conceivable that κ^2 always equals κ; this can be proved using AC (and turns out to be equivalent to AC). Without AC the following can be proved:

(i) $\aleph_0^2 = \aleph_0$ ().

(ii) More generally, if $\aleph$ is any aleph (), then $\aleph^2 = \aleph$.

(iii) $(2^{\aleph_0})^2 = 2^{\aleph_0}$ ().

XVI.1.8.18. We now turn to the other cardinal property. Let κ be a cardinal, and X a set of cardinality κ. Let Φ be the set of all finite subsets of X. We have $card(\Phi) \geq card(X)$. If κ is finite, the inequality is strict. For infinite sets, it is conceivable that equality always holds (it holds for $\aleph_0$ by ()); if $card(\Phi) = card(X)$, we say κ has the *finite subset property*.

If $\kappa^2 = \kappa \geq \aleph_0$ and X can be totally ordered, it is not difficult to show that κ has the finite subset property (II.5.6.12.). Thus $\aleph_0$ (in fact, any $\aleph$) and $2^{\aleph_0}$ have the finite subset property. So it follows from AC that every infinite cardinal has the finite subset property. It cannot be proved in ZF that every infinite cardinal has the finite subset property (), i.e. some type of Choice is necessary.

XVI.1.8.19. THEOREM. ([?]; cf. [?], [GJ76, 9.2]) Let κ be a (necessarily infinite) cardinal with $\kappa^2 = \kappa$ and the finite subset property, and X a set of cardinality κ. Then there are 2^{2^κ} pairwise separated filters on X. In particular, there are $2^{2^{\aleph_0}}$ pairwise separated filters on $\mathbb{N}$.

PROOF: Let Φ be the set of all finite subsets of X, and Λ the set of all finite subsets of Φ. Set $\Omega = \Phi \times \Lambda$. Then from $\kappa^2 = \kappa$ and the finite subset property it follows that $card(\Phi) = card(\Lambda) = card(\Omega) = \kappa$.

We will define 2^{2^κ} pairwise separated filters on Ω, which will suffice since $card(\Omega) = card(X)$. First some notation: for $S \subseteq X$, define a subset $\mathcal{B}_S$ of Ω as follows:

$$\mathcal{B}_S = \{(F, \mathcal{L}) \in \Phi \times \Lambda : S \cap F \in \mathcal{L}\} .$$

Write $\mathcal{C}_S$ for the complement of $\mathcal{B}_S$ in Ω. Then, for $\mathcal{S} \subseteq \mathcal{P}(X)$, define

$$\mathfrak{B}_\mathcal{S} = \{\mathcal{B}_S : S \in \mathcal{S}\} \cup \{\mathcal{C}_S : S \notin \mathcal{S}\} \subseteq \mathcal{P}(\Omega) \ .$$

We claim that $\mathfrak{B}_\mathcal{S}$ has the finite intersection property (XVI.1.8.9.). If $\mathfrak{A}$ is a finite subset of $\mathfrak{B}_\mathcal{S}$, the distinct sets in $\mathfrak{A}$ can be numbered as follows:

$$\mathcal{B}_{S_1}, \ldots, \mathcal{B}_{S_k}, \mathcal{C}_{S_{k+1}}, \ldots, \mathcal{C}_{S_n}$$

where $S_1, \ldots, S_n$ are distinct subsets of X. For $i \neq j$ let x_{ij} be an element of X which is in one of S_i and S_j but not the other (note that there are only finitely many choices to be made here, so no form of AC is needed). Let F be the set of all the x_{ij} for $i \neq j$. Then the sets

$$S_1 \cap F, \ldots, S_n \cap F$$

are all distinct since if $i \neq j$, x_{ij} is in one of $S_i \cap F$ and $S_j \cap F$ but not the other. Set

$$\mathcal{L} = \{S_1 \cap F, \ldots, S_k \cap F\} \in \Lambda \ .$$

We have that $S_i \cap F \in \mathcal{L}$ for $1 \leq i \leq k$, and $S_j \cap K \notin \mathcal{L}$ for $k+1 \leq j \leq n$. So $(F, \mathcal{L}) \in \mathcal{B}_{S_i}$ for $1 \leq i \leq k$ and $(F, \mathcal{L}) \in \mathcal{C}_{S_j}$ for $k+1 \leq j \leq n$, and hence the intersection of the sets in $\mathfrak{A}$ is nonempty, proving the claim.

Thus, for each $\mathcal{S} \subseteq \mathcal{P}(X)$, there is a filter $\mathfrak{F}_\mathcal{S}$ on Ω containing $\mathfrak{B}_\mathcal{S}$ by XVI.1.8.10.. If $\mathcal{S}_1$ and $\mathcal{S}_2$ are distinct subsets of $\mathcal{P}(X)$, let S be a subset of X which is in one of them (say $\mathcal{S}_1$) but not the other. Then $\mathcal{B}_S \in \mathfrak{F}_{\mathcal{S}_1}$ and $\mathcal{C}_S \in \mathfrak{F}_{\mathcal{S}_2}$, so $\mathfrak{F}_{\mathcal{S}_1}$ and $\mathfrak{F}_{\mathcal{S}_2}$ are separated. The number of distinct filters constructed is $card(\mathcal{P}(\mathcal{P}(X))) = 2^{2^\kappa}$. 👍

Note that AC is not used in this result. Using AC, the conclusion holds for all infinite cardinals. The full AC is needed here, not just the BPI (since we need $\kappa^2 = \kappa$ for all κ). Combining this result with XVI.1.8.15., we obtain:

XVI.1.8.20. THEOREM. (AC) Let X be an infinite set, of cardinality κ. Then there are exactly 2^{2^κ} free ultrafilters on X.

PROOF: By XVI.1.8.15., each $\mathfrak{F}_\mathcal{S}$ from the proof of XVI.1.8.19. is contained in an ultrafilter. These ultrafilters are all distinct since the $\mathfrak{F}_\mathcal{S}$ are pairwise separated. Thus there are at least 2^{2^κ} ultrafilters on X. On the other hand, there are at most 2^{2^κ} filters on X since there are only 2^{2^κ} sets of subsets of X. Thus there are exactly 2^{2^κ} ultrafilters on X. Only κ of these are fixed, so there are 2^{2^κ} free ones. 👍

In particular, there are $2^{2^{\aleph_0}}$ distinct free ultrafilters on $\mathbb{N}$ (assuming AC, or at least BPI).

XVI.1.9. Exercises

XVI.1.9.1. Verify the assertions in the examples of XVI.1.8.6..

XVI.1.9.2. Verify the assertion in XVI.1.8.7. that every ultrafilter which is not free is of the form $\mathcal{F}_x$ for some x.

XVI.1.9.3. (a) Verify that the set $\mathcal{F}$ defined in the proof of XVI.1.8.10. is a filter.

(b) Show that two filters $\mathcal{F}$ and $\mathcal{G}$ on a set X are separated if and only if $\mathcal{F} \cup \mathcal{G}$ fails to have the finite intersection property. Thus if $\mathcal{F}$ and $\mathcal{G}$ are not separated, they are contained in a common larger filter on X.

XVI.1.9.4. (a) Show that if $\{\mathcal{F}_i : i \in I\}$ is a chain (under inclusion) of filters on a set X, i.e. $\mathcal{F}_i \subseteq \mathcal{F}_j$ or $\mathcal{F}_j \subseteq \mathcal{F}_i$ for all $i, j \in I$, then $\cup_{i \in I} \mathcal{F}_i$ is a filter on X.

(b) Let $\mathcal{F}$ be a filter on a set X, and let $\mathcal{Y} \subseteq \mathcal{P}(\mathcal{P}(X))$ be the set of all filters on X containing $\mathcal{F}$. Partially order $\mathcal{Y}$ by inclusion. Show that every chain in $\mathcal{Y}$ has an upper bound in $\mathcal{Y}$.

(c) Conclude from Zorn's Lemma that $\mathcal{Y}$ contains a maximal element $\mathcal{G}$. Show that $\mathcal{G}$ is an ultrafilter on X.

XVI.1.9.5. Let $(\mathcal{X}, \leq)$ be a partially ordered set. Define an *order ideal* in $(\mathcal{X}, \leq)$ to be a subset $\mathcal{I}$ which is hereditary, i.e. if $y \in \mathcal{I}$, $x \in \mathcal{X}$, and $x \leq y$, then $x \in \mathcal{I}$, and such that any two elements (hence any finite set of elements) of $\mathcal{I}$ have an upper bound in $\mathcal{I}$. Define a *filter* on $(\mathcal{X}, \leq)$ to be the complement of a nonempty proper order ideal.

(a) Let X be a set. Show that order ideals and filters in $(\mathcal{P}(X), \subseteq)$ under this definition are exactly the order ideals and filters under the definitions of this section.

(b) Extend the results of this section as far as possible to filters in general partially ordered sets.

XVI.2. Borel Sets and Borel Measurable Functions

For measure theory purposes, the Borel sets are the most important subsets of $\mathbb{R}$ or $\mathbb{R}^n$. The class of Borel sets is large enough to contain every subset of $\mathbb{R}$ which can be described "naturally", and small enough that measure-theoretic operations behave well on them. Similarly, the Borel measurable functions are the natural ones to consider for integration theory. (Later we will expand the classes of Borel sets and Borel measurable functions to the larger classes of Lebesgue measurable sets and functions, primarily for technical reasons.)

We will give two characterizations of the Borel sets in this section, and a third important one in the next section, along with corresponding descriptions of the Borel measurable functions. Most of the work can be done in a general topological space, but we will emphasize the most important case of $\mathbb{R}$ in examples.

XVI.2.1. Borel Sets

XVI.2.1.1. Definition. Let $(X, \mathcal{T})$ be a topological space. Let $\mathcal{B}_\mathcal{T}$ be the σ-algebra on X generated by $\mathcal{T}$, i.e. the smallest σ-algebra containing the open sets of X. $\mathcal{B}_\mathcal{T}$ is called the class of *Borel sets* of $(X, \mathcal{T})$.

We also often write $\mathcal{B}_X$ for $\mathcal{B}_\mathcal{T}$ when it is the set and not the topology which must be emphasized (although $\mathcal{B}_X$ depends on $\mathcal{T}$ and not just on X!) When additional precision is needed, we write $\mathcal{B}_\mathcal{T}(X)$. When an understood set X such as $\mathbb{R}$ has a natural topology, we will usually just write $\mathcal{B}$ for the class of Borel sets on X, suppressing the subscript $\mathcal{T}$ or X.

XVI.2.1.2. The following sets are Borel sets in $(X, \mathcal{T})$:

(i) All open sets in X.

(i') All closed sets in X.

(ii) Any countable intersection of open sets in X. Such a subset of X is called a G_δ *set*.

(ii') Any countable union of closed sets in X. Such a subset of X is called an F_σ *set*.

(iii) Any countable union of G_δ sets in X. Such a subset of X is called a $G_{\delta\sigma}$ *set*.

(iii') Any countable intersection of F_σ sets in X. Such a subset of X is called a $F_{\sigma\delta}$ *set*.

(iv) $G_{\delta\sigma\delta}$, $F_{\sigma\delta\sigma}$, $G_{\delta\sigma\delta\sigma}$, $\cdots$ sets can be defined analogously.

XVI.2.1.3. In most cases, and in particular if $X = \mathbb{R}$, this process never terminates and must be repeated transfinitely (uncountably many times) to obtain all Borel sets (XVI.3.5.4.). (It is not obvious that these classes are never σ-algebras after finitely or countably many steps for $X = \mathbb{R}$; see () for a proof.) The complement of a G_δ set is an F_σ set, the complement of an $F_{\sigma\delta}$ is a $G_{\delta\sigma}$, etc. For measure theoretic purposes, it is rarely necessary to consider sets more complicated than G_δ's or F_σ's (which are complicated enough already!)

The notation F_σ, G_δ, etc., is due to Hausdorff. There is a more modern notation, described in XVI.3.4., which is used by descriptive set theorists; but Hausdorff's notation is still commonly used by analysts.

It should be obvious that the class of Borel sets in $\mathbb{R}$ includes a great many subsets. For example:

XVI.2.1.4.

(i) Every singleton set $\{x\}$ in $\mathbb{R}$ (or, more generally, in a T_1 space X) is a closed set, hence a Borel set. Every countable subset of $\mathbb{R}$ is thus an F_σ, hence a Borel set, since it is a countable union of singleton sets. In particular, $\mathbb{Q}$ is an F_σ, and hence $\mathbb{J}$ is a G_δ. (It turns out, although it is not obvious, that $\mathbb{Q}$ is not a G_δ, so $\mathbb{J}$ is not an F_σ; see XIII.12.1.5..)

(ii) Every open interval in $\mathbb{R}$ is an F_σ: $(a,b) = \cup_{n=1}^{\infty}[a+\frac{1}{n}, b-\frac{1}{n}]$, and similarly for $(a,+\infty)$ or $(-\infty,b)$. Since every open set in $\mathbb{R}$ is a countable union of open intervals, every open set in $\mathbb{R}$ is an F_σ, and thus every closed set is a G_δ. Of course, every open set is a G_δ, so every closed set is an F_σ.

It is not true that every open set in a general topological space X is an F_σ, but it is true if X is metrizable ().

(iii) A half-open interval in $\mathbb{R}$ is neither open nor closed, but is both an F_σ and a G_δ, and hence a Borel set. The proof is very similar to the proof of (ii), and is left as an exercise.

(iv) An arbitrary (not necessarily countable) union of actual intervals can be written as a countable union of possibly different intervals (), hence is a Borel set (in fact, an F_σ.)

An interesting and useful consequence of the construction of the Borel sets is:

XVI.2.1.5. Proposition. Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}$ (or any metrizable space) such that

(1) $\mathcal{A}$ contains all closed sets.

(2) $\mathcal{A}$ is closed under countable unions and countable intersections.

Then $\mathcal{A}$ contains all Borel sets.

Proof: $\mathcal{A}$ contains all F_σ's. Since every open set in $\mathbb{R}$ is an F_σ, $\mathcal{A}$ contains all open sets, and hence all G_δ's. Hence $\mathcal{A}$ contains all $F_{\sigma\delta}$'s, all $G_{\delta\sigma}$'s, ..., and by transfinite induction all Borel sets. 👍

XVI.2.1.6. Note that such an $\mathcal{A}$ need not be a σ-algebra, i.e. need not be closed under complements. An example is the collection of analytic sets (). A similar proof shows that a collection of subsets of a metrizable space which contains all open sets and is closed under countable unions and intersections contains all Borel sets.

XVI.2.1.7. Here is an important technicality. If X is a topological space and A is a subset of X, then A has a relative topology (). Write $\mathcal{B}_X$ for the σ-algebra of Borel sets in X. What subsets of A should be called "Borel sets in A"? There are two σ-algebras of subsets of A which are natural candidates: $\mathcal{B}_A$, the σ-algebra generated by the (relatively) open sets in A, the Borel sets for the relative topology, and

$$\mathcal{B}_X^A = \{A \cap B : B \in \mathcal{B}_X\}$$

which could be called the "relatively Borel sets in A." (It is easily checked that $\mathcal{B}_X^A$ is a σ-algebra of subsets of A.) Fortunately, these two σ-algebras coincide:

XVI.2.1.8. PROPOSITION. Let X be a topological space, $A \subseteq X$, and $\mathcal{B}_A$ and $\mathcal{B}_X^A$ as above. Then $\mathcal{B}_A = \mathcal{B}_X^A$.

PROOF: Every relatively open set in A is of the form $A \cap U$ for an open set $U \subseteq X$. Since $U \in \mathcal{B}_X$, every relatively open set in A is in the σ-algebra $\mathcal{B}_X^A$. Since $\mathcal{B}_A$ is the smallest σ-algebra on A containing the relatively open sets, $\mathcal{B}_A \subseteq \mathcal{B}_X^A$. On the other hand, set

$$\mathcal{D} = \{B_1 \cup B_2 : B_1 \in \mathcal{B}_A, B_2 \in \mathcal{B}_{A^c}\} .$$

It is easily checked that $\mathcal{D}$ is a σ-algebra of subsets of X. If U is an open set in X, then $U = (A \cap U) \cup (A^c \cap U)$, and $A \cap U$ is relatively open in A, hence in $\mathcal{B}_A$, and similarly $A^c \cap U \in \mathcal{B}_{A^c}$, so $U \in \mathcal{D}$. It follows that $\mathcal{B}_X \subseteq \mathcal{D}$. Thus

$$\mathcal{B}_X^A = \{A \cap B : B \in \mathcal{B}_X\} \subseteq \{A \cap D : D \in \mathcal{D}\} = \mathcal{B}_A .$$



XVI.2.1.9. Note that in XVI.2.1.8., A could be any subset of X, not necessarily a Borel set. But if A is a Borel set in X, then $\mathcal{B}_X^A = \{B \in \mathcal{B}_X : B \subseteq A\}$. Thus the phrase "Borel subset of A" is totally unambiguous in this case. And if $A \in \mathcal{B}_X$ the σ-algebra $\mathcal{D}$ is exactly $\mathcal{B}_X$; if $A \notin \mathcal{B}_X$, then $\mathcal{D}$ is strictly larger than $\mathcal{B}_X$ since $A \in \mathcal{D}$ (in fact $\mathcal{D}$ is exactly the σ-algebra generated by $\mathcal{B}_X$ and A).

XVI.2.2. Real-Valued Measurable Functions

XVI.2.2.1. Recall () that if $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are measurable spaces, then $f : X \to Y$ is $(\mathcal{A}, \mathcal{B})$-measurable if $f^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$. But how should we define measurability for real-valued (or extended real-valued or complex-valued) functions, i.e. what σ-algebra should we take on the range space? Experience has shown that it is almost always most useful and appropriate to take the Borel sets, and we will always do this unless specifically stated otherwise:

XVI.2.2.2. DEFINITION. Let $(X, \mathcal{A})$ be a measurable space. A real-valued function $f : X \to \mathbb{R}$ is *$\mathcal{A}$-measurable* if it is a measurable as a function from $(X, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B})$ in the sense of (), i.e. if $f^{-1}(B) \in \mathcal{A}$ for every Borel set B in $\mathbb{R}$. Denote by $\mathfrak{B}_{\mathcal{A}}$ the set of $\mathcal{A}$-measurable functions from X to $\mathbb{R}$.

We can similarly define $\mathcal{A}$-measurable functions from a measurable space $(X, \mathcal{A})$ to a topological space $(Y, \mathcal{T})$ to be functions which are $(\mathcal{A}, \mathcal{B}_{\mathcal{T}})$-measurable in the sense of (). Thus we may also define $\mathcal{A}$-measurable functions from X to $\mathbb{R}^n$.

The next proposition is a special case of XVI.1.6.3.:

XVI.2.2.3. PROPOSITION. (a) Let $(X, \mathcal{A})$ and $(Y, \mathcal{C})$ be measurable spaces, $f : X \to Y$ an $(\mathcal{A}, \mathcal{C})$-measurable function, and $g : Y \to \mathbb{R}$ a $\mathcal{C}$-measurable function. Then $g \circ f : X \to \mathbb{R}$ is $\mathcal{A}$-measurable. (Here $\mathbb{R}$ may be replaced by any topological space.)

It is often convenient to use an apparently weaker criterion for measurability of real-valued functions. Here are some of the most useful ones:

XVI.2.2.4. Proposition. Let $(X, \mathcal{A})$ be a measurable space and $f : X \to \mathbb{R}$ a function. Then f is $\mathcal{A}$-measurable if and only if any of the following conditions are satisfied:

(i) $f^{-1}(U) \in \mathcal{A}$ for every open set $U \subseteq \mathbb{R}$.

(ii) $f^{-1}(E) \in \mathcal{A}$ for every closed set $E \subseteq \mathbb{R}$.

(iii) $f^{-1}(U) \in \mathcal{A}$ for every open interval $U \subseteq \mathbb{R}$.

(iv) $f^{-1}((-\infty, a)) \in \mathcal{A}$ for every $a \in \mathbb{R}$.

(v) $f^{-1}((-\infty, a]) \in \mathcal{A}$ for every $a \in \mathbb{R}$.

(vi) $f^{-1}((a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{R}$.

(vii) $f^{-1}([a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{R}$.

(viii) $f^{-1}((-\infty, a)) \in \mathcal{A}$ for every $a \in \mathbb{Q}$.

(ix) $f^{-1}((-\infty, a]) \in \mathcal{A}$ for every $a \in \mathbb{Q}$.

(x) $f^{-1}((a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{Q}$.

(xi) $f^{-1}([a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{Q}$.

This proposition is an immediate consequence of XVI.1.6.9., since each of the collections of sets listed generates $\mathcal{B}$ as a σ-algebra.

XVI.2.2.5. Proposition. Let $(X, \mathcal{A})$ be a measurable space, and $A \subseteq X$. Then χ_A is an $\mathcal{A}$-measurable function if and only if $A \in \mathcal{A}$.

Proof: Since χ_A takes only the values 0 and 1, $\chi_A^{-1}((a, +\infty))$ is either $\emptyset$ (if $a \geq 1$), A (if $0 \leq a < 1$), or X (if $a < 0$). Apply XVI.2.4.4.(vi). 👍

More generally, we have:

XVI.2.2.6. Proposition. Let $(X, \mathcal{A})$ be a measurable space, and $\phi = \sum_{i=1}^{n} \alpha_i \chi_{A_i}$ be a simple function from X to $\mathbb{R}$. If each $A_i \in \mathcal{A}$, then ϕ is $\mathcal{A}$-measurable. The converse holds if the α_i are distinct.

The proof is almost identical to the proof of XVI.2.2.5., and is left as an exercise.

Suprema and Pointwise Limits of Measurable Functions

A simple consequence of () is:

XVI.2.2.7. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space, and (f_n) a sequence of $\mathcal{A}$-measurable functions from X to $\mathbb{R}$. If (f_n) is pointwise bounded (i.e. for every $x \in X$, $\{f_n(x) : n \in \mathbb{N}\}$ is bounded above), then $\sup f_n$ is $\mathcal{A}$-measurable. If (f_n) is pointwise bounded below, then $\inf f_n$ is $\mathcal{A}$-measurable.

PROOF: Let $a \in \mathbb{R}$. Then

$$\{x : \sup f_n(x) \leq a\} = \bigcap_{n\in\mathbb{N}} \{x : f_n(x) \leq a\}$$

$$\{x : \inf f_n(x) \geq a\} = \bigcap_{n\in\mathbb{N}} \{x : f_n(x) \geq a\}$$

and the sets on the right are in $\mathcal{A}$, which is a σ-algebra. 👍

One of the most important facts about the collection of $\mathcal{A}$-measurable real-valued functions on a measurable space $(X, \mathcal{A})$ is that it is closed under pointwise limits:

XVI.2.2.8. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space and (f_n) a sequence of real-valued functions on X converging pointwise to $f : X \to \mathbb{R}$. If each f_n is $\mathcal{A}$-measurable, so is f.

PROOF: By (), we need only show that $A := \{x \in X : f(x) \leq a\} \in \mathcal{A}$ for every $a \in \mathbb{R}$. Fix a. For $m, n \in \mathbb{N}$, set $A_{m,n} = \{x \in A : f_n(x) \leq a + \frac{1}{m}\}$. Then

$$A = \bigcap_{m\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} \bigcap_{n\geq k} A_{m,n}$$

and by assumption, each $A_{m,n} \in \mathcal{A}$ and $\mathcal{A}$ is a σ-algebra. 👍

See XVI.2.6.12. for an alternate proof, and XVI.2.4.8.–XVI.2.4.11. for additional properties of the set of $\mathcal{A}$-measurable functions.

Approximation by Simple Functions

Every real-valued measurable function is a pointwise limit of simple functions, in a controlled way:

XVI.2.2.9. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space, and $f : X \to \mathbb{R}$ an $\mathcal{A}$-measurable function. Then

(i) There is a sequence (f_n) of simple $\mathcal{A}$-measurable functions converging pointwise to f.

(ii) If f is bounded, there is a sequence (f_n) of simple $\mathcal{A}$-measurable functions converging uniformly to f.

(iii) If f is nonnegative, there is an increasing sequence (f_n) of simple $\mathcal{A}$-measurable functions converging pointwise to f, uniformly if f is bounded.

PROOF: Fix $n \in \mathbb{N}$. For each $k \in \mathbb{Z}$, let

$$E_{kn}(f) = f^{-1}\left(\left[\frac{k}{n}, \frac{k+1}{n}\right)\right).$$

Then each $E_{kn}(f)$ is in $\mathcal{A}$ since it is the preimage of a Borel set, and for fixed n we have that $\mathbb{R}$ is the disjoint union of the $E_{kn}(f)$'s. Let $f_n(x) = \frac{k}{n}$ if $x \in E_{kn}(f)$. Since $\frac{k}{n} \leq f(x) < \frac{k+1}{n}$ for $x \in E_{kn}(f)$, we have that $f_n(x) \leq f(x)$ and $f(x) - f_n(x) < \frac{1}{n}$ for every x. Thus the sequence (f_n) actually converges uniformly to f. If f is bounded, each f_n is simple, proving (ii). However, f_n is not simple if f is unbounded. To obtain simple functions, let $g_n(x) = f_n(x)$ if $x \in E_{kn}(f)$ for $-n \leq k \leq n$ and $g_n(x) = 0$ otherwise. Then g_n is simple, and $g_n \to f$ pointwise on $\mathbb{R}$ (but not uniformly unless f is bounded.) In case (iii), take a subsequence of the n's, each dividing the next (e.g. powers of 2). 👍

XVI.2.3. Real-Valued Measurable Functions on δ-Rings

XVI.2.3.1. If X and Y are sets, $\mathcal{R}$ is a collection of subsets of X, and $\mathcal{S}$ a collection of subsets of Y, Definition () makes sense (but is relatively uninteresting in this generality) to say that a function $f : X \to Y$ is $(\mathcal{R}, \mathcal{S})$-measurable if $f^{-1}(S) \in \mathcal{R}$ for all $S \in \mathcal{S}$. But if f is a real-valued function on $\mathcal{R}$, the usual definition () of measurability for f with $Y = \mathbb{R}$, $\mathcal{S} = \mathcal{B}$ is no good (never satisfied) unless $X \in \mathcal{R}$, since $f^{-1}(\mathbb{R}) = X$. So even if $\mathcal{R}$ is a σ-ring, this definition must be modified unless $\mathcal{R}$ is a σ-algebra.

XVI.2.3.2. There are several meaningful ways to generalize the definition of measurability. If $\mathcal{D}$ is a collection of subsets of a set X, normally at least a δ-ring, let $\mathcal{A}$ be the σ-algebra generated by $\mathcal{D}$ and $\mathcal{M}$ the σ-algebra of locally $\mathcal{D}$-measurable sets (). Then if $\mathcal{D}$ is a δ-ring, $\mathcal{A} \subseteq \mathcal{M}$ and $\mathcal{D}$ is a hereditary subset of both $\mathcal{A}$ and $\mathcal{M}$ ().

XVI.2.3.3. DEFINITION. Let $\mathcal{D}$ be a δ-ring and $\mathcal{A}$ and $\mathcal{M}$ as above, and let $f : X \to \mathbb{R}$ be a function.
(i) The function f is *locally $\mathcal{D}$-measurable* if it is $\mathcal{M}$-measurable, i.e.

$$A \cap f^{-1}(B) = \{x \in A : f(x) \in B\} \in \mathcal{D}$$

for every $A \in \mathcal{D}$ and Borel set $B \subseteq \mathbb{R}$.
(ii) The function f is *strictly $\mathcal{D}$-measurable* if $N := \{x \in X : f(x) \neq 0\} \in \mathcal{D}$ and

$$N \cap f^{-1}(B) = f^{-1}(B \setminus \{0\}) \in \mathcal{D}$$

for all Borel sets $B \subseteq \mathbb{R}$. (Note that $N = f^{-1}(\mathbb{R} \setminus \{0\})$ so the second condition includes the first.)

XVI.2.3.4. There is a third condition, that f be $\mathcal{A}$-measurable; we could call this *$\mathcal{D}$-measurability*. Since $\mathcal{A} \subseteq \mathcal{M}$, $\mathcal{D}$-measurability implies local $\mathcal{D}$-measurability. If f is strictly $\mathcal{D}$-measurable, then $N^c = \{x \in X : f(x) = 0\} \in \mathcal{A}$, and if B is a Borel set, then $f^{-1}(B)$ is either $N \cap f^{-1}(B)$ (if $0 \notin B$) or $N^c \cup (N \cap f^{-1}(B))$ (if $0 \in B$), and in either case $f^{-1}(B) \in \mathcal{A}$. Thus strictly $\mathcal{D}$-measurable implies $\mathcal{D}$-measurable implies locally $\mathcal{D}$-measurable. If $\mathcal{D}$ is a σ-algebra, then $\mathcal{A} = \mathcal{M} = \mathcal{D}$ and all three kinds of measurability coincide with ordinary measurability. If $\mathcal{D}$ is a δ-ring which is not a σ-algebra, strict $\mathcal{D}$-measurability is the most important of the three: it means that the function is measurable and supported on a set in $\mathcal{D}$, or equivalently that it is supported on a set $N \in \mathcal{D}$ and its restriction to N is measurable in the usual sense with respect to the σ-algebra $\mathcal{D}_N$ of subsets of N. These are functions which can potentially be integrated ().

XVI.2.3.5. For a δ-ring $\mathcal{D}$ which is not a σ-ring, there are two other occasionally useful notions:

A function $f : X \to \mathbb{R}$ is *σ-strictly $\mathcal{D}$-measurable* if it is strictly measurable with respect to the σ-ring generated by $\mathcal{D}$.

A function $f : X \to \mathbb{R}$ is *almost strictly $\mathcal{D}$-measurable* if it is $\mathcal{D}$-measurable and, for every $\epsilon > 0$, $\{x \in X : |f(x)| \geq \epsilon\} \in \mathcal{D}$.

We have (strictly $\mathcal{D}$-measurable) $\Rightarrow$ (almost strictly $\mathcal{D}$-measurable) $\Rightarrow$ (σ-strictly $\mathcal{D}$-measurable) $\Rightarrow$ ($\mathcal{D}$-measurable) $\Rightarrow$ (locally $\mathcal{D}$-measurable). If $\mathcal{D}$ is a σ-ring, the first three coincide, but if $\mathcal{D}$ is not a σ-ring they are distinct. For suppose (A_n) is a sequence of sets in $\mathcal{D}$ for which $A := \cup_n A_n \notin \mathcal{D}$. Then

$\sum_{n=1}^{\infty} 2^{-n} \chi_{A_n}$ is almost strictly $\mathcal{D}$-measurable but not strictly $\mathcal{D}$-measurable.

χ_A is σ-strictly $\mathcal{D}$-measurable but not almost strictly $\mathcal{D}$-measurable.

Almost strictly $\mathcal{D}$-measurable functions can also potentially be integrated ().

XVI.2.3.6. The various types of measurability can be defined analogously for extended real-valued and complex-valued functions. A complex-valued function is measurable in any of the senses if and only if both its real and imaginary parts are measurable in the same sense.

XVI.2.4. Borel Measurable Functions

Along with the Borel sets themselves, the Borel measurable functions are basic objects for study in measure theory.

XVI.2.4.1. Definition. If $(X, \mathcal{T})$ is a topological space, a function $f : X \to \mathbb{R}$ is *Borel measurable* if it is $\mathcal{B}_{\mathcal{T}}$-measurable, i.e. the inverse image of every Borel set in $\mathbb{R}$ is a Borel set in $(X, \mathcal{T})$. Denote by $\mathfrak{B}_{\mathcal{T}}$ the set of Borel measurable functions from X to $\mathbb{R}$. Write $\mathfrak{B}$ for the set of Borel measurable functions from $\mathbb{R}$ to $\mathbb{R}$. We may similarly define Borel measurability for functions from X to $\mathbb{R}^n$, and more generally for functions from X to any topological space $(Y, \mathcal{S})$.

XVI.2.4.2. We most commonly use the term "Borel measurable" for functions from $\mathbb{R}$ to $\mathbb{R}$ (or from $\mathbb{R}^n$ to $\mathbb{R}^m$). We will show that all "reasonable" functions from $\mathbb{R}$ to $\mathbb{R}$ are Borel measurable. In fact, roughly speaking, any function from $\mathbb{R}$ to $\mathbb{R}$ which can be defined or described without using the Axiom of Choice is Borel measurable; in particular, any function which can be defined by a countable number of algebraic, order, or pointwise limit operations is Borel measurable. (Strictly speaking, the first part of this statement is not true; it is possible, but quite complicated, to explicitly describe a subset of $\mathbb{R}$ which is not a Borel set, and hence a function which is not Borel measurable; see ().)

The next proposition is a special case of XVI.2.2.3.:

XVI.2.4.3. Proposition. Let $(X, \mathcal{T})$, $(Y, \mathcal{S})$, and $(Z, \mathcal{R})$ be topological spaces, $f : X \to Y$ and $g : Y \to Z$ Borel measurable functions. Then $g \circ f : X \to Z$ is Borel measurable.

XVI.2.2.4. applies to Borel measurability:

XVI.2.4.4. PROPOSITION. Let $f : \mathbb{R} \to \mathbb{R}$ a function. Then f is Borel measurable if and only if any of the following conditions are satisfied:

(i) $f^{-1}(U) \in \mathcal{B}$ for every open set $U \subseteq \mathbb{R}$.

(ii) $f^{-1}(E) \in \mathcal{B}$ for every closed set $E \subseteq \mathbb{R}$.

(iii) $f^{-1}(U) \in \mathcal{B}$ for every open interval $U \subseteq \mathbb{R}$.

(iv) $f^{-1}((-\infty, a)) \in \mathcal{B}$ for every $a \in \mathbb{R}$.

(v) $f^{-1}((-\infty, a]) \in \mathcal{B}$ for every $a \in \mathbb{R}$.

(vi) $f^{-1}((a, +\infty)) \in \mathcal{B}$ for every $a \in \mathbb{R}$.

(vii) $f^{-1}([a, +\infty)) \in \mathcal{B}$ for every $a \in \mathbb{R}$.

(viii) $f^{-1}((-\infty, a)) \in \mathcal{B}$ for every $a \in \mathbb{Q}$.

(ix) $f^{-1}((-\infty, a]) \in \mathcal{B}$ for every $a \in \mathbb{Q}$.

(x) $f^{-1}((a, +\infty)) \in \mathcal{B}$ for every $a \in \mathbb{Q}$.

(xi) $f^{-1}([a, +\infty)) \in \mathcal{B}$ for every $a \in \mathbb{Q}$.

XVI.2.4.5. COROLLARY. Every continuous function from a topological space $(X, \mathcal{T})$ to $\mathbb{R}$ is Borel measurable. In particular, every continuous function from $\mathbb{R}$ to $\mathbb{R}$ is Borel measurable.

PROOF: Apply (i): if f is continuous and $U \subseteq \mathbb{R}$ is open, then $f^{-1}(U)$ is open in $(X, \mathcal{T})$, and every open set is a Borel set. ☝

The next special case of XVI.1.6.11. (using XVI.2.4.5.) shows that "piecewise-continuous" functions are Borel measurable.

XVI.2.4.6. COROLLARY. Let $(X, \mathcal{T})$ be a topological space, and (E_n) a sequence of Borel subsets of X with $\cup_{n=1}^{\infty} E_n = X$. Let $f : X \to \mathbb{R}$ be a function. If $f|_{E_n}$ is continuous (regarding E_n as a topological space with the relative topology) for each n, then f is Borel measurable.

Many very discontinuous functions are Borel measurable. For example:

XVI.2.4.7. PROPOSITION. If $B \subseteq \mathbb{R}$, then χ_B is a Borel measurable function if and only if B is a Borel set in $\mathbb{R}$.

This is a special case of XVI.2.2.5..

From XVI.2.4.3., in combination with XVI.2.4.5., we obtain an important property of the set $\mathfrak{B}_{\mathcal{A}}$ of $\mathcal{A}$-measurable functions:

XVI.2.4.8. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space, $f \in \mathfrak{B}_{\mathcal{A}}$, and g a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$. Then $g \circ f$ is in $\mathfrak{B}_{\mathcal{A}}$. In particular, if $h : \mathbb{R} \to \mathbb{R}$ is continuous, then $h \circ f$ is in $\mathfrak{B}_{\mathcal{A}}$.

XVI.2.4.9. If f is a real-valued function, we define the *positive and negative parts* of f to be f_+ and f_-, where $f_+(x) = \max(f(x), 0)$ and $f_-(x) = \max(-f(x), 0)$. In other words, $f_+ = f$ wherever f is nonnegative, and is zero where f is negative; and $f_- = |f|$ where f is negative and zero where f is nonnegative. Then $f = f_+ - f_-$, $|f| = f_+ + f_-$, and $f_+ f_- = 0$.

XVI.2.4.10. COROLLARY. Let $(X, \mathcal{A})$ be a measurable space, and $f \in \mathfrak{B}_{\mathcal{A}}$. Then f^2, f_+, f_-, and $|f|$ are in $\mathfrak{B}_{\mathcal{A}}$. If $f \geq 0$, then $\sqrt{f} \in \mathfrak{B}_{\mathcal{A}}$.

PROOF: $f^2 = g \circ f$, where $g(t) = t^2$. Similarly, the other functions are the compositions of f with the functions $\max(t, 0)$, $\max(-t, 0)$, $|t|$, and $\sqrt{|t|}$ respectively. 👍

This corollary can also be easily proved directly from XVI.2.4.4..

Our next goal is to show that $\mathfrak{B}_{\mathcal{A}}$ is closed under the usual algebraic operations: sum, product, and scalar multiple. (We say that $\mathfrak{B}_{\mathcal{A}}$ is an *algebra* of functions.) The only step which takes some work is showing closure under addition.

XVI.2.4.11. THEOREM. Let $(X, \mathcal{A})$ be a measurable space, $f, g \in \mathfrak{B}_{\mathcal{A}}$. Then $f + g$ and fg are in $\mathfrak{B}_{\mathcal{A}}$. If $\alpha \in \mathbb{R}$, then $\alpha f \in \mathfrak{B}_{\mathcal{A}}$.

PROOF: Closure under scalar multiplication is obvious from XVI.2.4.4.(vi)-(viii).

We show closure under addition. Let $h = f + g$. To show that $h \in \mathfrak{B}_{\mathcal{A}}$, then by XVI.2.4.4.(vi) it suffices to show that $h^{-1}((a, \infty)) \in \mathcal{A}$ for every $a \in \mathbb{R}$. But if a is fixed, then

$$\begin{aligned} h^{-1}((a, \infty)) &= \{x : f(x) > a - g(x)\} \\ &= \{x : f(x) > r > a - g(x) \text{ for some } r \in \mathbb{Q}\} \\ &= \{x : f(x) > r \text{ and } g(x) > a - r \text{ for some } r \in \mathbb{Q}\} \\ &= \bigcup_{r \in \mathbb{Q}} \left[f^{-1}((r, \infty)) \cap g^{-1}((a - r, \infty)) \right]. \end{aligned}$$

Since f and g are $\mathcal{A}$-measurable, and $\mathbb{Q}$ is countable, the last set is in $\mathcal{A}$.

A trick using XVI.2.4.10. gives closure under multiplication:

$$fg = \frac{1}{2}[(f + g)^2 - f^2 - g^2].$$

👍

Measurable Functions to $\mathbb{R}^n$

As previously noted, we can define $\mathcal{A}$-measurability for functions from a measurable space $(X, \mathcal{A})$ to $\mathbb{R}^n$. We do not get anything essentially new by doing this, however:

XVI.2.4.12. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space, and $f : X \to \mathbb{R}^n$ a function. Write $f_k : X \to \mathbb{R}$ $(1 \leq k \leq n)$ for the k'th coordinate functions of f, i.e. $f(x) = (f_1(x), \ldots, f_n(x))$ for $x \in X$. Then f is an $\mathcal{A}$-measurable function from X to $\mathbb{R}^n$ if and only if each f_k is an $\mathcal{A}$-measurable function from X to $\mathbb{R}$.

PROOF: We have $f_k = \pi_k \circ f$, where $\pi_k : \mathbb{R}^n \to \mathbb{R}$ is the projection onto the k'th coordinate. Since π_k is continuous, f_k is $\mathcal{A}$-measurable if f is $\mathcal{A}$-measurable by XVI.2.4.8.. Conversely, suppose each f_k is $\mathcal{A}$-measurable. Let $B_1, \ldots, B_n$ be Borel sets in $\mathbb{R}$, and $B = B_1 \times B_2 \times \cdots \times B_n$ the corresponding measurable rectangle in $\mathbb{R}^n$. Then $f^{-1}(B) = f_1^{-1}(B_1) \cap \cdots \cap f_n^{-1}(B_n)$. Since $f_k^{-1}(B_k) \in \mathcal{A}$ for each k, $f^{-1}(B) \in \mathcal{A}$. Measurable rectangles generate the σ-algebra of Borel sets in $\mathbb{R}^n$ (), so f is $\mathcal{A}$-measurable by XVI.1.6.9.. 👍

XVI.2.5. Another Characterization of Borel Sets and Borel Measurable Functions

It turns out that the Borel measurable functions have a simple and useful alternate characterization as the "sequential completion" of the class of continuous functions. The results of this section can be generalized to functions on arbitrary topological spaces, but we will stick to functions from $\mathbb{R}$ to $\mathbb{R}$ for simplicity.

XVI.2.5.1. DEFINITION. Let $\mathfrak{C}$ be the set of continuous functions from $\mathbb{R}$ to $\mathbb{R}$. Let $\bar{\mathfrak{C}}$ be the smallest class of functions from $\mathbb{R}$ to $\mathbb{R}$ containing $\mathfrak{C}$ and which is closed under pointwise sequential limits, i.e. such that if (f_n) is a sequence in $\bar{\mathfrak{C}}$ and $f_n \to f$ pointwise, then $f \in \bar{\mathfrak{C}}$.

A subset A of $\mathbb{R}$ is a *BC set* if $\chi_A \in \bar{\mathfrak{C}}$.

The abbreviation BC stands for "Baire class."

The definition of $\bar{\mathfrak{C}}$ makes sense as the intersection of all classes of functions with the two properties. This collection of classes is nonempty since the class of all functions from $\mathbb{R}$ to $\mathbb{R}$ has the properties.

We will show that the functions in $\bar{\mathfrak{C}}$ are precisely the Borel measurable functions, and the BC sets are precisely the Borel sets. Such sets and functions are very natural to consider in measure and integration, where pointwise limits often arise.

XVI.2.5.2. **Caution:** It is not true that $\bar{\mathfrak{C}}$ is just the set $\mathfrak{BC}_1$ of pointwise limits of sequences of continuous functions. This set $\mathfrak{BC}_1$ is contained in $\bar{\mathfrak{C}}$, but is not closed under pointwise sequential limits (although this is not obvious; see Exercise XVI.2.10.1..) Let $\mathfrak{BC}_2$ be the set of pointwise limits of sequences of functions in $\mathfrak{BC}_1$, and inductively for each ordinal α let $\mathfrak{BC}_{\alpha+1}$ be the set of pointwise limits of sequences of functions in $\mathfrak{BC}_\alpha$, with $\mathfrak{BC}_\alpha = \cup_{\beta<\alpha}\mathfrak{BC}_\beta$ when α is a limit ordinal. Then $\bar{\mathfrak{C}}$ is the union of the $\mathfrak{BC}_\alpha$ over all countable α, and it turns out that $\bar{\mathfrak{C}} \neq \mathfrak{BC}_\alpha$ for any countable α (XVI.2.7.2.). The functions in $\mathfrak{BC}_\alpha$ are said to be of *Baire class* α. ✋

XVI.2.6. Algebraic Properties of $\bar{\mathfrak{C}}$

We first show that, like $\mathfrak{B}_{\mathcal{A}}$ (XVI.2.4.11.), $\bar{\mathfrak{C}}$ is closed under the usual algebraic operations: sum, product, and scalar multiple (i.e. $\bar{\mathfrak{C}}$ is an *algebra* of functions.) It is a standard fact from Calculus that $\mathfrak{C}$ is closed under these operations. It can be easily shown by transfinite induction that each $\mathfrak{BC}_\alpha$ is also closed under the operations, and hence so is $\bar{\mathfrak{C}}$. A cleaner and more elementary proof can be given, however:

XVI.2.6.1. Proposition. Let $f, g \in \bar{\mathfrak{C}}$ and $\alpha \in \mathbb{R}$. Then $f+g$, fg, and αf are all in $\bar{\mathfrak{C}}$.

Proof: First fix $g \in \mathfrak{C}$, and let $\mathfrak{D} = \{f \in \bar{\mathfrak{C}} : f+g \in \bar{\mathfrak{C}}\}$. We claim that $\mathfrak{D} = \bar{\mathfrak{C}}$. Clearly $\mathfrak{D} \subseteq \bar{\mathfrak{C}}$, so we need only show that $\mathfrak{C} \subseteq \mathfrak{D}$ and that $\mathfrak{D}$ is closed under pointwise sequential limits. If $f \in \mathfrak{C}$, then $f+g \in \mathfrak{C} \subseteq \bar{\mathfrak{C}}$, so $f \in \mathfrak{D}$. And if (f_n) is a sequence in $\mathfrak{D}$ (i.e. $f_n \in \bar{\mathfrak{C}}$ and $f_n + g \in \bar{\mathfrak{C}}$), and $f_n \to f$ pointwise, then $f_n \in \bar{\mathfrak{C}}$ for all n, so $f \in \bar{\mathfrak{C}}$; also, $f_n + g \to f+g$ pointwise, and $f_n + g \in \bar{\mathfrak{C}}$, so $f+g \in \bar{\mathfrak{C}}$, i.e. $f \in \mathfrak{D}$. This proves the claim, and therefore the sum of a function in $\mathfrak{C}$ and a function in $\bar{\mathfrak{C}}$ is in $\bar{\mathfrak{C}}$.

Now suppose $g \in \bar{\mathfrak{C}}$, and repeat the argument with $\mathfrak{E} = \{f \in \bar{\mathfrak{C}} : f+g \in \bar{\mathfrak{C}}\}$. As before, we show that $\mathfrak{C} \subset \mathfrak{E}$ (by the first part of the proof) and that $\mathfrak{E}$ is closed under pointwise sequential limits, and thus $\mathfrak{E} = \bar{\mathfrak{C}}$. This shows that $\bar{\mathfrak{C}}$ is closed under addition.

The proof that $\bar{\mathfrak{C}}$ is closed under multiplication is virtually identical, replacing "sum" by "product" everywhere, and is left as an exercise ().

The case of scalar multiplication is a special case of the product, regarding the scalar α as the constant function with value α. 👍

XVI.2.6.2. Corollary. The class of BC sets in $\mathbb{R}$ is a σ-algebra.

Proof: Suppose A is a BC set, i.e. $\chi_A \in \bar{\mathfrak{C}}$. Then $\chi_{A^c} = 1 - \chi_A \in \bar{\mathfrak{C}}$, so A^c is a BC set. If A and B are BC sets, then $\chi_{A\cap B} = \chi_A \chi_B \in \bar{\mathfrak{C}}$, so $A \cap B$ is a BC set. Finally, if (A_n) is a sequence of BC sets and $A = \cap_{n=1}^{\infty} A_n$, set $f_n = \chi_{A_1 \cap \cdots \cap A_n} = \chi_{A_1}\chi_{A_2}\cdots\chi_{A_n}$. Then $f_n \in \bar{\mathfrak{C}}$ and $f_n \to \chi_A$ pointwise, so $\chi_A \in \bar{\mathfrak{C}}$ and so A is a BC set. Thus the class of BC sets is closed under complements and countable intersections. It clearly contains $\mathbb{R}$ since $\chi_{\mathbb{R}}$ is the constant function 1, which is in $\mathfrak{C}$. 👍

XVI.2.6.3. Proposition. Every open interval in $\mathbb{R}$ is a BC set.

Proof: Let (a,b) be an open interval in $\mathbb{R}$. We need to show that $\chi_{(a,b)} \in \bar{\mathfrak{C}}$; we will find a sequence of continuous functions converging pointwise to $\chi_{(a,b)}$. Let f_n be the function with the following graph:

It is easy to check that $f_n \to \chi_{(a,b)}$ pointwise. 👍

XVI.2.6.4. Corollary. Every Borel set is a BC set.

Proof: The class of BC sets is a σ-algebra (XVI.2.6.2.) which contains the open intervals (XVI.2.6.3.). The smallest such σ-algebra is the Borel sets. 👍

Now we show that $\bar{\mathfrak{C}}$ is closed under composition:

XVI.2.6.5. PROPOSITION. Let $f, g \in \bar{\mathfrak{C}}$. Then $f \circ g \in \bar{\mathfrak{C}}$.

PROOF: The proof is very similar to the proof of XVI.2.6.1.. First fix $f \in \mathfrak{C}$, and let $\mathfrak{D} = \{g \in \bar{\mathfrak{C}} : f \circ g \in \bar{\mathfrak{C}}\}$. As in XVI.2.6.1., $\mathfrak{C} \subseteq \mathfrak{D}$ since a composition of continuous functions is continuous. Also, if $g_n \in \mathfrak{D}$ and $g_n \to g$ pointwise, and $x \in \mathbb{R}$, if $y_n = g_n(x)$ and $y = g(x)$ we have that $y_n \to y$, and thus $[f \circ g_n](x) = f(y_n) \to f(y) = [f \circ g](x)$ since f is continuous; thus $f \circ g_n \to f \circ g$ and so $f \circ g \in \bar{\mathfrak{C}}$, $f \in \mathfrak{D}$. Thus $\mathfrak{D} = \bar{\mathfrak{C}}$, i.e. $f \circ g \in \bar{\mathfrak{C}}$ if $f \in \mathfrak{C}$, $g \in \bar{\mathfrak{C}}$.

Now fix $g \in \bar{\mathfrak{C}}$ and let $\mathfrak{E} = \{f \in \bar{\mathfrak{C}} : f \circ g \in \bar{\mathfrak{C}}\}$. By the first part of the proof, $\mathfrak{C} \subseteq \mathfrak{E}$. Also, if (f_n) is a sequence in $\mathfrak{E}$ converging pointwise to f, then $f \in \bar{\mathfrak{C}}$, and $f_n \circ g \to f \circ g$ pointwise (this is true with no restriction on g), and thus $f \circ g \in \bar{\mathfrak{C}}$, $f \in \mathfrak{E}$. Thus $\mathfrak{E} = \bar{\mathfrak{C}}$ and the proposition is proved. 👍

XVI.2.6.6. COROLLARY. If $g \in \bar{\mathfrak{C}}$, then $|g| \in \bar{\mathfrak{C}}$.

PROOF: We have that $|g| = f \circ g$, where $f(x) = |x|$, which is a continuous function. 👍

See problem () for another proof of this result.

A very important consequence of this corollary comes from the following lemma.

XVI.2.6.7. LEMMA. If $x, y \in \mathbb{R}$, then $\max(x, y) = \frac{1}{2}(x + y + |x - y|)$ and $\min(x, y) = \frac{1}{2}(x + y - |x - y|)$. Thus, if f and g are functions from $\mathbb{R}$ to $\mathbb{R}$, then $\max(f, g) = \frac{1}{2}(f+g+|f-g|)$ and $\min(f, g) = \frac{1}{2}(f+g-|f-g|)$.

PROOF: The proof is a straightforward case-by-case verification which is left as an exercise. 👍

XVI.2.6.8. COROLLARY. If f and g are in $\bar{\mathfrak{C}}$, then so are $\max(f, g)$ and $\min(f, g)$.

XVI.2.6.9. A sequence (f_n) of functions from $\mathbb{R}$ to $\mathbb{R}$ is *pointwise bounded* if, for every $x \in \mathbb{R}$, the set $\{f_n(x) : n \in \mathbb{N}\}$ is bounded in $\mathbb{R}$. If (f_n) is pointwise bounded, then $\sup_n f_n$ and $\inf_n f_n$ are defined as functions from $\mathbb{R}$ to $\mathbb{R}$: $[\sup_n f_n](x) = \sup\{f_n(x) : n \in \mathbb{N}\}$, $[\inf_n f_n](x) = \inf\{f_n(x) : n \in \mathbb{N}\}$.

XVI.2.6.10. COROLLARY. Let (f_n) be a pointwise bounded sequence in $\bar{\mathfrak{C}}$. Then $\sup_n f_n$ and $\inf_n f_n$ are in $\bar{\mathfrak{C}}$.

PROOF: Let $g_n = \max(f_1, \ldots, f_n)$ and $h_n = \min(f_1, \ldots, f_n)$. Then $g_n, h_n \in \bar{\mathfrak{C}}$ by XVI.2.6.8., and $g_n \to \sup_n f_n$, $h_n \to \inf_n f_n$ pointwise. 👍

In fact, it turns out that any algebra of functions which is closed under pointwise sequential limits is also closed under countable suprema and infima (Problem ().)

A clever trick yields a converse to this:

XVI.2.6.11. PROPOSITION. Let $\mathfrak{F}$ be a set of real-valued functions on a set X. If $\mathfrak{F}$ is closed under suprema and infima of pointwise bounded sequences, then $\mathfrak{F}$ is closed under pointwise sequential limits.

PROOF: Suppose (f_n) is a sequence in $\mathfrak{F}$, and $f_n \to f$ pointwise. Then, for each $x \in X$, the sequence $(f_n(x))$ is convergent and therefore bounded; thus (f_n) is pointwise bounded. Also, for $x \in X$,

$$f(x) = \lim_n f_n(x) = \limsup_n f_n(x) = \inf_n [\sup_{k \geq n} f_k(x)]$$

and so, if $g_n = \sup_{k \geq n} f_k$, then $g_n \in \mathfrak{F}$ for all n, and therefore $f = \inf_n [\sup_{k \geq n} f_k] = \inf_n g_n \in \mathfrak{F}$. 👍

This applies in particular to the collection of $\mathcal{A}$-measurable real-valued functions on any measurable space $(X, \mathcal{A})$:

XVI.2.6.12. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space, and $\mathfrak{B}_{\mathcal{A}}$ the set of $\mathcal{A}$-measurable real-valued functions. Then $\mathfrak{B}_{\mathcal{A}}$ is closed under suprema and infima of pointwise bounded sequences, and hence under pointwise sequential limits.

PROOF: Let (f_n) be a locally bounded sequence in $\mathfrak{B}_{\mathcal{A}}$, and let $g = \sup_n f_n$. We will show that $g^{-1}((a, +\infty)) \in \mathcal{A}$ for every $a \in \mathbb{R}$, and use XVI.2.4.4.(vi). If $x \in X$, then $g(x) > a$ if and only if $f_n(x) > a$ for some n, and thus $g^{-1}((a, +\infty)) = \cup_{n=1}^{\infty} f_n^{-1}((a, +\infty))$. By assumption, $f_n^{-1}((a, +\infty)) \in \mathcal{A}$ for all n, so $g^{-1}((a, +\infty))$ is a countable union of sets in $\mathcal{A}$.

The proof for $\inf_n f_n$ is nearly identical, using XVI.2.4.4.(iv). 👍

The set of $\mathcal{A}$-measurable functions on $(X, \mathcal{A})$ is also an algebra (XVI.2.4.11.).

As a consequence, we get a relationship between $\bar{\mathfrak{C}}$ and Borel measurability:

XVI.2.6.13. COROLLARY. Every function in $\bar{\mathfrak{C}}$ is Borel measurable.

PROOF: If $\mathfrak{B}$ is the set of Borel measurable functions, then $\mathfrak{B}$ contains $\mathfrak{C}$ by XVI.2.4.5., and is closed under pointwise sequential limits by XVI.2.6.12.; thus $\mathfrak{B}$ contains $\bar{\mathfrak{C}}$. 👍

XVI.2.6.14. COROLLARY. Every BC set is a Borel set.

PROOF: If A is a BC set, then $\chi_A \in \bar{\mathfrak{C}}$, and hence χ_A is Borel measurable. But this implies that A is a Borel set by XVI.2.2.5.. 👍

Combining XVI.2.6.14. with XVI.2.6.4., we obtain one of the principal facts we needed to prove:

XVI.2.6.15. THEOREM. A subset of $\mathbb{R}$ is a Borel set if and only if it is a BC set.

XVI.2.6.16. Corollary. A simple function is Borel measurable if and only if it is a BC function.

It still remains to prove that every Borel measurable function is in $\bar{\mathfrak{C}}$. Since every simple Borel measurable function is in $\bar{\mathfrak{C}}$ (XVI.2.6.16.), it suffices to show:

XVI.2.6.17. Lemma. Every Borel measurable function is a pointwise limit of simple Borel measurable functions.

This is a special case of XVI.2.2.9..

So we have shown:

XVI.2.6.18. Theorem. $\bar{\mathfrak{C}} = \mathfrak{B}$.

XVI.2.7. The Borel Sets Are Not Countably Constructible

In this section, we will show that none of the Baire classes $\mathfrak{BC}_\alpha$ for countable α are equal to the class of all Borel measurable functions, and hence the set $\mathcal{BC}_\alpha$ of sets of Baire class α (for α countable) is not the entire class of Borel sets. Thus the classes $\mathfrak{BC}_\alpha$ and $\mathcal{BC}_\alpha$ are strictly increasing in α for α countable, i.e. the entire class of Borel sets or Borel measurable functions cannot be constructed in finitely, or even countably, many steps.

This result was proved by Lebesgue in his 1905 paper [**?**], and was the beginning of what today is called *descriptive set theory* (). The following slick refinement of Lebesgue's original argument, due to W. Sierpiński, which uses an "ultimate space-filling curve," is adapted from [vRS82]. For a rather different argument, see [Bil95, p. 30-32].

XVI.2.7.1. Definition. Let $\mathfrak{F}$ be a set of Borel measurable functions from $\mathbb{R}$ to $\mathbb{R}$. A *catalog* for $\mathfrak{F}$ is a Borel measurable function $F : [0,1] \times \mathbb{R} \to \mathbb{R}$ such that, for every $f \in \mathfrak{F}$, there is an $s \in [0,1]$ for which $f(t) = F(s,t)$ for all $t \in \mathbb{R}$.

If $\mathfrak{F}$ has a catalog, then obviously so does any subset of $\mathfrak{F}$.

We will prove:

XVI.2.7.2. Theorem. For each countable α, $\mathfrak{BC}_\alpha$ has a catalog. But $\mathfrak{B}$ does not have a catalog. So $\mathfrak{BC}_\alpha \neq \mathfrak{B}$.

The proof will proceed in a number of steps.

XVI.2.7.3. Proposition. For each n, let $\mathfrak{F}_n$ be a set of Borel measurable functions, and let $\mathfrak{F} = \cup_n \mathfrak{F}_n$. If each $\mathfrak{F}_n$ has a catalog, then $\mathfrak{F}$ has a catalog.

Proof: Let F_n be a catalog for $\mathfrak{F}_n$. Let h_n be a linear (affine) function from $[\frac{1}{2n}, \frac{1}{2n-1}]$ onto $[0,1]$, and define $G_n : [0,1] \times \mathbb{R} \to \mathbb{R}$ by setting $G_n(s,t) = F_n(h_n(s),t)$ if $s \in [\frac{1}{2n}, \frac{1}{2n-1}]$ and $G_n(s,t) = 0$ otherwise. Then G_n is Borel measurable by () and (). The sequence of partial sums of $\sum_n G_n$ converges pointwise on $[0,1] \times \mathbb{R}$ to a function G which is thus Borel measurable; and G is a catalog for $\mathfrak{F}$ since, if $f \in \mathcal{F}_n$ and $f(t) = F_n(s,t)$, then $f(t) = G(h_n^{-1}(s),t)$. 👍

XVI.2.7.4. Corollary. Any countable set of Borel measurable functions has a catalog.

Proof: It is only necessary to observe that if f is a single Borel measurable function, then $\mathfrak{F} = \{f\}$ has a catalog. Set $F(s,t) = f(t)$ for all s. 👍

XVI.2.7.5. If $\mathfrak{F}$ is a set of Borel measurable functions, write $\mathfrak{F}^*$ for the set of functions (also Borel measurable) which are pointwise limits of sequences of functions in $\mathfrak{F}$. For example, $\mathfrak{BC}_\alpha^* = \mathfrak{BC}_{\alpha+1}$. In general, $\mathfrak{F} \subseteq \mathfrak{F}^*$ (use constant sequences).

The most difficult part of the proof of the theorem is to show that if $\mathfrak{F}$ has a catalog, so does $\mathfrak{F}^*$. This will be done using the continuous function ψ of XIII.18.4.4.. We first need a useful variation of the notion of limit of a sequence:

XVI.2.7.6. DEFINITION. Let (a_n) be a sequence of numbers. Define $\mathrm{Lim}_{n\to\infty} a_n = \lim_{n\to\infty} a_n$ if the limit exists, and $\mathrm{Lim}_{n\to\infty} a_n = 0$ if $\lim_{n\to\infty} a_n$ does not exist.

Thus $\mathrm{Lim}_{n\to\infty} a_n$ exists for every sequence (a_n) of numbers. If $(a_n) \subseteq [0,1]$, then $\mathrm{Lim}_{n\to\infty} a_n \in [0,1]$.

XVI.2.7.7. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space, and (f_n) a sequence of real-valued $\mathcal{A}$-measurable functions on X. Let $C = \{x \in X : (f_n(x)) \text{ converges}\}$. Then $C \in \mathcal{A}$.

PROOF: If $x \in X$, then $x \in C$ if and only if $(f_n(x))$ is a Cauchy sequence. Thus

$$C = \bigcap_{k\in\mathbb{N}} \bigcup_{r\in\mathbb{Q}} \bigcup_{n\in\mathbb{N}} \bigcap_{m\geq n} f_m^{-1}\left(\left[r - \frac{1}{k}, r + \frac{1}{k}\right]\right) .$$

👍

XVI.2.7.8. COROLLARY. Let $(X, \mathcal{A})$ be a measurable space, and (f_n) a sequence of real-valued $\mathcal{A}$-measurable functions on X. For $x \in X$, set $f(x) = \mathrm{Lim}_{n\to\infty} f_n(x)$. Then f is $\mathcal{A}$-measurable.

PROOF: If C is as above, then $C \in \mathcal{A}$, and $f_n \to f$ on C, so $f|_C$ is measurable. f is constant on $X \setminus C$, hence measurable. Thus f is measurable by (). 👍

XVI.2.7.9. THEOREM. If $\mathfrak{F}$ is a set of Borel measurable functions from $\mathbb{R}$ to $\mathbb{R}$, and $\mathfrak{F}$ has a catalog, then $\mathfrak{F}^*$ also has a catalog.

PROOF: Let F be a catalog for $\mathfrak{F}$, and let $\psi : [0,1] \to [0,1]^{\mathbb{N}}$ be an ultimate space-filling curve as in **??**, with coordinate functions ψ_n $(n \in \mathbb{N})$. For each $s \in [0,1]$, $t \in \mathbb{R}$, set $G_n(s,t) = F(\psi_n(s), t)$, and $G(s,t) = \mathrm{Lim}_{n\to\infty} G_n(s,t)$. Since $h_n(s,t) = (\psi_n(s), t)$ is continuous and hence Borel measurable, so is $G_n = F \circ h_n$, and therefore G is Borel measurable by XVI.2.7.8..

We claim G is a catalog for $\mathfrak{F}^*$. Suppose (f_n) is a sequence in $\mathfrak{F}$ with $f_n \to f$ pointwise. For each n there is an $s_n \in [0,1]$ such that $f_n(t) = F(s_n, t)$ for all t. There is an $s \in [0,1]$ such that $s_n = \psi_n(s)$ for all n (). Then $G(s,t) = f(t)$ for all $t \in \mathbb{R}$. 👍

We next show that $\mathfrak{BC}_\alpha$ has a catalog for α countable.

XVI.2.7.10. DEFINITION. If $\mathfrak{F}$ and $\mathfrak{G}$ are sets of Borel measurable functions, we say $\mathfrak{G}$ is *locally uniformly dense in* $\mathfrak{F}$ if, for every $f \in \mathfrak{F}$, $\epsilon > 0$, and $n \in \mathbb{N}$, there is a $g \in \mathfrak{G}$ with $|f(t) - g(t)| < \epsilon$ for all $t \in [-n, n]$.

We usually apply this definition when $\mathfrak{G} \subseteq \mathfrak{F}$.

XVI.2.7.11. EXAMPLE. Here is the most important example for us. Let $\mathfrak{Q}$ be the set of polynomials with rational coefficients. Then by the Weierstrass Approximation Theorem (), $\mathfrak{Q}$ is locally uniformly dense in the set $\mathfrak{C}$ of continuous functions from $\mathbb{R}$ to $\mathbb{R}$.

XVI.2.7.12. PROPOSITION. If $\mathfrak{G}$ is locally uniformly dense in $\mathfrak{F}$, then $\mathfrak{F}^* \subseteq \mathfrak{G}^*$. (If $\mathfrak{G} \subseteq \mathfrak{F}$, it follows that $\mathfrak{G}^* = \mathfrak{F}^*$.)

PROOF: Suppose $f_n \in \mathfrak{F}$ and $f_n \to f$ pointwise. For each n, choose $g_n \in \mathfrak{G}$ with $|f_n(t) - g_n(t)| < \frac{1}{n}$ for all $t \in [-n, n]$. Then it is easily checked that $g_n \to f$ pointwise. 👍

In particular, we have $\mathfrak{Q}^* = \mathfrak{C}^* = \mathfrak{BC}_1$.

XVI.2.7.13. PROPOSITION. $\mathfrak{BC}_\alpha$ has a catalog for every countable α.

PROOF: $\mathfrak{Q}$ is countable, so it has a catalog by **??**. Thus $\mathfrak{Q}^* = \mathfrak{BC}_1$ has a catalog by XVI.2.7.9..

Now proceed by transfinite induction. If $\mathfrak{BC}_\alpha$ has a catalog, then $\mathfrak{BC}_{\alpha+1}$ has a catalog by XVI.2.7.9.. And if α is a countable limit ordinal, then $\mathfrak{BC}_\alpha = \cup_{\beta<\alpha}\mathfrak{BC}_\beta$, a countable union, so $\mathfrak{BC}_\alpha$ has a catalog by **??**. 👍

The next result finishes the proof of XVI.2.7.2.. The proof is strongly reminiscent of Cantor's diagonalization argument.

XVI.2.7.14. PROPOSITION. $\mathfrak{B}$ does not have a catalog.

PROOF: Suppose F is a catalog for $\mathfrak{B}$. Define $f : \mathbb{R} \to \mathbb{R}$ by $f(t) = F(t,t) + 1$ if $t \in [0,1]$, and $f(t) = 0$ otherwise. Then $f|_{[0,1]}$ is a composition of two Borel measurable functions, hence Borel measurable; so f is Borel measurable by (). Since F is a catalog, there is an $s \in [0,1]$ such that $f(t) = F(s,t)$ for all t. But $f(s) = F(s,s) + 1 \neq F(s,s)$, a contradiction. 👍

This completes the proof of XVI.2.7.2.. We have the following consequence:

XVI.2.7.15. COROLLARY. For any countable α, $\mathcal{BC}_\alpha$ is a proper subset of $\mathcal{B}$. Thus the $\mathcal{BC}_\alpha$ are strictly increasing for α countable.

PROOF: Suppose $\mathcal{BC}_\alpha = \mathcal{B}$ for some countable α. Since $\mathfrak{BC}_\alpha$ is an algebra of functions (), it follows that every simple Borel measurable function is in $\mathfrak{BC}_\alpha$. Hence every Borel measurable function is in $\mathfrak{BC}_{\alpha+1}$ by (), a contradiction.

If $\alpha < \beta < \omega_1$ and $\mathcal{BC}_\alpha = \mathcal{BC}_\beta$, any countable intersection of sets in $\mathcal{BC}_\alpha$ is in $\mathcal{BC}_{\alpha+1} \subseteq \mathcal{BC}_\beta = \mathcal{BC}_\alpha$ by (), and so it follows from () that $\mathcal{BC}_\alpha$ is a σ-algebra and hence $\mathcal{BC}_\alpha = \mathcal{B}$, a contradiction. 👍

We then have *a fortiori* that $\mathfrak{BC}_\alpha \neq \mathfrak{BC}_\beta$ for $\alpha < \beta < \omega_1$.

XVI.2.8. The Multiplicity Function

In many settings where a function is being analyzed, it is important to consider the "size" of inverse images of points; this provides, among other things, a description of how far the function is from being one-to-one. The most basic way to do this is via the "multiplicity function" (sometimes called the "counting function"):

XVI.2.8.1. DEFINITION. Let $f : X \to Y$ be a function. The *multiplicity function* of f is the function $N_f : Y \to \{0\} \cup \mathbb{N} \cup \{+\infty\}$ defined by

$$N_f(y) = \begin{cases} card(f^{-1}(\{y\})) & \text{if} \quad f^{-1}(\{y\}) \text{ is finite} \\ +\infty & \text{if} \quad f^{-1}(\{y\}) \text{ is infinite} \end{cases} .$$

The Multiplicity Function is most commonly used when $Y = \mathbb{R}$ (or $\mathbb{R}^n$). Note that in this function, we do not distinguish between cases where $f^{-1}(\{y\})$ is countable and cases where it is uncountable. We can sometimes take this into account by a finer invariant; cf. ().

XVI.2.8.2. EXAMPLES. (i) N_f is the constant function 1 if and only if f is a bijection. If f is one-to-one, then N_f takes only the values 0 and 1.
(ii) If $f : \mathbb{R} \to \mathbb{R}$ is given by $f(x) = x^2$, then

$$N_f(y) = \begin{cases} 0 & \text{if} \quad y < 0 \\ 1 & \text{if} \quad y = 0 \\ 2 & \text{if} \quad y > 0 \end{cases} .$$

(iii) If $f : [0,1] \to [0,1]$ is a coordinate function of a space-filling curve (), then f is continuous and $N_f(y) = +\infty$ for all $y \in [0,1]$ ().

XVI.2.8.3. If $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are measurable spaces and $f : X \to Y$ is measurable, it is not true in general (cf. XVI.2.8.7.) even if $X = Y = \mathbb{R}$, that N_f is a measurable function (since its range is a countable set, N_f will be measurable if and only if, for each $n \in \{0\} \cup \mathbb{N} \cup \{+\infty\}$ we have that $\{y \in Y : N_f(y) = n\}$ is in $\mathcal{B}$). But in some important special cases this can be shown. The next result is the simplest example, and perhaps the most widely used; the proof can be adapted to broader contexts.

XVI.2.8.4. PROPOSITION. Let f be a real-valued continuous function on an an interval I in $\mathbb{R}$. Then N_f is an extended real-valued Borel measurable function.

PROOF: First suppose I is a bounded interval from a to b (with or without endpoints). For each $n \in \mathbb{N}$, define a function $\phi_n : \mathbb{R} \to \mathbb{N}$ by taking a partition $\{x_0 = a, x_1, \ldots, x_{2^n} = b\}$ of I into 2^n subintervals $J_1, \ldots, J_{2^n}$ of equal length, with $x_k \in J_k$, $x_k \notin J_{k+1}$ for $1 \leq k \leq 2^n - 1$, $a \in J_1$ if $a \in I$, and $b \in J_{2^n}$ if $b \in I$. Then set

$$\phi_n = \sum_{k=1}^{2^n} \chi_{f(J_k)} .$$

We have that $f(J_k)$ is an interval, hence a Borel set; thus ϕ_n is a Borel measurable function from $\mathbb{R}$ to $\mathbb{R}$. For any $y \in \mathbb{R}$, $\phi_n(y)$ is the number of subintervals whose intersection with $f^{-1}(\{y\})$ is nonempty. It is easily seen that $\phi_n \leq \phi_{n+1}$ for each n, and that $N_f = \sup_n \phi_n$. Thus N_f is Borel measurable.

For a general I, write $I = \cup I_m$, where each I_m is a bounded interval and $I_m \subseteq I_{m+1}$. Let $f_m = f|_{I_m}$. Then $N_f = \sup_m N_{f_m}$, and each f_m is Borel measurable, hence so is N_f. ☝

XVI.2.8.5. We can try to adapt this argument to a more general setting of a Borel measurable function f (e.g. a continuous function) between topological spaces X and Y. There are two things which need to be done:

(i) We need a sequence of finite partitions of X, each refining the previous one, into "nice" subsets (e.g. Borel sets), such that distinct elements of X are separated by sufficiently fine partitions in the sequence. This can be done systematically using Borel sets in any Polish space (), hence in any standard Borel space (), e.g. in any Borel subset of $\mathbb{R}^n$; for Borel sets in $\mathbb{R}^n$ there is an obvious approach using grids similar to the approach in the proof of XVI.2.8.4..

(ii) We need to know that if $J_1, \ldots, J_r$ are the sets in the n'th partition, then $f(J_k)$ is a Borel subset of Y. This is false in general (). However, if Y is a Polish space (e.g. $\mathbb{R}^n$), then $f(J_k)$ is an analytic subset of Y if J_k is Borel in X ().

Combining these observations with (), we obtain the most general result which can be easily stated:

XVI.2.8.6. THEOREM. Let $(X, \mathcal{A})$ be a standard Borel space (e.g. a Borel subset of $\mathbb{R}^n$), and $f : X \to \mathbb{R}^m$ a Borel measurable function. Then N_f is a Lebesgue measurable extended real-valued function on $\mathbb{R}^m$.

Lebesgue measurability of N_f can be replaced by μ-measurability, for any measure μ which is the completion of a Borel measure on $\mathbb{R}^m$, e.g. Hausdorff measure $\mathcal{H}^s$ for any s.

XVI.2.8.7. EXAMPLE. This result, which applies to Borel measurable functions from $\mathbb{R}$ to $\mathbb{R}$, cannot be extended to Lebesgue measurable functions from $\mathbb{R}$ to $\mathbb{R}$:

Let K be the usual Cantor set, and $K' \subseteq [0,1]$ a generalized Cantor set of positive measure. Let g be a strictly increasing continuous function from $[0,1]$ onto $[0,1]$ with $g(K) = K'$ (). Let Y be a subset of K' which is not Lebesgue measurable, and set $A = g^{-1}(Y)$. Then A has measure 0 since it is a subset of K. Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} g(x) & \text{if} \quad x \in A \\ 2 & \text{if} \quad x \notin A \end{cases} .$$

Then f is Lebesgue measurable (it equals the constant function 2 a.e.), but N_f is not Lebesgue measurable since

$$\{y \in \mathbb{R} : N_f(y) = 1\} = Y .$$

Note that although $f = 2$ a.e., we do not have $N_f = N_2$ a.e. In fact, N_2 (i.e. N_h, where h is the constant function 2) is Borel measurable (it is constant a.e.)

For continuous real-valued functions on a closed bounded interval, there is a close relationship between the Multiplicity Function and the notion of bounded variation (XIX.4.3.1.).

XVI.2.9. Standard Borel Spaces

XVI.2.9.1. DEFINITION. Let $(X, \mathcal{A})$ be a Borel space (measurable space). Then $(X, \mathcal{A})$ is a *standard Borel space* if there is a Polish topology $\mathcal{T}$ on X with $\mathcal{A} = \mathcal{B}_{\mathcal{T}}$.

XVI.2.9.2. EXAMPLES. (i) If X is any countable set, then $(X, \mathcal{P}(X))$ is a standard Borel space.
(ii) $(\mathbb{R}, \mathcal{B})$ is a standard Borel space.

We will see (XVI.2.9.9.) that these are the only ones, i.e. every standard Borel space is isomorphic to one (and only one) of these.

Borel Sets as Clopen Sets

The next remarkable theorem allows us to consider Borel sets in a standard Borel space as clopen sets with respect to a Polish topology.

XVI.2.9.3. THEOREM. Let $(X, \mathcal{T})$ be a Polish space (), and A a Borel set in $(X, \mathcal{T})$. Then there is a Polish topology $\mathcal{S}$ on X stronger than $\mathcal{T}$ such that $\mathcal{B}_{\mathcal{S}}(X) = \mathcal{B}_{\mathcal{T}}(X)$ and such that A is clopen in $(X, \mathcal{S})$.

PROOF: First note that if A is closed in $(X, \mathcal{T})$ and $\mathcal{S}$ is the topology on X generated by $\mathcal{T} \cup \{A\}$, then $(X, \mathcal{S})$ is homeomorphic to the separated union of A and A^c with their relative topologies, each of which is a Polish space, so $(X, \mathcal{S})$ is Polish. To see that $\mathcal{B}_{\mathcal{S}} = \mathcal{B}_{\mathcal{T}}$, first note that $\mathcal{B}_{\mathcal{T}} \subseteq \mathcal{B}_{\mathcal{S}}$. If $\{U_n : n \in \mathbb{N}\}$ is a countable base for $\mathcal{T}$, then $\{U_n \cap A, U_n \cap A^c : n \in \mathbb{N}\}$ is a countable base for $\mathcal{S}$, so every open set in $\mathcal{S}$ is a countable union of these, and is hence in $\mathcal{B}_{\mathcal{T}}$.
Now let $\mathcal{A}$ be the set of all $A \in \mathcal{B}_{\mathcal{T}}$ such that the conclusion holds for A. Then $\mathcal{A}$ contains all closed sets in $(X, \mathcal{T})$, so to show that $\mathcal{A}$ is all of $\mathcal{B}_{\mathcal{T}}$ it suffices to show $\mathcal{A}$ is a σ-algebra. $\mathcal{A}$ is clearly closed under complements, so we need only show $\mathcal{A}$ is closed under countable intersections. Let (A_n) be a sequence in $\mathcal{A}$. For each n let $\mathcal{T}_n$ be a Polish topology on X stronger than $\mathcal{T}$ such that $\mathcal{B}_{\mathcal{T}_n} = \mathcal{B}_{\mathcal{T}}$ and such that A_n is clopen in $(X, \mathcal{T}_n)$. If $\mathcal{T}_\infty$ is the weakest topology on X stronger than all the $\mathcal{T}_n$, then $\mathcal{T}_\infty$ is a Polish topology on X by XIII.10.5.4. and each A_n is closed (clopen) in $(X, \mathcal{T}_\infty)$. If for each n $\{U_k^{(n)} : k \in \mathbb{N}\}$ is a countable base for $(X, \mathcal{T}_n)$, then $\{U_k^{(n)} : k, n \in \mathbb{N}\}$ is a countable subbase for $\mathcal{T}_\infty$, and thus every open set in $(X, \mathcal{T}_\infty)$ is a countable union of finite intersections of these and is hence in $\mathcal{B}_{\mathcal{T}} = \mathcal{B}_{\mathcal{T}_n}$ for each n. So $\mathcal{B}_{\mathcal{T}_\infty} = \mathcal{B}_{\mathcal{T}}$.
Then $A = \cap_n A_n$ is closed in $\mathcal{T}_\infty$, so applying the first part of the proof there is a Polish topology $\mathcal{S}$ on X stronger than $\mathcal{T}_\infty$, with $\mathcal{B}_{\mathcal{S}} = \mathcal{B}_{\mathcal{T}_\infty} = \mathcal{B}_{\mathcal{T}}$ and A clopen in $(X, \mathcal{S})$. Thus $A \in \mathcal{A}$. 👍

With additional applications of XIII.10.5.4., the result can be extended:

XVI.2.9.4. COROLLARY. Let $(X, \mathcal{T})$ be a Polish space (), and (A_n) a sequence of Borel sets in $(X, \mathcal{T})$. Then there is a Polish topology $\mathcal{S}$ on X stronger than $\mathcal{T}$ such that $\mathcal{B}_{\mathcal{S}}(X) = \mathcal{B}_{\mathcal{T}}(X)$ and such that A_n is clopen in $(X, \mathcal{S})$ for each n. The topology $\mathcal{S}$ can be chosen so $(X, \mathcal{S})$ is zero-dimensional ().

PROOF: Apply XVI.2.9.3. to A_n to obtain $\mathcal{T}_n$ for each n, and apply XIII.10.5.4. to get $\mathcal{T}_\infty$. To make a zero-dimensional space, take a countable base for $\mathcal{T}_\infty$, extend the topology to make each basic open set clopen, and iterate, and apply XVI.2.9.3. once more at the end. 👍

XVI.2.9.5. This gets very close to XIII.10.5.5.: in the current language, this result implies that if $(X, \mathcal{T})$ is a perfect Polish space, there is a Polish topology $\mathcal{S}$ on X stronger than $\mathcal{T}$ such that $\mathcal{B}_{\mathcal{S}} = \mathcal{B}_{\mathcal{T}}$ and $(X, \mathcal{S})$ is homeomorphic to $\mathcal{N}$. Actually XVI.2.9.4. is very close to this, since it can be shown that $\mathcal{N}$ is the unique perfect zero-dimensional Polish space in which every compact subset has empty interior ().

There are several important corollaries:

XVI.2.9.6. Corollary. Let $(X, \mathcal{T})$ be a Polish space, and A a Borel set in X. If A is uncountable, then A contains a subset homeomorphic to the Cantor set, and in particular $card(A) = 2^{\aleph_0}$.

Proof: There is a Polish topology $\mathcal{S}$ on X stronger than $\mathcal{T}$ in which A is clopen, hence a Polish space in the relative topology. Then $(A, \mathcal{S}_A)$ contains a subset C homeomorphic to the Cantor set. Since C is compact in the relative $\mathcal{S}$-topology and Hausdorff in the relative $\mathcal{T}$-topology, the topologies coincide on C. 👍

XVI.2.9.7. Corollary. Let $(X, \mathcal{A})$ be a standard Borel space, If $Y \in \mathcal{A}$, then $(Y, \mathcal{A}_Y)$ is also a standard Borel space.

Proof: There is a Polish topology $\mathcal{T}$ on X such that $\mathcal{A} = \mathcal{B}_{\mathcal{T}}$ and Y is clopen. Then $(Y, \mathcal{T}_Y)$ is also Polish. We have $\mathcal{B}_Y = \{Y \cap A : A \in \mathcal{B}_X\} = \mathcal{A}_Y$ by (). 👍

XVI.2.9.8. Corollary. Let $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ be Polish spaces, and $f : X \to Y$ a measurable (Borel) function. Then there are zero-dimensional Polish topologies $\mathcal{S}_X$ and $\mathcal{S}_Y$ on X and Y, stronger than $\mathcal{T}_X$ and $\mathcal{T}_Y$ respectively with the same Borel sets, such that $f : (X, \mathcal{S}_X) \to (T, \mathcal{S}_Y)$ is continuous.

Proof: Let $\{V_n : n \in \mathbb{N}\}$ be a countable base for the topology of Y. Extend the topologies on Y and X to make each V_n and $f^{-1}(V_n)$ clopen. 👍

Uniqueness of Standard Borel Spaces

XVI.2.9.9. Theorem. Any two uncountable standard Borel spaces are isomorphic. Two standard Borel spaces are isomorphic if and only if they have the same cardinality, and the cardinality can only be countable or $2^{\aleph_0}$.

Proof: We need only prove the first statement. Let $(X, \mathcal{A})$ be an uncountable Borel space. Then there is a Polish topology $\mathcal{T}$ on X such that $\mathcal{A} = \mathcal{B}_{\mathcal{T}}$. By the Cantor-Bendixson Theorem (), X is the disjoint union of a countable discrete open set and a perfect closed subset. By extending $\mathcal{T}$, we may assume these subsets are clopen, i.e. X is the separated union of a countable discrete space and a perfect Polish space. Applying XIII.10.5.5., we may assume the perfect subspace is $\mathcal{N}$, i.e. $(X, \mathcal{T})$ is the separated union of $\mathcal{N}$ and a countable discrete space.

So it suffices to show that if X is such a space, then $(X, \mathcal{B}_X)$ is isomorphic to $(\mathcal{N}, \mathcal{B}_{\mathcal{N}})$. $(\mathcal{N}, \mathcal{B}_{\mathcal{N}})$ is clearly isomorphic to a Borel subspace (the perfect part) of $(X, \mathcal{B}_X)$. Conversely, let D be a countably infinite subset of $\mathcal{N}$, and $Y = D^c$. Then $(Y, \mathcal{B}_Y)$ is an uncountable standard Borel space by XVI.2.9.7., thus has

a topology $\mathcal{S}$ such that $(Y, \mathcal{S})$ is homeomorphic to a separated union of $\mathcal{N}$ and a countable discrete space. Thus $\mathcal{N} = D \cup Y$ has a stronger topology $\mathcal{S}$ with the same Borel sets homeomorphic to the separated union of $\mathcal{N}$ and a countable discrete space. Clearly $(X, \mathcal{B}_X)$ is isomorphic to a Borel subspace of $(\mathcal{N}, \mathcal{S})$. Apply the Schröder-Bernstein Theorem for measurable spaces (XVI.1.6.16.) to conclude $(X, \mathcal{B}_X)$ is isomorphic to $(\mathcal{N}, \mathcal{B}_\mathcal{N})$. 👍

XVI.2.10. Exercises

XVI.2.10.1. Let $\mathfrak{BC}_\alpha$ be the set of functions of Baire class α (). A subset A of $\mathbb{R}$ is a *set of Baire class* α if $\chi_A \in \mathfrak{BC}_\alpha$. Let $\mathcal{BC}_\alpha$ denote the set of subsets of $\mathbb{R}$ of Baire class α.
(a) Show that $\cup_{\alpha<\omega_1}\mathcal{BC}_\alpha$ is precisely the set of Borel sets.
(b) Show by transfinite induction that $\mathfrak{BC}_\alpha$ is an algebra of functions for each α.
(c) Show that each $\mathcal{BC}_\alpha$ is an algebra of sets. [Hint: $\chi_{A\cap B} = \chi_A\chi_B$.]
(d) If $f \in \mathfrak{BC}_1$ and U is an open set in $\mathbb{R}$, then $f^{-1}(U)$ is an F_σ. (The converse is also true: if $f^{-1}(U)$ is an F_σ for every open set U, then $f \in \mathfrak{BC}_1$ [].) [If $f_n \to f$ pointwise and $\alpha \in \mathbb{R}$, then

$$f^{-1}((\alpha,\infty)) = \bigcup_n \bigcup_m \bigcap_{k\geq m} \{t : f_k(t) \geq \alpha + \frac{1}{n}\}.]$$

(e) Show that $A \in \mathcal{BC}_1$ if and only if A is both an F_σ and a G_δ. [If A is both an F_σ and a G_δ, show that there is an increasing sequence (F_n) of closed sets, and a decreasing sequence (U_n) of open sets, such that $A = \cup_n F_n = \cap_n U_n$. If f_n is a continuous function which is 1 on F_n and 0 outside U_n (), then $f_n \to \chi_A$. Deduce the converse from (d).] In particular, any open or closed set is in $\mathcal{BC}_1$ ().
(f) Show that a countable union or countable intersection of sets in $\mathcal{BC}_\alpha$ is in $\mathcal{BC}_{\alpha+1}$. In particular, any F_σ or G_δ is in $\mathcal{BC}_2$.
(g) Since $\mathbb{Q}$ is not a G_δ (), $\mathbb{Q} \notin \mathcal{BC}_1$. But $\mathbb{Q} \in \mathcal{BC}_2$ since it is an F_σ.
(h) Use the argument of (d) to show that if $f \in \mathfrak{BC}_2$ and U is open, then $f^{-1}(U)$ is a $G_{\delta\sigma}$. Conclude that if A is in $\mathcal{BC}_2$, then A is both a $G_{\delta\sigma}$ and an $F_{\sigma\delta}$. Generalize to $\mathfrak{BC}_n$ and $\mathcal{BC}_n$.
In fact, if $\alpha < \beta$ are countable ordinals, then $\mathcal{BC}_\alpha$ is a proper subset of $\mathcal{BC}_\beta$, and hence $\mathcal{BC}_\alpha$ is a proper subset of the class of Borel sets (Appendix B).

XVI.2.10.2. Let f be any function from $\mathbb{R}$ to $\mathbb{R}$. Show that the set of points of continuity of f is a G_δ. [For each n, let U_n be the set of all $a \in \mathbb{R}$ for which there is a $\delta > 0$ such that $|f(x) - f(y)| < \frac{1}{n}$ whenever $|x - a| < \delta$ and $|y - a| < \delta$. Show that U_n is open.]

XVI.2.10.3. (a) If I is an interval, show that there is a Borel measurable $f : \mathbb{R} \to \mathbb{R}$ such that I is exactly the set of discontinuities for f. [If I is closed, consider $\chi_\mathbb{Q}\chi_I$. If I is open, let g be a continuous function which is nonzero precisely on I, and consider $\chi_\mathbb{Q} g$.]
(b) If B is an F_σ containing no interval, show there is a Borel measurable f whose set of discontinuities is precisely B. [Write $B = \cup F_n$, with F_n closed, and consider $\sum 2^{-n}\chi_{F_n}$.]
(c) Show that if A is a G_δ, then there is a Borel measurable $f : \mathbb{R} \to \mathbb{R}$ such that A is exactly the set of points where f is continuous.

XVI.2.10.4. Let $f : \mathbb{R} \to \mathbb{R}$ be continuous.
(a) For $n, k \in \mathbb{N}$, $c \in \mathbb{R}$, let $U_{n,k,c}$ be the set of all a for which there is an x with

$$0 < |x - a| < \frac{1}{n} \quad \text{and} \quad \frac{f(x) - f(a)}{x - a} > c - \frac{1}{k}$$

Show that $U_{n,k,c}$ is open and $\cap_{n,k} U_{n,k,c}$ is the set

$$V_c = \left\{ a \in \mathbb{R} : \limsup_{x \to a} \frac{f(x) - f(a)}{x - a} \geq c \right\}$$

So V_c is a G_δ.
(b) Similarly, if $d \in \mathbb{R}$, show that

$$W_d = \left\{ a \in \mathbb{R} : \liminf_{x \to a} \frac{f(x) - f(a)}{x - a} \leq d \right\}$$

is a G_δ.
(c) Show that the set D of points where f is differentiable is an $F_{\sigma\delta}$. [The complement of D is

$$\left(\bigcup_{\substack{c,d \in \mathbb{Q} \\ d<c}} V_c \cap W_d \right) \cup \left(\bigcap_{c \in \mathbb{Q}} V_c \right) \cup \left(\bigcap_{d \in \mathbb{Q}} W_d \right) .]$$

(d) What if f is not assumed to be continuous? (cf. XVI.2.10.2.)

XVI.2.10.5. Let f be a real-valued function on an interval I which has only countably many discontinuities. Show that f is Baire class 1. [Let (x_n) be a dense sequence in I containing all discontinuities of f, and let f_n be a piecewise-linear continuous function agreeing with f on $\{x_1, \ldots, x_n\}$. Show that $f_n \to f$ pointwise.]

XVI.2.10.6. Let I be an interval in $\mathbb{R}$, and $f : I \to \mathbb{R}$. Extend f to $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by setting $\tilde{f}$ identically 0 outside I. Show that $\tilde{f}$ is Baire class 1 on $\mathbb{R}$ if and only if f is Baire class 1 on I.

XVI.2.10.7. Let I be an interval in $\mathbb{R}$. Let $\mathcal{D}$ be a vector space of real-valued functions on I containing the constant functions, whose uniform closure $\bar{\mathcal{D}}$ has the property that if $f \in \mathcal{D}$, then $|f| \in \bar{\mathcal{D}}$, and let $\tilde{\mathcal{D}}$ be the set of all pointwise limits of sequences in $\mathcal{D}$. Show that $\tilde{\mathcal{D}}$ is closed under uniform limits, as follows.
(a) If f is a pointwise limit of a sequence in $\bar{\mathcal{D}}$, show that f is a pointwise limit of a sequence in $\mathcal{D}$. Thus we may, and will, assume $\mathcal{D}$ is uniformly closed.
(b) If $f \in \mathcal{D}$, then $f_+ = f \vee 0 = \frac{1}{2}(f + |f|) \in \mathcal{D}$. Show that if $a \leq b$, then $(f \vee a) \wedge b \in \mathcal{D}$.
(c) Let (f_n) a sequence in $\mathcal{D}$ which converges uniformly to f on I. Passing to a subsequence, we may assume that $|f_{n+1}(x) - f_n(x)| \leq 2^{-n-1}$ for all n and all $x \in I$. Set $g_1 = f_1$ and $g_{n+1} = f_{n+1} - f_n$ for $n \geq 1$. For each n let (h_{nk}) be a sequence in $\mathcal{D}$ with $h_{nk} \to g_n$ pointwise. Use (b) to show that we may assume $|h_{nk}(x)| \leq 2^{-n}$ for all $n > 1$, all k, and all $x \in I$.
(d) For each k, the infinite series $\sum_{n=1}^{\infty} h_{nk}$ converges uniformly. Let $h_k = \sum_{n=1}^{\infty} h_{nk}$. Then $h_k \in \mathcal{D}$. Show that $h_k \to f$ pointwise.

(e) The hypotheses are satisfied if $\mathcal{D}$ is an algebra of bounded functions on I and contains the constant functions (XX.10.2.9.).

(f) The hypotheses are satisfied if $\mathcal{D} = \mathcal{C}(I)$. Conclude that $\mathfrak{BC}_1(I)$ is uniformly closed.

(g) Show by induction on n that if $g \in \mathfrak{BC}_n(I)$ and f is continuous, then $f \circ g$ is in $\mathfrak{BC}_n(I)$. Conclude that $\mathcal{D} = \mathfrak{BC}_n(I)$ satisfies the hypotheses, and hence that $\mathfrak{BC}_{n+1}(I)$ is closed under uniform limits.

(h) Show that the result of the problem holds if "uniform" is replaced by "u.c." throughout. Thus the class $\mathfrak{BC}_n(I)$ is closed under u.c. limits for each $n \in \mathbb{N}$.

XVI.2.10.8. Let I be an interval in $\mathbb{R}$.

(a) Show that every step function on I is in $\mathfrak{BC}_1(I)$. [Use XVI.2.10.5..]

(b) Show that every regulated function (V.8.6.7.) on I is in $\mathfrak{BC}_1(I)$. [Use XVI.2.10.5. and XVI.2.10.7., or XVI.2.10.5. and V.8.6.7.(a).]

XVI.3. Analytic Sets

There is an important class of subsets of $\mathbb{R}$ (or more general topological spaces) which is slightly larger than the class of Borel sets.

XVI.3.1. Standard Borel Spaces

XVI.3.1.1. Recall that a topological space is a *Polish space* if it is homeomorphic to a separable complete metric space.

The next result says that, although there is an enormous variety of topologically distinct uncountable Polish spaces, they are all "identical" from a measure-theoretic standpoint:

XVI.3.1.2. THEOREM. Let X and Y be uncountable Polish spaces. Then the measurable spaces $(X, \mathcal{B}_X)$ and $(Y, \mathcal{B}_Y)$ are isomorphic (). In particular, $(X, \mathcal{B}_X)$ is isomorphic to $(\mathbb{R}, \mathcal{B})$ and to $([0,1], \mathcal{B}_{[0,1]})$.

This theorem is an easy consequence of the following result of KURATOWSKI [**?**], which is a sharpening of XIII.10.4.7.. The proof is rather involved, although fairly elementary, and we omit it; see, for example, [**?**, 2.4.1].

XVI.3.1.3. THEOREM. Let X be an uncountable Polish space which is dense in itself. Then there is an injective continuous function from $\mathcal{N}$ onto X such that the image of every open set is an F_σ.

Such a function is obviously a Borel isomorphism. To obtain XVI.3.1.2., note that by the Cantor-Bendixson Theorem (XIII.2.5.4.) every Polish space is a disjoint union of a countable open subset and a closed subset which is dense-in-itself.

XVI.3.1.4. So to finish the proof of XVI.3.1.2., note that if Y is $\mathcal{N}$ with a countable number of points removed, then Y is a G_δ in $\mathcal{N}$, hence a Polish space, and Y is dense in itself; so Y is Borel isomorphic to $\mathcal{N}$. Since every countable subset of a T_1-space is a Borel set, it follows that $(\mathcal{N}, \mathcal{B}_\mathcal{N})$ is isomorphic as a Borel space to $(\mathcal{N}, \mathcal{B}_\mathcal{N}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ (). Any uncountable Polish space X decomposes into a countable set and a Borel subset Z Borel isomorphic to $\mathcal{N}$ by XVI.3.1.3., so

$$(X, \mathcal{B}_X) \cong (Z, \mathcal{B}_Z) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N})) \cong (\mathcal{N}, \mathcal{B}_\mathcal{N}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N})) \cong (\mathcal{N}, \mathcal{B}_\mathcal{N}) \ .$$

See problems ()–() for an alternate proof of XVI.3.1.2..

In particular, we have

XVI.3.1.5. COROLLARY. Every uncountable Polish space has cardinality $2^{\aleph_0}$.

XVI.3.1.6. DEFINITION. A measurable space $(X, \mathcal{B})$ is a *standard Borel space* if it is isomorphic to $([0,1], \mathcal{B}_{[0,1]})$, hence to $(Y, \mathcal{B}_Y)$ for any uncountable Polish space Y.

Sometimes measurable spaces of the form $(X, \mathcal{P}(X))$ for a countable set X are also called standard Borel spaces. These are the measurable spaces of the form $(X, \mathcal{B}_X)$ for countable Polish spaces X.

Subspaces of Standard Borel Spaces

XVI.3.1.7. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space which is countably generated and separates points (i.e. if $x, y \in X$, there is an $A \in \mathcal{A}$ with $x \in A$ and $y \in A^c$). Then there is a subset Y of $\mathcal{C}$ such that $(X, \mathcal{A})$ is Borel isomorphic to $(Y, \mathcal{B}_Y)$.

PROOF: Let $\{A_n : n \in \mathbb{N}\}$ generate $\mathcal{A}$ as a σ-algebra. For $x \in X$, define $f(x) \in \mathcal{C}$ by $f(x)(n) = 1$ if $x \in A_n$, $f(x) = 0$ if $x \notin A_n$. It is easily checked that f is a Borel isomorphism onto $f(X)$. 👍

Such an $(X, \mathcal{A})$ is not necessarily a standard Borel space, even if uncountable (). But if the set Y is an uncountable G_δ in $\mathcal{C}$, then it follows from XIII.10.5.3. that $(X, \mathcal{A})$ is a standard Borel space; We will extend this to show that $(X, \mathcal{A})$ is a standard Borel space if Y is merely an uncountable Borel set in $\mathcal{C}$.

XVI.3.1.8. THEOREM. Let $(X, \mathcal{T})$ be a Polish space and A a Borel set in $(X, \mathcal{T})$. Then there is a Polish topology $\mathcal{S}$ on X which is stronger than $\mathcal{T}$, in which A is a clopen set, and such that every Borel set in $(X, \mathcal{S})$ is a Borel set in $(X, \mathcal{T})$.

PROOF: Let $\mathcal{A}$ be the set of all Borel subsets of X for which such an extension of $\mathcal{T}$ is possible. If F is closed in X, let $\mathcal{S} = \{U, U \cap F : U \in \mathcal{T}\}$. Then $\mathcal{S}$ is a topology on X stronger than $\mathcal{T}$, F is clopen in $\mathcal{S}$, and $\mathcal{S} \subseteq \mathcal{B}_\mathcal{T}$, so $\mathcal{B}_\mathcal{S} = \mathcal{B}_\mathcal{T}$. Also, $(X, \mathcal{S})$ is homeomorphic to $(F, \mathcal{T}_F) \sqcup (F^c, \mathcal{T}_{F^c})$, and $(F, \mathcal{T}_F)$ and $(F^c, \mathcal{T}_{F^c})$ are Polish by (), so $(X, \mathcal{S})$ is Polish. Thus all closed sets in X are in $\mathcal{A}$. $\mathcal{A}$ is clearly closed under complements.

To show that $\mathcal{A}$ is closed under countable unions, suppose (A_n) is a sequence in $\mathcal{A}$. For each n let $\mathcal{S}_n$ be a Polish topology on X stronger than $\mathcal{T}$, with the same Borel sets, in which A_n is clopen. Let $\mathcal{S}_\infty$ be the topology on X generated by $\cup_n \mathcal{S}_n$; then $\mathcal{S}_\infty \subseteq \mathcal{B}_\mathcal{T}$, so $\mathcal{B}_{\mathcal{S}_\infty} = \mathcal{B}_\mathcal{T}$. If σ_n is a $[0,1]$-valued complete metric giving $\mathcal{S}_n$, then $\sigma = \sum_n 2^{-n}\sigma_n$ is a complete metric giving $\mathcal{S}_\infty$. If D_n is a countable dense subset of $(X, \mathcal{S}_n)$, then $D = \cup_n D_n$ is a countable dense subset of $(X, \mathcal{S}_\infty)$. Thus $(X, \mathcal{S}_\infty)$ is Polish. We have that $A = \cup_n A_n$ is open in $(X, \mathcal{S}_\infty)$, so by the first part of the proof there is a Polish topology $\mathcal{S}$ stronger than $\mathcal{S}_\infty$, with the same Borel sets, such that A is clopen in $(X, \mathcal{S})$. Thus $A \in \mathcal{A}$ and $\mathcal{A}$ is a σ-algebra containing $\mathcal{T}$, hence containing $\mathcal{B}_\mathcal{T}$. 👍

XVI.3.1.9. COROLLARY. An uncountable measurable subspace of a standard Borel space is a standard Borel space.

An interesting consequence is the following, which "settles" the Continuum Hypothesis for Borel sets:

XVI.3.1.10. COROLLARY. Let X be a Polish space. Any uncountable Borel subset of X has cardinality $2^{\aleph_0}$.

XVI.3.2. Analytic Sets

XVI.3.2.1. DEFINITION. Let X be a Polish space. A subset Y of X is *analytic* if there is a continuous function from $\mathcal{N}$ onto Y.

XVI.3.2.2. The official definition of an analytic set in a general topological space X is a little different: it is a subset of X of the form $\pi_X(F)$, where F is a closed subset of $X \times \mathcal{N}$. If $Y \subset X$ and there is a continuous function f from $\mathcal{N}$ onto Y, then Y is analytic (let F be the graph of f). Conversely, if X is Polish and $Y = \pi_X(F)$ for a closed subset F of $X \times \mathcal{N}$, then F is Polish, so there is a continuous function g from $\mathcal{N}$ onto F by XIII.10.4.7., and $\pi_X \circ g$ is a continuous function from $\mathcal{N}$ onto Y. Thus the two definitions are equivalent for X Polish.

By XIII.10.4.7., if X is a Polish space, then X itself is analytic in X; and the property of being analytic is independent of the containing space.

It is obvious from the definition that continuous images of analytic sets are analytic:

XVI.3.2.3. Proposition. If X and Z are Polish spaces, $f : X \to Z$ is continuous, and Y is an analytic subset of X, then $f(Y)$ is an analytic subset of Z.

One of the most important results about analytic sets is:

XVI.3.2.4. Theorem. Let X be a Polish space. The collection of analytic subsets of X is closed under countable unions and countable intersections.

Proof: Closure under unions is fairly simple. Let E_n be an analytic set for each n, and $f_n : \mathcal{N} \to E_n$ a surjective continuous function. We can regard f_n as being defined on $\{n\} \times \mathcal{N} \subseteq \mathbb{N} \times \mathcal{N}$. Putting these together, we get a map f from $\mathbb{N} \times \mathcal{N} \cong \mathcal{N}$ onto $\cup E_n$.

Closure under intersections is trickier, and cannot be done simply by taking complements because it is not clear (and in fact turns out to be false) that the complement of an analytic set is analytic. So we proceed as follows. Let ψ be a homeomorphism from $\mathcal{N}$ onto $\mathcal{N}^{\mathbb{N}}$, with coordinate functions ψ_n. (This ψ is an analog of the "ultimate space-filling curve" of ().) Let E_n be an analytic subset of X for each n and f_n a continuous function from $\mathcal{N}$ onto E_n. Let

$$Y = \{s \in \mathcal{N} : f_n(\psi_n(s)) = f_m(\psi_m(s)) \text{ for all } n, m \in \mathbb{N}\} .$$

Y is a closed subset of $\mathcal{N}$. For $s \in Y$, set $f(s) = f_1(\psi_1(s)) = f_n(\psi_n(s))$ for all $n \in \mathbb{N}$. We have that $f(s) \in f_n(\mathcal{N}) = E_n$ for all n, hence the range of f is contained in $\cap E_n$. If $x \in \cap E_n$, then for each n there is an $s_n \in \mathcal{N}$ with $f_n(s_n) = x$. Then there is an $s \in \mathcal{N}$ with $s_n = \psi_n(s)$ for all n; then $s \in Y$ and $f(s) = x$. So the range of f is exactly $\cap E_n$. Since Y is closed in $\mathcal{N}$, it is analytic, and hence $\cap E_n$ is analytic by XVI.3.2.3.. 👍

XVI.3.2.5. Corollary. Let X be a Polish space. Then every Borel subset of X is analytic.

Proof: Every closed subset of X is a Polish space, hence analytic by XIII.10.4.7.. The result then follows from XVI.3.2.4. and (). 👍

This result also follows easily from XVI.3.1.8.: if B is a Borel set in X, there is a Polish topology $\mathcal{S}$ on X stronger than the given topology $\mathcal{T}$ for which B is clopen. B is thus Polish in the relative topology from $\mathcal{S}$; hence by XIII.10.4.7. there is an $\mathcal{S}$-continuous map from $\mathcal{N}$ onto B, which is also $\mathcal{T}$-continuous.

The converse of XVI.3.2.5. is not quite true, since we will see that the complement of an analytic set is not necessarily analytic. But the converse is not "too far" from being true. To see this, we first show that disjoint analytic sets can be separated by Borel sets:

XVI.3.2.6. THEOREM. [LUZIN SEPARATION THEOREM] Let X be a Polish space, and E and F disjoint analytic subsets of X. Then there is a Borel subset B of X such that $E \subseteq B$ and $F \subseteq B^c$.

PROOF: First note that if (Y_n) and (Z_m) are sequences of subsets of X which can be separated by Borel sets, i.e. for each n, m there is a Borel set $B_{n,m}$ with $Y_n \subseteq B_{n,m}$ and $Z_m \subseteq B^c_{n,m}$, then there is a Borel set B separating $\cup_n Y_n$ and $\cup_m Z_m$: $B = \cup_n \cap_m B_{n,m}$ works.

Write $E = \phi(\mathcal{N})$ and $F = \psi(\mathcal{N})$ for continuous functions ϕ, ψ. If $(k_1, \dots, k_n) \in \mathbb{N}^n$, set

$$\mathcal{N}_{k_1,\dots,k_n} = \{(s_1, s_2, \dots) \in \mathcal{N} : s_j = k_j \text{ for } 1 \leq j \leq n\} .$$

and put $E_{k_1,\dots,k_n} = \phi(\mathcal{N}_{k_1,\dots,k_n})$, $F_{k_1,\dots,k_n} = \psi(\mathcal{N}_{k_1,\dots,k_n})$.

Suppose E and F are not separated by a Borel set. Since $E = \cup_k E_k$ and $F = \cup_k F_k$, by the first part of the proof there must be s_1, t_1 such that E_{s_1} and F_{t_1} are not separated by a Borel set. Similarly, since $E_{s_1} = \cup_k E_{s_1,k}$ and $F_{t_1} = \cup_k F_{t_1,k}$, there must be s_2, t_2 such that E_{s_1,s_2} and F_{t_1,t_2} are not separated by Borel sets. Continuing inductively, sequences s and t in $\mathcal{N}$ are constructed such that $E_{s_1,\dots,s_n}$ and $F_{t_1,\dots,t_n}$ cannot be separated by a Borel set. Let $x = \phi(s)$ and $y = \phi(t)$. Then $x \in E$ and $y \in F$, so $x \neq y$. Let U and V be disjoint neighborhoods of x and y in X. By continuity, we must have $E_{s_1,\dots,s_n} = \phi(\mathcal{N}_{s_1,\dots,s_n}) \subseteq U$ and $F_{t_1,\dots,t_n} = \psi(\mathcal{N}_{t_1,\dots,t_n}) \subseteq V$, so U separates $E_{s_1,\dots,s_n}$ and $F_{t_1,\dots,t_n}$, a contradiction. 👍

Note that no form of Choice is needed in this proof: each s_j and t_j can just be chosen to be the smallest one for which the conditions hold (specifically, choose s_j to be the smallest possible for which there is a t_j, and then t_j the smallest possible for this s_j).

XVI.3.2.7. COROLLARY. Let E be an analytic subset of a Polish space X. Then E is a Borel set if and only if E^c is also analytic.

XVI.3.2.8. EXAMPLE. It is not so easy to construct an analytic set which is not a Borel set. Here is a construction of an analytic subset A of $\mathcal{N}$ whose complement in $\mathcal{N}$ is not analytic, hence A is not Borel. Since $\mathcal{N}$ is homeomorphic to the set of irrational numbers $\mathbb{J}$, A may be regarded as an analytic subset of $\mathbb{R}$ whose complement in $\mathbb{R}$ is not analytic (since if it were, $(\mathbb{R} \setminus A) \cap \mathbb{J}$ would be analytic by XVI.3.2.4.). The construction uses a version of Cantor's diagonalization argument.

We first construct a "universal" open set W in $\mathcal{N}^3$ which is a catalog for open sets in $\mathcal{N}^2$, i.e. for every open set V in $\mathcal{N}^2$ there is a $t \in \mathcal{N}$ such that $(r, s) \in V \iff (r, s, t) \in W$. The basic open sets in $\mathcal{N}^2$ (products of sets of the form $\mathcal{N}^{k_1,\dots,k_m}_{n_1,\dots,n_m}$ ()) can be arranged in a sequence (U_n) (this can be done explicitly without any form of AC; cf. Exercise **??**). For each $n \in \mathbb{N}$, define

$$W_n = \{(r, s, t) \in \mathcal{N}^3 | (r, s) \in U_{t_n}\}$$

where as usual we write $t = (t_1, t_2, \dots)$. Then W_n is open: if $(r, s, t) \in W_n$, then

$$\{(r', s', t') \in \mathcal{N}^3 | (r', s') \in U_{t_n}, t'_n = t_n\}$$

is a neighborhood of (r,s,t) contained in W_n. Set $W = \cup_n W_n$. Then W is open in $\mathcal{N}^3$. If V is an open set in $\mathcal{N}^2$, let $t = (t_1, t_2, \dots)$ be the set of n's (in increasing order) such that $U_n \subseteq V$. Then $V = \cup_n U_{t_n}$, so

$$(r,s) \in V \iff (r,s) \in U_{t_n} \text{ for some } n \iff (r,s,t) \in W_n \text{ for some } n \iff (r,s,t) \in W$$

so W is a catalog for open sets in $\mathcal{N}^2$.

Let $G = \mathcal{N}^3 \setminus W$. Then G is a catalog for closed sets if $\mathcal{N}^2$: if $F \subset \mathcal{N}^2$ is closed, let $V = \mathcal{N}^2 \setminus F$. Then there is a $t \in \mathcal{N}$ such that

$$(r,s) \in F \iff (r,s) \notin V \iff (r,s,t) \notin W \iff (r,s,t) \in G \ .$$

Now let $C = \pi_{13}(G) \subseteq \mathcal{N}^2$ be the projection of G onto the first and third coordinates. Then C is an analytic subset of $\mathcal{N}^2$ since G is closed in $\mathcal{N}^2 \times \mathcal{N}$. We claim that C is a catalog for analytic subsets of $\mathcal{N}$. If B is an analytic subset of $\mathcal{N}$, there is a closed set $F \subseteq \mathcal{N}^2$ with $B = \pi_1(F)$. There is then a $t \in \mathcal{N}$ such that $(r,s) \in F \iff (r,s,t) \in G$. We then have, for $r \in \mathcal{N}$,

$$r \in B \iff (r,s) \in F \text{ for some } s \iff (r,s,t) \in G \text{ for some } s \iff (r,t) \in C \ .$$

Finally, let $D = \{(t,t) | (t,t) \in C\} \subseteq \mathcal{N}^2$ and let $A = \pi_1(D)$. D is the intersection of C with the diagonal in $\mathcal{N}^2$, which is a closed, hence analytic, subset of $\mathcal{N}^2$; hence D is analytic in $\mathcal{N}^2$ by XVI.3.2.4., and so A is analytic in $\mathcal{N}$. But $A^c = \mathcal{N} \setminus A$ cannot be analytic, since if it were there would be a $t \in \mathcal{N}$ such that $r \in A^c \iff (r,t) \in C$. But then, for this t,

$$t \in A^c \iff (t,t) \in C \iff (t,t) \in D \iff t \in A$$

which is a contradiction.

XVI.3.3. Suslin Sets and the Suslin Operation

XVI.3.3.1. There is a curious operation on sets called the *Suslin operation*, which is reminiscent of the construction in the proof of XIII.10.4.7.. For each finite sequence $(s_1, \dots, s_n)$ of natural numbers, let $A_{s_1,\dots,s_n}$ be a subset of a set X. The collection of finite sequences in $\mathbb{N}$ is countable, so there are countably many such sets in all. Define

$$S(\{A_{s_1,\dots,s_n}\}) = \bigcup_{s \in \mathcal{N}} \bigcap_{n=1}^{\infty} A_{s_1,\dots,s_n} \ .$$

Although the operation begins with a countable collection of sets, the union is over all sequences (uncountably many) in $\mathcal{N}$.

XVI.3.3.2. Definition. Let X be a Polish space. A *Suslin set* is a subset of X obtained by applying the Suslin operation to a collection of closed subsets of X.

XVI.3.3.3. THEOREM. A subset of a Polish space is a Suslin set if and only if it is analytic.

PROOF: Suppose A is a Suslin set in a Polish space X. Let $A = \cup_{s\in\mathcal{N}} \cap_n F_{s|n}$, where the $F_{s|n}$ are closed sets. Set

$$F = \{(x, s) \in X \times \mathcal{N} : x \in F_{s|n} \text{ for all } n\} \subseteq X \times \mathcal{N} .$$

Then F is a closed subset of $X \times \mathcal{N}$, and $A = \pi_X(F)$, so A is analytic.

Conversely, suppose A is analytic. Fix a complete metric ρ on X. Let $\phi : \mathcal{N} \to A$ be continuous. For each $(s_1, \ldots, s_n)$ set $F_{s_1,\ldots,s_n} = \overline{\phi(\mathcal{N}_{s_1,\ldots,s_n})}$. Then $F_{s_1,\ldots,s_n}$ is closed and for each $s \in \mathcal{N}$, $F_{s|(n+1)} \subseteq F_{s|n}$ for all n, and $diam(F_{s|n}) \to 0$ as $n \to \infty$ since ϕ is continuous. Thus for every $s \in \mathcal{N}$ there is a unique $x_s \in \cap_n F_{s|n}$ since X is complete. Fix s and set $x = x_s$. Then for each n there is an $x_n \in \phi(\mathcal{N}_{s_1,\ldots,s_n})$ with $\rho(x_n, x) < 2^{-n}$, and there is an $s^{(n)} \in \mathcal{N}_{s_1,\ldots,s_n}$ with $\phi(s^{(n)}) = x_n$. We have $s^{(n)} \to s$ since $s^{(n)}|n = s|n$ for all n; thus $\phi(s^{(n)}) \to \phi(s)$. But $\phi(s^{(n)}) = x_n \to x$, so $x = \phi(s)$, i.e. $x \in A$. So $S(\{F_{s_1,\ldots,s_n}\}) \subseteq A$. The reverse containment is obvious, so $A = S(\{F_{s_1,\ldots,s_n}\}) \in \mathcal{S}(\mathcal{F})$. 👍

XVI.3.3.4. THEOREM. Let $(X, \mathcal{A}, \mu)$ be a complete measure space. Then $\mathcal{A}$ is closed under the Suslin operation.

XVI.3.3.5. COROLLARY. Let X be a Polish space, and μ a complete measure defined on a σ-algebra $\mathcal{A}$ of subsets of X containing the Borel sets. Then $\mathcal{A}$ contains all analytic subsets of X.

XVI.3.3.6. The history of Suslin sets and the Suslin operation is unfortunate and somewhat murky, and is an important facet of the story of analysis in the Soviet Union. SUSLIN died of typhus in 1919 at age 24, having published only one paper during his lifetime. LUZIN, his mentor, then refined, expanded on, and published some of SUSLIN'S work (with credit to SUSLIN). Later, ALEKSANDROV claimed that he was responsible for the Suslin operation and the notion of analytic set, and that in the original names "A-operation" and "A-set", used by both SUSLIN and LUZIN, the "A" stood for "Aleksandrov", not "analytic", the term eventually used by LUZIN (LUZIN said this name had been proposed by LEBESGUE). There is some (inconclusive) evidence to support ALEKSANDROV'S claim, but also persuasive evidence to the contrary, and most contemporary scholarship attributes the concept of analytic set to SUSLIN and LUZIN (HAUSDORFF also played a significant and rather independent role); cf. [Lor01]. The relations between ALEKSANDROV and LUZIN began well but soured early, and contained a lot of personal friction and political intrigue during the Stalin era. ALEKSANDROV was an active participant in the attempted political persecution of LUZIN known as the "Luzin affair." Accounts of these and related events vary widely in point of view, and it is quite difficult to get an unbiased picture of what really happened. The story is certainly a complicated one, involving (according to references) politics, mysticism, and homosexuality. See [GK09], [ZD07], and [Sin03] for pieces of the story.

There are many fascinating anecdotes about these personalities in these references (many of which involve their fanatical and occasionally tragic swimming adventures). For example, the usually reserved KOLMOGOROV, a close friend of ALEKSANDROV, once slapped LUZIN in the face during a meeting of the Academy of Sciences (according to [GK09], in response to an insult by LUZIN); the incident came to the personal attention of STALIN, who shrugged it off, saying, "Well, such things occur with us as well." (Fist-fights in the Politburo?) It should be noted that both before and after this incident, KOLMOGOROV had kind things to say about LUZIN, as did ALEKSANDROV earlier in his career.

XVI.3.3.7. Universally measurable sets

XVI.3.4. Addison's Notation and the Projective Hierarchy

In 1958 J. Addison [Add59] introduced coherent new notation not only for the sets in the hierarchy of the construction of the Borel sets, but also for analytic sets and a continuing hierarchy of subsets of $\mathbb{R}$ (or a general topological space). Indexing by ordinals had been part of the subject since the work of Baire, Lebesgue, Hausdorff, and the Russian school.

XVI.3.4.1. Fix a topological space X, usually $\mathbb{R}$ or $\mathcal{N}$. Define the classes $\mathbf{\Sigma}^0_\alpha$ and $\mathbf{\Pi}^0_\alpha$ of subsets of X inductively for $0 < \alpha < \omega_1$ as follows:

$\mathbf{\Sigma}^0_1$ is the class of open sets.

$\mathbf{\Pi}^0_1$ is the class of closed sets.

$\mathbf{\Sigma}^0_\alpha$ is the class of countable unions of sets in $\cup_{\beta<\alpha}\mathbf{\Pi}^0_\beta$.

$\mathbf{\Pi}^0_\alpha$ is the class of countable intersections of sets in $\cup_{\beta<\alpha}\mathbf{\Sigma}^0_\beta$.

Thus $\mathbf{\Sigma}^0_2$ is the class of F_σ's, $\mathbf{\Pi}^0_2$ the class of G_δ's, $\mathbf{\Sigma}^0_3$ the class of $G_{\delta\sigma}$'s, $\mathbf{\Pi}^0_3$ the class of $F_{\sigma\delta}$'s, etc.

A set A is called a $\mathbf{\Sigma}^0_\alpha$-set ($\mathbf{\Pi}^0_\alpha$-set) if $A \in \mathbf{\Sigma}^0_\alpha$ ($A \in \mathbf{\Pi}^0_\alpha$).

XVI.3.4.2. If X is Polish, we have that $\mathbf{\Sigma}^0_\beta \subseteq \mathbf{\Pi}^0_\alpha$ and $\mathbf{\Pi}^0_\beta \subseteq \mathbf{\Sigma}^0_\alpha$ if $\beta < \alpha$. The complement of a $\mathbf{\Sigma}^0_\alpha$-set is a $\mathbf{\Pi}^0_\alpha$-set, and vice versa. Since $\mathbf{\Sigma}^0_1 \subseteq \mathbf{\Sigma}^0_2$ (), by induction we have $\mathbf{\Sigma}^0_\beta \subseteq \mathbf{\Sigma}^0_\alpha$ and $\mathbf{\Pi}^0_\beta \subseteq \mathbf{\Pi}^0_\alpha$ for $\beta < \alpha$. The unions $\cup_{\alpha<\omega_1}\mathbf{\Sigma}^0_\alpha$ and $\cup_{\alpha<\omega_1}\mathbf{\Pi}^0_\alpha$ are both equal to the class of Borel sets (the Countable AC is needed to show that these unions form a σ-algebra and hence contain all Borel sets (XVI.3.5.4.)).

XVI.3.4.3. It follows easily from the results of () that for $X = \mathbb{R}$, these classes strictly increase in α. This was first shown by Lebesgue [**?**], who proved that $\mathbf{\Sigma}^0_\alpha \not\subseteq \mathbf{\Pi}^0_\alpha$ for any $\alpha < \omega_1$ (see ()).

We also define $\mathbf{\Delta}^0_\alpha = \mathbf{\Sigma}^0_\alpha \cap \mathbf{\Pi}^0_\alpha$.

XVI.3.4.4. Now for $n \in \mathbb{N} \cup \{0\}$ define $\mathbf{\Sigma}^1_n$ and $\mathbf{\Pi}^1_n$ inductively as follows.

$\mathbf{\Sigma}^1_0 = \mathbf{\Sigma}^0_1$ is the class of open sets.

$\mathbf{\Pi}^1_n$ is the class of complements of $\mathbf{\Sigma}^1_n$-sets.

$\mathbf{\Sigma}^1_{n+1}$ is the class of π_X images of $\mathbf{\Pi}^1_n$-sets in $X \times \mathcal{N}$.

$\mathbf{\Delta}^1_n = \mathbf{\Sigma}^1_n \cap \mathbf{\Pi}^1_n$.

Thus $\mathbf{\Pi}^1_0$ is the class of closed sets, and $\mathbf{\Sigma}^1_1$ is the class of analytic sets (). $\mathbf{\Delta}^1_1$ is the class of Borel sets (). It follows that $\mathbf{\Sigma}^1_0 \cup \mathbf{\Pi}^1_0 \subseteq \mathbf{\Delta}^1_1$ (there is also a simple direct proof of this) and hence by induction $\mathbf{\Sigma}^1_n \cup \mathbf{\Pi}^1_n \subseteq \mathbf{\Delta}^1_{n+1}$ for all n.

A set in $\cup_n\mathbf{\Sigma}^1_n = \cup_n\mathbf{\Pi}^1_n = \cup_n\mathbf{\Delta}^1_n$ is called a *projective set*.

XVI.3.4.5. To see why the projection operation is natural, and that this hierarchy is analogous to the Borel sets, note that an F_σ ($\mathbf{\Sigma}^0_2$-set) in X is a π_X projection of a closed set ($\mathbf{\Pi}^0_1$-set) in $X \times \mathbb{N}$, and more generally a $\mathbf{\Sigma}^0_{\alpha+1}$-set in X is the π_X projection of a $\mathbf{\Pi}^0_\alpha$-set in $X \times \mathbb{N}$.

XVI.3.4.6. This hierarchy can be extended transfinitely to define $\mathbf{\Sigma}^1_\alpha$ and $\mathbf{\Pi}^1_\alpha$ for $\alpha < \omega_1$. What about $\mathbf{\Sigma}^n_\alpha$, $\mathbf{\Sigma}^\beta_\alpha$, etc.?

XVI.3.4.7. Addison's notation is strongly reminiscent of similar notation used in mathematical logic and model theory, and was in fact inspired by analogies and connections with the logic notation.

XVI.3.5. Exercises

XVI.3.5.1. Let $\mathcal{C} = \{0,1\}^{\mathbb{N}}$ ($\mathcal{C}$ is homeomorphic to the Cantor set ()). Define a map $f : \mathcal{N} \to \mathcal{C}$ by

$$f(n_1, n_2, \dots) = (\underbrace{0, \dots, 0}_{n_1 - 1}, 1, \underbrace{0, \dots, 0}_{n_2 - 1}, 1, \dots) \ .$$

(a) Show that f is a homeomorphism of $\mathcal{N}$ with a subset of $\mathcal{C}$ with countable complement.

(b) Show that $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is Borel isomorphic to $(\mathcal{N}, \mathcal{B}_\mathcal{N}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$.

(c) Since $\mathcal{N}$ is also homeomorphic to the irrational numbers in $[0,1]$, $([0,1], \mathcal{B}_{[0,1]})$ is Borel isomorphic to $(\mathcal{N}, \mathcal{B}_\mathcal{N}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Thus $([0,1], \mathcal{B}_{[0,1]})$ is Borel isomorphic to $(\mathcal{C}, \mathcal{B}_\mathcal{C})$.

(d) Show that $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is also Borel isomorphic to $([0,1]^{\mathbb{N}}, \mathcal{B}_{[0,1]^{\mathbb{N}}})$. [$\mathcal{C}$ is homeomorphic to $\mathcal{C}^{\mathbb{N}}$.]

(e) Since $(\mathbb{N}, \mathcal{P}(\mathbb{N})) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ is Borel isomorphic to $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, $(\mathcal{C}, \mathcal{B}_\mathcal{C}) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ is Borel isomorphic to $(\mathcal{C}, \mathcal{B}_\mathcal{C})$. [Use (b).]

XVI.3.5.2. If (X, ρ) is a perfect complete metric space, show that there is a homeomorphism from $\mathcal{C}$ onto a (compact) subspace of X. Thus $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is Borel isomorphic to a measurable subspace of $(X, \mathcal{B}_X)$.

XVI.3.5.3. (a) Let X be a Polish space. Then X is homeomorphic to a G_δ in the Hilbert cube (() and ().) Thus $(X, \mathcal{B}_X)$ is Borel isomorphic to a measurable subspace of $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ by ()(d).

(b) If X is an uncountable Polish space, then by () and () $(X, \mathcal{B}_X)$ is Borel isomorphic to $(Y, \mathcal{B}_Y) \oplus (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ for some perfect (uncountable) Polish space Y. Thus $(X, \mathcal{B}_X)$ is Borel isomorphic to a measurable subspace of $(\mathcal{C}, \mathcal{B}_\mathcal{C})$.

(c) If X is an uncountable Polish space, then X has a (compact) subspace homeomorphic to $\mathcal{C}$ (() and ()), and thus $(\mathcal{C}, \mathcal{B}_\mathcal{C})$ is Borel isomorphic to a measurable subspace of $(X, \mathcal{B}_X)$.

(d) Apply the Schröder-Bernstein Theorem for measurable spaces (XVI.1.7.3.) to conclude that if X is any uncountable Polish space, then $(X, \mathcal{B}_X)$ is Borel isomorphic to $(\mathcal{C}, \mathcal{B}_\mathcal{C})$.

XVI.3.5.4. Explain why the Countable AC is needed to show that the unions $\mathcal{A} = \cup_{\alpha<\omega_1} \mathbf{\Sigma}^0_\alpha = \cup_{\alpha<\omega_1} \mathbf{\Pi}^0_\alpha$ (XVI.3.4.2.) form a σ-algebra. (See II.8.6.10..)

Chapter XVII

Measure Theory

XVII.1. Measures

It should come as no surprise that the fundamental concept of measure theory is the notion of measure. Measures live on measurable spaces (), and assign to each measurable set a nonnegative extended real-number "size" for the set. The crucial property a measure must have is countable additivity.

XVII.1.1. Definitions

XVII.1.1.1. DEFINITION. Let $(X, \mathcal{A})$ be a measurable space. A *measure* on $(X, \mathcal{A})$ is a function $\mu : \mathcal{A} \to [0, \infty]$ satisfying

(i) $\mu(\emptyset) = 0$.

(ii) If $A_1, A_2, \ldots$ is a sequence of pairwise disjoint sets in $\mathcal{A}$, then

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k) \left[= \sup_n \sum_{k=1}^{n} \mu(A_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(A_k)\right]$$

(μ is *countably additive*).

A *measure space* is a triple $(X, \mathcal{A}, \mu)$, where $(X, \mathcal{A})$ is a measurable space and μ is a measure on $(X, \mathcal{A})$.

XVII.1.1.2. Note that specification of the σ-algebra is part of the definition of a measure and of a measure space. While we frequently downplay this specification, especially in contexts where there is some natural σ-algebra of measurable sets, it cannot be ignored. Thus, for example, if $(X, \mathcal{A}, \mu)$ is a measure space and $\mathcal{B}$ is a proper sub-σ-algebra of $\mathcal{A}$, and $\nu = \mu|_{\mathcal{B}}$, then ν is technically a different measure than μ, and the measure spaces $(X, \mathcal{A}, \mu)$ and $(X, \mathcal{B}, \nu)$ are different measure spaces. See XVII.1.6.12..

XVII.1.1.3. Condition (i) in the definition is almost redundant: the only function satisfying (ii) but not (i) is the function μ with $\mu(A) = \infty$ for all $A \in \mathcal{A}$. (To see this, let $A \in \mathcal{A}$, set $A_1 = A$ and $A_n = \emptyset$ for $n > 1$, and apply (ii).)

By a similar argument, we can see:

XVII.1.1.4. PROPOSITION. Let μ be a measure on $(X, \mathcal{A})$. Then μ is *finitely additive*: if $A_1, \ldots, A_n$ are pairwise disjoint sets in $\mathcal{A}$, then

$$\mu\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} \mu(A_k) \ .$$

In particular, if $A, B \in \mathcal{A}$ are disjoint, then $\mu(A \cup B) = \mu(A) + \mu(B)$.

PROOF: Set $A_k = \emptyset$ for $k > n$, and apply countable additivity. 👍

XVII.1.1.5. We can define a *finitely additive measure* on a measurable space (X, μ) to be a finitely additive function $\mu : \mathcal{A} \to [0, \infty]$ with $\mu(\emptyset) = 0$. (Actually we only need for $\mathcal{A}$ to be an algebra of subsets of X for this definition to make sense.) XVII.1.1.4. says that every measure is a finitely additive measure; but there are many finitely additive measures which are not measures (e.g. XVII.1.1.6.(vii)–(viii)). A few authors call a finitely additive measure a "measure" (measures in our sense are then called "countably additive measures"), but this terminology is nonstandard. ✋

The theory of finitely additive measures is important in some contexts, and the theory of measure and integration can be developed surprisingly far in this generality (cf. [DS88, Chapter 3]). However, countable additivity is really what gives the theory of measure and integration its power. Countable additivity should be regarded as a sort of continuity condition (cf. XVII.1.2.7.).

Although we cannot yet give any really interesting examples of measures (the interesting ones take considerable effort to construct), here are some simple examples which range from uninteresting to moderately interesting:

XVII.1.1.6. EXAMPLES. (i) Let $(X, \mathcal{A})$ be any measurable space. Set $\mu(A) = 0$ for all $A \in \mathcal{A}$. Then μ is a measure on $(X, \mathcal{A})$, called the *zero measure*.

(ii) Let X be any set. For $A \subseteq X$, set $\mu(A) = 0$ if $A = \emptyset$ and $\mu(A) = \infty$ if $A \neq \emptyset$. Then μ is a measure on $(X, \mathcal{P}(X))$. (This measure can be restricted to any σ-algebra on X.)

(iii) Let X be an uncountable set. If $A \subseteq X$, set $\mu(A) = 0$ if A is countable and $\mu(A) = \infty$ if A is uncountable. Then μ is a measure on $(X, \mathcal{P}(X))$.

(iv) Let X be an uncountable set, $\mathcal{A}$ the σ-algebra of countable/cocountable subsets of X (). If $A \in \mathcal{A}$, set $\mu(A) = 0$ if A is countable and $\mu(A) = 1$ if A is cocountable. Then μ is a measure on $(X, \mathcal{A})$.

(v) Let X be any set, and p a fixed element of X. For $A \subseteq X$, set $\delta_p(A) = 1$ if $p \in A$ and $\delta_p(A) = 0$ if $p \notin A$. Then δ_p is a measure on $(X, \mathcal{P}(X))$, called the *(unit) point mass* (or *Dirac measure*) at p.

(vi) Let X be any set. If $A \subseteq X$, set $\mu(A) = n$ if A is a finite set with n elements, and $\mu(A) = \infty$ if A is infinite. Then μ is a measure on $(X, \mathcal{P}(X))$, called *counting measure* on X. This measure, while very simple, is important, especially in the case $X = \mathbb{N}$; integration theory for this measure is really exactly the theory of infinite series.

(vii) Let X be an infinite set. If $A \subseteq X$, set $\mu(A) = 0$ if A is finite and $\mu(A) = \infty$ if A is infinite. Then μ is a finitely additive measure on $(X, \mathcal{P}(X))$ which is not a measure.

(viii) Let $\mathcal{F}$ be a free ultrafilter () on $\mathbb{N}$. For $A \subseteq \mathbb{N}$, set $\mu(A) = 1$ if $A \in \mathcal{F}$ and $\mu(A) = 0$ if $A \notin \mathcal{F}$. Then μ is a finitely additive measure on $(X, \mathcal{P}(X))$ which is not a measure. The same construction works if $\mathbb{N}$ is replaced by any infinite set whose cardinality is less than the first measurable cardinal.

XVII.1.2. Elementary Properties

Here we develop some of the most basic properties of measures.

XVII.1.2.1. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space. Then

(i) μ is *monotone*: if $A, B \in \mathcal{A}$, $B \subseteq A$, then $\mu(B) \leq \mu(A)$.

(ii) μ is *countably subadditive*: if (A_k) is a sequence in $\mathcal{A}$, then

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mu(A_k) \ .$$

PROOF: (i): We have $A = B \cup (A \setminus B)$, a disjoint union, and $(A \setminus B) \in \mathcal{A}$, so

$$\mu(A) = \mu(B) + \mu(A \setminus B)$$

by finite additivity, and $\mu(A \setminus B) \geq 0$.
(ii): Disjointize the sequence (A_n) as in (), i.e. set $B_1 = A_1$ and $B_n = A_n \setminus (\cup_{k=1}^{n-1} A_k)$ for $n > 1$. Then the B_n are in $\mathcal{A}$ and are pairwise disjoint, $B_n \subseteq A_n$, and $\cup_{n=1}^{\infty} B_n = \cup_{n=1}^{\infty} A_n$. So, by countable additivity and (i),

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) \leq \sum_{n=1}^{\infty} \mu(A_n) \ .$$



As with XVII.1.1.4., we also have

XVII.1.2.2. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a measure space. Then μ is *finitely subadditive*: if $A_1, \ldots, A_n \in \mathcal{A}$, then

$$\mu\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} \mu(A_k) \ .$$

In particular, if $A, B \in \mathcal{A}$, then $\mu(A \cup B) \leq \mu(A) + \mu(B)$.

XVII.1.2.3. Note that by a nearly identical proof, a finitely additive measure is also monotone and finitely subadditive. A finitely additive measure need not be countably subadditive: the measure in XVII.1.1.6.(vii) is a counterexample. In fact, a finitely additive measure which is countably subadditive is countably additive, i.e. a measure (Exercise XVII.1.10.1.).

Decompositions of Measure Spaces

The next observation is an immediate consequence of countable additivity. We will often use this result to break up a measure space into a finite or countable number of pieces.

XVII.1.2.4. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, I a finite or countable index set, and $\{X_i : i \in I\}$ a collection of pairwise disjoint measurable subsets of X with $X = \cup_{i\in I} X_i$. Then for any $A \in \mathcal{A}$ we have

$$\mu(A) = \sum_{i\in I} \mu(A \cap X_i) \ .$$

XVII.1.2.5. Note that countability of the index set I is essential. In many examples such as in XVII.1.1.6.(iii), or $\mathbb{R}$ with Lebesgue measure, the entire space X can be decomposed as a disjoint union of singleton subsets, each of which is measurable and has measure zero.

Measures of Differences

XVII.1.2.6. If $(X, \mathcal{A}, \mu)$ is a measure space, and $A, B \in \mathcal{A}$ with $B \subseteq A$, and $\mu(B) < \infty$ (e.g. if $\mu(A) < \infty$), then we have that

$$\mu(A \setminus B) = \mu(A) - \mu(B)$$

where we interpret $\infty - b = \infty$ for $b \in \mathbb{R}$. However, if $\mu(B) = \infty$ (and hence $\mu(A) = \infty$ also by monotonicity), we cannot draw any conclusion about $\mu(A \setminus B)$ (cf. Exercise XVII.1.10.2.). Thus we must be careful in subtracting measures of sets when the measures are infinite.

Upward Continuity

The next property is the first instance of how countable additivity can be regarded as a continuity property. It is called the Baby Monotone Convergence Theorem because it is actually the special case of the Monotone Convergence Theorem for indicator functions.

XVII.1.2.7. PROPOSITION. [BABY MONOTONE CONVERGENCE THEOREM] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (A_n) an increasing sequence in $\mathcal{A}$, i.e. $A_n \subseteq A_{n+1}$ for all n. Then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sup_n \mu(A_n) = \lim_{n\to\infty} \mu(A_n)$$

(μ is *upward continuous*).

PROOF: Disjointize the sequence (A_n) by setting $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n > 1$. Then the B_n are in $\mathcal{A}$ and are pairwise disjoint, $A_n = \cup_{k=1}^{n} B_k$ for all n, and $\cup_{n=1}^{\infty} A_n = \cup_{k=1}^{\infty} B_k$. Then

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \mu\left(\bigcup_{k=1}^{\infty} B_k\right) = \sum_{k=1}^{\infty} \mu(B_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(B_k) = \lim_{n\to\infty} \mu(A_n) \ .$$

XVII.1.2.8. A finitely additive measure μ is not upward continuous in general, although if (A_n) is an increasing sequence of measurable sets we always have

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) \geq \sup_n \mu(A_n) = \lim_{n\to\infty} \mu(A_n)$$

by monotonicity. Example XVII.1.1.6.(vii) is a counterexample to equality. In fact, it is easily seen that upward continuity is equivalent to countable additivity for a finitely additive measure (Exercise XVII.1.10.1.).

Downward Continuity

What about downward continuity? The following example shows that it does not hold in general.

XVII.1.2.9. Example. Let μ be counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Let $A_n = \{k \in \mathbb{N} : k \geq n\}$. Then $A_n \supseteq A_{n+1}$ for all n. We have $\mu(A_n) = \infty$ for all n. But $\cap_{n=1}^{\infty} A_n = \emptyset$, so

$$0 = \mu\left(\bigcap_{n=1}^{\infty} A_n\right) < \inf_n \mu(A_n) = \lim_{n\to\infty} \mu(A_n) = \infty \ .$$

The difficulty stems from the failure of subtraction for measures of sets (cf. XVII.1.2.6.). In fact, we have:

XVII.1.2.10. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and (A_n) a decreasing sequence in $\mathcal{A}$, i.e. $A_n \supseteq A_{n+1}$ for all n. If $\mu(A_1) < \infty$, then

$$\mu\left(\bigcap_{n=1}^{\infty} A_n\right) = \inf_n \mu(A_n) = \lim_{n\to\infty} \mu(A_n)$$

(μ is *downward continuous* on sets of finite measure).

Proof: Set $B_n = A_1 \setminus A_n$ for each n. Then the B_n are in $\mathcal{A}$ and are increasing. If $A = \cap_{n=1}^{\infty} A_n$ and $B = \cup_{n=1}^{\infty} B_n$, we have $B = A_1 \setminus A$. We thus have, using XVII.1.2.6. and XVII.1.2.7.,

$$\mu(A_1) - \mu(A) = \mu(B) = \lim_{n\to\infty} \mu(B_n) = \lim_{n\to\infty} [\mu(A_1) - \mu(A_n)] = \mu(A_1) - \lim_{n\to\infty} \mu(A_n)$$

and the result follows. 👍

Note that, by passing to a tail of the sequence, it suffices that $\mu(A_n) < \infty$ for some n.

The First Borel-Cantelli Lemma

We can obtain various more sophisticated results about measures. For example, the next result is particularly important in probability, where it is called the *first Borel-Cantelli Lemma* (). Recall () that if (A_n) is a sequence of subsets of a set X, then

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \left(\bigcup_{k=n}^{\infty} A_k\right)$$

is the set of $x \in X$ which are in infinitely many of the A_n.

XVII.1.2.11. PROPOSITION. [FIRST BOREL-CANTELLI LEMMA] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (A_n) a sequence in $\mathcal{A}$. If

$$\sum_{k=1}^{\infty} \mu(A_k) < \infty \ ,$$

then

$$\mu(\limsup_n A_n) = 0 \ .$$

PROOF: Let $\epsilon > 0$. Then there is an m such that

$$\sum_{k=1}^{\infty} \mu(A_k) - \sum_{k=1}^{m} \mu(A_k) = \sum_{k=m+1}^{\infty} \mu(A_k) < \epsilon$$

(in fact this holds for all sufficiently large m). For such an m, we have

$$\limsup_n A_n \subseteq \bigcup_{k=m+1}^{\infty} A_k$$

so by monotonicity and countable subadditivity we have

$$\mu(\limsup_n A_n) \leq \mu\left(\bigcup_{k=m+1}^{\infty} A_k\right) \leq \sum_{k=m+1}^{\infty} \mu(A_k) < \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $\mu(\limsup_n A_n) = 0$. 👍

XVII.1.3. Null Sets and Completeness

XVII.1.3.1. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space. A set $A \in \mathcal{A}$ is a *null set* (or *μ-null set*) if $\mu(A) = 0$.

XVII.1.3.2. In general mathematics, the term "null set" is often used as a synonym for "empty set", but in analysis it is more useful to use the definition we have given. Note that the empty set is always a null set by the analysis definition, but many measure spaces also have other null sets which are nonempty. Of course, whether a set is a null set depends on the measure: if μ and ν are measures on a measurable space $(X, \mathcal{A})$, then a μ-null set is not necessarily a ν-null set.

The idea behind the terminology is that null sets are "negligibly small" for measure-theoretic purposes. A general philosophy in measure theory is that sets or functions can be modified on a null set with no effect. In fact, we will eventually feel quite free to ignore what happens on null sets, although at least at first one should be careful about this.

Unions of Null Sets

XVII.1.3.3. PROPOSITION. A countable union of null sets is a null set. (In particular, a finite union of null sets is a null set.)

This is an immediate consequence of countable subadditivity.

Subsets of Null Sets and Completeness

XVII.1.3.4. It may be surprising that a subset of a null set is not necessarily a null set. It is a trivial consequence of monotonicity that any *measurable* subset of a null set is a null set. However, there are many examples of measure spaces containing null sets with nonmeasurable subsets. Here is perhaps the simplest example:

XVII.1.3.5. Example. Let $\mathcal{A}$ be the σ-algebra of countable/cocountable subsets of $\mathbb{R}$. If $A \in \mathcal{A}$, let $\mu(A) = 1$ if $1 \in A$ and $\mu(A) = 0$ if $1 \notin A$. It is easily checked that μ is a measure on $(\mathbb{R}, \mathcal{A})$ (it is the restriction to $\mathcal{A}$ of the Dirac measure δ_1 of XVII.1.1.6.(v)). We have that $\mathbb{R} \setminus \{1\}$ is a null set, but $(1, +\infty)$ and $(-\infty, 1)$ are not null sets since they are not measurable, i.e. not in $\mathcal{A}$ (each is an uncountable set with uncountable complement).

This example may seem artificial since we arbitrarily chose a smaller σ-algebra than the natural one on which μ is defined (namely $\mathcal{P}(\mathbb{R})$). There are more natural examples such as Lebesgue measure on the Borel sets (), although in a certain sense every example is unnatural (XVII.1.3.7.).

XVII.1.3.6. Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space. Then $(X, \mathcal{A}, \mu)$ is a *complete measure space* if every subset of a null set is a null set (i.e. measurable).

XVII.1.3.7. Theorem. Let $(X, \mathcal{A}, \mu)$ be a measure space, and let $\bar{\mathcal{A}}$ be the σ-algebra of subsets of X generated by $\mathcal{A}$ and all subsets of μ-null sets. Then there is a unique extension $\bar{\mu}$ of μ to $(X, \bar{\mathcal{A}})$, and $(X, \bar{\mathcal{A}}, \bar{\mu})$ is complete. In fact, $\bar{\mathcal{A}}$ is the collection of all subsets of X of the form $A \cup D$, where $A \in \mathcal{A}$ and D is a subset of a μ-null set.

Proof: Let $\mathcal{B}$ be the collection of all subsets of X of the form $A \cup D$, where $A \in \mathcal{A}$ and D is a subset of a μ-null set. Note that $\emptyset$ and X are in $\mathcal{B}$, and it is easy (using XVII.1.3.3.) to check that $\mathcal{B}$ is closed under countable unions. Let $B = A \cup D \in \mathcal{B}$, where $A \in \mathcal{A}$ and $D \subseteq C \in \mathcal{A}$ with $\mu(C) = 0$. Then

$$X \setminus B = (X \setminus [A \cup C]) \cup (C \setminus B)$$

and $X \setminus [A \cup C] \in \mathcal{A}$, $C \setminus B \subseteq C$. Thus $\mathcal{B}$ is a σ-algebra and $\mathcal{A} \subseteq \mathcal{B} \subseteq \bar{\mathcal{A}}$; $\mathcal{B}$ contains all subsets of μ-null sets, so $\mathcal{B} = \bar{\mathcal{A}}$.
For $B = A \cup D$ as above, set $\bar{\mu}(B) = \mu(A)$. This is well defined: if $B = A_1 \cup D_1 = A_2 \cup D_2$ with $D_j \subseteq C_j$, $\mu(C_j) = 0$, then

$$A_1 \setminus C_2 \subseteq A_2 \subseteq A_1 \cup C_1$$

so $\mu(A_1) = \mu(A_2)$. It is clear that $\bar{\mu}$ extends μ and easy to check that $\bar{\mu}$ is countably additive. For uniqueness, if ν extends μ and $B = A \cup D$ with $D \subseteq C$, $\mu(C) = 0$, we must have

$$\mu(A) = \nu(A) \leq \nu(B) \leq \nu(A \cup C) = \mu(A \cup C) \leq \mu(A) + \mu(C) = \mu(A)$$

by monotonicity and subadditivity. Since every subset of a $\bar{\mu}$-null set is a subset of a μ-null set and hence in $\bar{\mathcal{A}}$, $(X, \bar{\mathcal{A}}, \bar{\mu})$ is complete. 👍

XVII.1.3.8. DEFINITION. The measure space $(X, \bar{\mathcal{A}}, \bar{\mu})$ is called the *completion* of $(X, \mathcal{A}, \mu)$ and, by slight abuse of terminology, $\bar{\mu}$ is called the *completion* of μ.

Almost Everywhere

XVII.1.3.9. It is convenient to use the phrase *almost everywhere*, abbreviated *a.e.*, to mean "except on a null set." Thus, for example, if we say "the functions f and g agree almost everywhere," we mean "there is a null set A such that $f(x) = g(x)$ for all $x \notin A$." When the measure μ needs to be specified, we say "μ-almost everywhere," written μ-a.e.

However, if we are dealing with an incomplete measure space, there is a subtlety. Saying "the functions f and g agree almost everywhere" does not necessarily mean "the set of all x for which $f(x) = g(x)$ does not hold (i.e. one or both of $f(x)$ and $g(x)$ is undefined, or both are defined but unequal) has measure zero," because this set could be nonmeasurable (it is only a subset of a set of measure 0). In fact, if f and g agree almost everywhere, and f is measurable, then g is not necessarily even measurable (consider the case where f is identically zero and g is the indicator function of a nonmeasurable subset of a null set). This difficulty disappears when working with complete measure spaces, and is one reason for completing measure spaces (although taking completions also introduces different subtleties; cf. () or ()).

XVII.1.3.10. We can also say that two measurable sets A and B in a measure space $(X, \mathcal{A}, \mu)$ are *almost disjoint* (or *μ-almost disjoint*) if $A \cap B$ is a null set.

The next technical result is very useful and is often applied in the later development of measure and integration:

XVII.1.3.11. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, with completion $(X, \bar{\mathcal{A}}, \bar{\mu})$. If f is an extended real-valued $\bar{\mathcal{A}}$-measurable function, then there is an $\mathcal{A}$-measurable function g with $g = f$ μ-almost everywhere.

PROOF: The proof is a good example of the technique of *bootstrapping*, proving something for all measurable functions by first proving it for indicator functions, then simple functions, and finally general functions by taking pointwise limits. In this case, suppose $f = \chi_C$, where $C \in \bar{\mathcal{A}}$. Then $C = A \cup B$, where $A \in \mathcal{A}$ and $B \subseteq D$ for some μ-null set $D \in \mathcal{A}$. Then $\chi_A = \chi_C$ on $X \setminus B$ and *a fortiori* on $X \setminus D$, and χ_A is $\mathcal{A}$-measurable. Next, if $f_1, \ldots, f_n$ are $\bar{\mathcal{A}}$-measurable functions equal a.e. to $\mathcal{A}$-measurable functions $g_1, \ldots, g_n$ respectively, and if $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, then $\alpha_1 f_1 + \cdots + \alpha_n f_n$ is equal a.e. to the $\mathcal{A}$-measurable function $\alpha_1 g_1 + \cdots + \alpha_n g_n$ since a finite union of null sets is a null set. In particular, any $\bar{\mathcal{A}}$-measurable simple function is equal a.e. to an $\mathcal{A}$-measurable simple function.

Now let f be a general $\bar{\mathcal{A}}$-measurable function. There is a sequence (f_n) of $\bar{\mathcal{A}}$-measurable simple functions converging pointwise to f. For each n, there is an μ-null set D_n and an $\mathcal{A}$-measurable simple function g_n such that $f_n = g_n$ on $X \setminus D_n$. Let $D = \cup_{n=1}^{\infty} D_n$. Then D is a μ-null set. Let $h_n = g_n = f_n$ on $X \setminus D$ and $h_n = 0$ on D. Then h_n is $\mathcal{A}$-measurable, and (h_n) converges pointwise to a function h which is $\mathcal{A}$-measurable and agrees with f on $X \setminus D$ (in fact, $h_n = g_n \chi_{X \setminus D} = f_n \chi_{X \setminus D}$ and $h = f \chi_{X \setminus D}$).

XVII.1.4. Finite, σ-Finite, and Semifinite Measures

Sets of infinite measure cause difficulties in measure theory. We cannot simply wish them away, however, since many of the most natural and important measure spaces (e.g. $\mathbb{R}$ with Lebesgue measure) have sets of infinite measure. We do want to work with sets of finite measure as much as possible, however. In this subsection we set up some terminology and develop a framework for doing this.

Finite Measure Spaces

The nicest measure spaces are ones with no sets of infinite measure:

XVII.1.4.1. DEFINITION. A measure space $(X, \mathcal{A}, \mu)$ is *finite*, and the measure μ is *finite*, if $\mu(X) < \infty$. The number $\mu(X)$ is the *total mass* of the measure space. The measure space is a *probability measure space*, and μ is a *probability measure*, if $\mu(X) = 1$.

The term "probability measure" comes of course from probability theory (), but the term is used throughout analysis even in contexts unrelated to probability. Any (nonzero) finite measure can be renormalized to a probability measure by taking an appropriate scalar multiple.

XVII.1.4.2. If $(X, \mathcal{A}, \mu)$ is a finite measure space with total mass m, then $\mu(A) \leq m$ for any $A \in \mathcal{A}$ by monotonicity. In particular, if μ is a probability measure, then $0 \leq \mu(A) \leq 1$ for all $A \in \mathcal{A}$.

σ-Finite Measures

It turns out to be almost as good for a set to be a countable union of sets of finite measure, and measure spaces which can be decomposed into countably many finite pieces are well behaved.

XVII.1.4.3. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space. A set $A \in \mathcal{A}$ is of *σ-finite measure* if there is a sequence (A_n) of sets in $\mathcal{A}$ with $A = \cup_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for all n. The measure space $(X, \mathcal{A}, \mu)$ is *σ-finite*, and μ is a *σ-finite measure*, if X is a set of σ-finite measure, i.e. if there is a sequence (X_n) in $\mathcal{A}$ with $X = \cup_{n=1}^{\infty} X_n$ and $\mu(X_n) < \infty$ for all n.

Note that in this definition, the sets A_n need not have any relation to each other, except they must of course all be measurable subsets of A. In particular, they do not have to be either nested or disjoint. However, if desired it can always be arranged that they are nested or that they are disjoint:

XVII.1.4.4. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $A \in \mathcal{A}$. The following are equivalent:

(i) A has σ-finite measure.

(ii) There is an increasing sequence (B_n) in $\mathcal{A}$ with $A = \cup_{n=1}^{\infty} B_n$ and $\mu(B_n) < \infty$ for all n.

(iii) There is a pairwise disjoint sequence (E_n) in $\mathcal{A}$ with $A = \cup_{n=1}^{\infty} E_n$ and $\mu(E_n) < \infty$ for all n.

PROOF: This is just a consequence of standard nesting or disjointizing operations. (ii) $\Rightarrow$ (i) and (iii) $\Rightarrow$ (i) are trivial. To show (i) $\Rightarrow$ (ii) and (i) $\Rightarrow$ (iii), set $B_n = \cup_{k=1}^{n} A_k$, $E_1 = A_1$, and $E_n = B_n \setminus B_{n-1}$ for $n > 1$. Note that $\mu(E_n) \leq \mu(B_n) \leq \sum_{k=1}^{n} \mu(A_k) < \infty$ by monotonicity and finite subadditivity. 👍

XVII.1.4.5. Corollary. Let $(X, \mathcal{A}, \mu)$ be a σ-finite measure space. Then

(i) There is an increasing sequence (Y_n) of sets in $\mathcal{A}$ with $X = \cup_{n=1}^{\infty} Y_n$ and $\mu(Y_n) < \infty$ for all n.

(ii) There is a pairwise disjoint sequence (Z_n) of sets in $\mathcal{A}$ with $X = \cup_{n=1}^{\infty} Z_n$ and $\mu(Z_n) < \infty$ for all n.

XVII.1.4.6. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space. Any measurable subset of a set of σ-finite measure has σ-finite measure. In particular, if the measure space is σ-finite, every measurable set in X has σ-finite measure.

Proof: If $A \in \mathcal{A}$ has σ-finite measure, let (A_n) be a sequence of sets in $\mathcal{A}$ with $A = \cup_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for all n. If $B \in \mathcal{A}$, $B \subseteq A$, set $B_n = B \cap A_n$ for each n. Then $B = \cup_{n=1}^{\infty} B_n$, and $\mu(B_n) < \infty$ by monotonicity. 👍

Decomposable Measures

The essential feature of σ-finite measure spaces is that the whole space can be chopped up into a countable number of disjoint pieces, each of finite measure. For some purposes it is sufficient that the space can be chopped up in a similar way even if there are uncountably many pieces.

XVII.1.4.7. Definition. A measure space $(X, \mathcal{A}, \mu)$ is *decomposable* (sometimes called *finitely decomposable*) if there is a collection $\{X_i : i \in I\}$ of pairwise disjoint sets in $\mathcal{A}$ with $X = \cup_i X_i$, $\mu(X_i) < \infty$ for all i, and

$$\mu(A) = \sum_{i \in I} \mu(A \cap X_i)$$

for all $A \in \mathcal{A}$ (where the sum is the supremum of the sums over finite subsets of I). By slight abuse of terminology we will also say that μ is decomposable.

XVII.1.4.8. It follows immediately from XVII.1.2.4. that a σ-finite measure space is decomposable; in fact, $(X, \mathcal{A}, \mu)$ is σ-finite if and only if it is decomposable with a countable set I. There are decomposable measures which are not σ-finite, e.g. counting measure on an uncountable set X (where we can take each X_i to be a singleton). There are also measures which are not decomposable, e.g. the ones in XVII.1.1.6.(ii)–(iii).

XVII.1.4.9. **Caution:** For a measure space to be decomposable, it is not enough that it decompose into a disjoint union of measurable sets of finite measure. The example of XVII.1.1.6.(iii) is a simple counterexample: the space is a disjoint union of singletons, each measurable and of finite measure. The problem is not that the subsets have measure 0; see XVII.1.10.7.. ✋

Semifinite Measures

There is a weaker condition than decomposability that is somewhat useful:

XVII.1.4.10. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space. A set $E \in \mathcal{A}$ is *semifinite*, or has *semifinite measure*, if, whenever $A \in \mathcal{A}$, $A \subseteq E$, and $\mu(A) = \infty$, there is a $B \in \mathcal{A}$, $B \subseteq A$, with $0 < \mu(B) < \infty$. $(X, \mathcal{A}, \mu)$ is a *semifinite measure space* if X has semifinite measure. By slight abuse of terminology we will also say that μ is semifinite.

XVII.1.4.11. It is obvious from the definition that every measurable subset of a set of semifinite measure also has semifinite measure. In particular, every measurable set in a semifinite measure space has semifinite measure.

XVII.1.4.12. PROPOSITION. Every decomposable measure space is semifinite. In particular, every σ-finite measure space is semifinite.

PROOF: Let $(X, \mathcal{A}, \mu)$ be decomposable. Let $\{X_i : i \in I\}$ be a pairwise disjoint collection of sets in $\mathcal{A}$ with $X = \cup_i X_i$, $\mu(X_i) < \infty$ for all i, and

$$\mu(A) = \sum_{i \in I} \mu(A \cap X_i)$$

for all $A \in \mathcal{A}$. If $\mu(A) = \infty$, then for any M there is a finite subset F of I such that $\sum_{i \in F} \mu(A \cap X_i) > M$. If $B = \cup_{i \in F}(A \cap X_i)$, then $B \subseteq A$ and by monotonicity

$$M < \mu(B) = \sum_{i \in F} \mu(A \cap X_i) \leq \sum_{i \in F} \mu(X_i) < \infty \ .$$



XVII.1.4.13. There are semifinite measures which are not decomposable (XVII.1.10.3., XVII.1.10.4.). There are also measures which are not semifinite, e.g. the ones in XVII.1.1.6.(ii)–(iii). Measure spaces which do not at least have many subsets of semifinite measure are so pathological that they are rarely useful; for example, integration theory for such measure spaces is completely degenerate.

XVII.1.4.14. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $E \in \mathcal{A}$ of semifinite measure. Then for every $A \in \mathcal{A}$, $A \subseteq E$, we have

$$\mu(A) = \sup\{\mu(B) : B \in \mathcal{A},\ B \subseteq A,\ \mu(B) < \infty\} \ .$$

Conversely, any E with this property has semifinite measure.

PROOF: The formula is trivial (in any measure space) by monotonicity if $\mu(A) < \infty$, so assume $\mu(A) = \infty$. Set

$$M = \sup\{\mu(B) : B \in \mathcal{A},\ B \subseteq A,\ \mu(B) < \infty\} \ .$$

If $M < \infty$, for each n let $B_n \in \mathcal{A}$, $B_n \subseteq A$, and $M - \frac{1}{n} \leq \mu(B_n) < \infty$. Set $D_n = \cup_{k=1}^n B_k$ for each n. Then (D_n) is an increasing sequence, $D := \cup_n D_n \subseteq A$, and

$$\mu(D_n) \leq \sum_{k=1}^n B_k < \infty$$

and therefore $\mu(D_n) \leq M$ for all n. We have $\mu(D) = \sup_n \mu(D_n) \leq M$ by XVII.1.2.7., and since $B_n \subseteq D$ for all n, by monotonicity we have

$$M - \frac{1}{n} \leq \mu(B_n) \leq \mu(D) \leq M$$

for all n. It follows that $\mu(D) = M$.

Since $\mu(A) = \infty$ and $\mu(D) < \infty$, we have $\mu(A \setminus D) = \infty$, and thus by semifiniteness of E there is an $C \in \mathcal{A}$, $C \subseteq A \setminus D$, with $0 < \mu(C) < \infty$. But then $C \cup D \subseteq A$, and

$$M < \mu(C \cup D) = \mu(C) + \mu(D) = \mu(C) + M < \infty$$

contradicting the definition of M. Thus $M = \infty$.

The converse statement is obvious. 👍

XVII.1.5. Discrete and Continuous Measures

XVII.1.5.1. Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space. An *atom* in $(X, \mathcal{A}, \mu)$ is a set $A \in \mathcal{A}$ with the property that $\mu(A) > 0$ and, whenever $B \in \mathcal{A}$, $B \subseteq A$, then either $\mu(B) = 0$ or $\mu(A \setminus B) = 0$.

An atom is a non-null measurable set which is "indivisible" up to null sets. The most obvious examples of atoms are (measurable) singleton subsets of positive measure, when such sets exist, although in general atoms need not be singletons. Measure spaces do not necessarily have any atoms; in fact, most interesting measure spaces do not.

XVII.1.6. Linear Combinations, Restrictions, and Extensions of Measures

We can take sums (finite or infinite) and scalar multiples of measures to obtain new measures, and we can in good circumstances restrict and extend measures. The arguments in this subsection are for the most part virtually trivial, with a few subtleties.

Addition of Measures

XVII.1.6.1. Proposition. Let μ and ν be measures on a measurable space $(X, \mathcal{A})$. Then $\mu + \nu$ is a measure on $(X, \mathcal{A})$. [By $\mu + \nu$ we mean the usual pointwise (setwise) sum, i.e. $(\mu + \nu)(A) = \mu(A) + \nu(A)$ for $A \in \mathcal{A}$, where as usual $x + \infty = \infty$ for $x \in [0, \infty]$.]

Proof: We have $(\mu + \nu)(\emptyset) = 0 + 0 = 0$. If (A_k) is a sequence of pairwise disjoint sets in $\mathcal{A}$, then

$$(\mu+\nu)\left(\bigcup_{k=1}^{\infty} A_k\right) = \mu\left(\bigcup_{k=1}^{\infty} A_k\right) + \nu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k) + \sum_{k=1}^{\infty} \nu(A_k) = \sum_{k=1}^{\infty} [\mu(A_k) + \nu(A_k)] = \sum_{k=1}^{\infty} (\mu+\nu)(A_k)$$

where combining the two infinite sums is justified by () since the terms are nonnegative. 👍

Note that we can only add measures defined on the *same* measurable space. We can even add a sequence of measures:

XVII.1.6.2. PROPOSITION. Let (μ_n) be a sequence of measures on a measurable space $(X, \mathcal{A})$. Define $\mu = \sum_{n=1}^{\infty} \mu_n$ on $(X, \mathcal{A})$ by

$$\mu(A) = \left(\sum_{n=1}^{\infty} \mu_n\right)(A) = \sum_{n=1}^{\infty} \mu_n(A)$$

for $A \in \mathcal{A}$. Then μ is a measure on $(X, \mathcal{A})$.

PROOF: We clearly have $\mu(\emptyset) = 0$. If (A_k) is a sequence of pairwise disjoint sets in $\mathcal{A}$, then

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{n=1}^{\infty} \mu_n\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{n=1}^{\infty}\left(\sum_{k=1}^{\infty} \mu_n(A_k)\right)$$

$$= \sum_{k=1}^{\infty}\left(\sum_{n=1}^{\infty} \mu_n(A_k)\right) = \sum_{k=1}^{\infty} \mu(A_k)$$

where the interchange of order in the double summation is justified by () since the terms are nonnegative. 👍

In fact, an arbitrary set of measures can be added in the same way. See () for this and more general results about combining measures.

Scalar Multiples of Measures

XVII.1.6.3. PROPOSITION. Let μ be a measure on a measurable space $(X, \mathcal{A})$, and $0 \leq \alpha < \infty$. Then $\alpha\mu$ is also a measure on $(X, \mathcal{A})$. [$\alpha\mu$ is defined by $[\alpha\mu](A) = \alpha[\mu(A)]$ for $A \in \mathcal{A}$; here by definition $0 \cdot \infty = 0$.]

PROOF: We obviously have $[\alpha\mu](\emptyset) = 0$. If (A_n) is a sequence of disjoint sets in $\mathcal{A}$, then

$$[\alpha\mu]\left(\bigcup_{k=1}^{\infty} A_k\right) = \alpha\left[\mu\left(\bigcup_{k=1}^{\infty} A_k\right)\right] = \alpha\sum_{k=1}^{\infty} \mu(A_k) = \sum_{k=1}^{\infty} \alpha[\mu(A_k)] = \sum_{k=1}^{\infty} [\alpha\mu](A_k)$$

where the next-to-last step is justified by (). 👍

Linear Combinations of Measures

Combining XVII.1.6.1. and XVII.1.6.3., we obtain:

XVII.1.6.4. PROPOSITION. Any nonnegative linear combination of measures on a measurable space is a measure. Specifically, if $\mu_1, \ldots, \mu_n$ are measures on a measurable space $(X, \mathcal{A})$, and $\alpha_1, \ldots, \alpha_n$ are nonnegative real numbers, then $\sum_{k=1}^{n} \alpha_k\mu_k$ is a measure on $(X, \mathcal{A})$, where $[\sum_{k=1}^{n} \alpha_k\mu_k](A) = \sum_{k=1}^{n} \alpha_k\mu_k(A)$ for $A \in \mathcal{A}$.

We have the same for countable sums, using XVII.1.6.2.:

XVII.1.6.5. PROPOSITION. If $\mu_1, \mu_2, \ldots$ is a sequence of measures on a measurable space $(X, \mathcal{A})$, and $\alpha_1, \alpha_2, \ldots$ is a sequence of nonnegative real numbers, then $\sum_{k=1}^{\infty} \alpha_k \mu_k$ is a measure on $(X, \mathcal{A})$, where $[\sum_{k=1}^{\infty} \alpha_k \mu_k](A) = \sum_{k=1}^{\infty} \alpha_k \mu_k(A)$ for $A \in \mathcal{A}$.

Restrictions and Compressions of Measures

XVII.1.6.6. DEFINITION. Let μ be a measure on a measurable space $(X, \mathcal{A})$, and $E \in \mathcal{A}$.

(i) The *restriction* of μ to E is the measure $\mu|_E$ on $(E, \mathcal{A}_E)$ defined by

$$[\mu|_E](A) = \mu(A)$$

for $A \in \mathcal{A}_E$ (i.e. $A \in \mathcal{A}$ and $A \subseteq E$).

(ii) The *compression* of μ to E is the measure $\mu \llcorner E$ on $(X, \mathcal{A})$ defined by

$$[\mu \llcorner E](A) = \mu(A \cap E)$$

for $A \in \mathcal{A}$.

The measures $\mu|_E$ and $\mu \llcorner E$ are in some sense the "same", but there is a technical difference: $\mu|_E$ is a measure on E, while $\mu \llcorner E$ is a measure on all of X which is supported on E, i.e. $[\mu \llcorner E](X \setminus E) = 0$.

It is a routine exercise (Exercise ()) to verify that $\mu|_E$ and $\mu \llcorner E$ are indeed measures.

XVII.1.6.7. If $(X, \mathcal{A}, \mu)$ is a measure space and $E \subseteq X$, the restriction of μ to E can still be defined even if $E \notin \mathcal{A}$, but it is a little more complicated. The collection $\mathcal{A}_E = \{A \cap E : A \in \mathcal{A}\}$ is a σ-algebra of subsets of E (although it can be quite small; the Countable AC is needed to prove this). If $A \in \mathcal{A}$, define

$$\mu|_E(A \cap E) = \sup\{\mu(B) : B \in \mathcal{A}, B \subseteq A \cap E\} \ .$$

If $A \cap E \in \mathcal{A}$ (and in particular if $E \in \mathcal{A}$), $\mu|_E(A \cap E) = \mu(A \cap E)$.

XVII.1.6.8. PROPOSITION. The function $\mu|_E$ is a measure on $(E, \mathcal{A}_E)$.

PROOF: We have $\mu|_E(\emptyset) = 0$, and $\mu|_E$ is monotone. For countable additivity, let (D_n) be a sequence of pairwise disjoint sets in $\mathcal{A}_E$, and set $D = \cup_n D_n$. If $\mu|_E(D_n) = \infty$ for some n, we have $\mu|_E = \infty$ by monotonicity. So we may assume $\mu|_E(D_n) < \infty$ for all n. Write $D_n = A_n \cap E$ for $A_n \in \mathcal{A}$ (using the Countable AC); then $D = A \cap E$, where $A = \cup_n A_n$. If $B \in \mathcal{A}$, $B \subseteq D$, set $B_n = B \cap A_n \subseteq D_n$. Then

$$\mu(B) = \sum_{n=1}^{\infty} \mu(B_n) \leq \sum_{n=1}^{\infty} \mu|_E(D_n)$$

and taking the supremum over all such B, we obtain

$$\mu|_E(D) \leq \sum_{n=1}^{\infty} \mu|_E(D_n) \ .$$

On the other hand, if $\epsilon > 0$, for each n let $B_n \in \mathcal{A}$ with $B \subseteq D_n$ and $\mu|_E(D_n) \leq \mu(B_n) + 2^{-n}\epsilon$, and set $B = \cup_n B_n$; then $B \in \mathcal{A}$, $B \subseteq D$, and

$$\mu|_E(D) \geq \mu(B) = \sum_{n=1}^{\infty} \mu(B_n) \geq \sum_{n=1}^{\infty} \mu|_E(D_n) - \epsilon$$

and since $\epsilon > 0$ is arbitrary, $\mu(D) \geq \sum_n \mu|_E(D_n)$. 👍

Extensions and Dilations of Measures

XVII.1.6.9. DEFINITION. Let $(X, \mathcal{A})$ be a measurable space, and $E \in \mathcal{A}$. If μ is a measure on $(X, \mathcal{A})$, and ν a measure on $(E, \mathcal{A}_E)$, then μ is an *extension* of ν to X if $\nu = \mu|_E$.

There is nothing unique about extensions of measures in general.

There is always a minimal extension of any measure:

XVII.1.6.10. DEFINITION. Let $(X, \mathcal{A})$ be a measurable space, and $E \in \mathcal{A}$. If ν is a measure on $(E, \mathcal{A}_E)$, the *dilation* of ν to $(X, \mathcal{A})$ is the measure $\nu \uparrow X$ defined by

$$[\nu \uparrow X](A) = \nu(A \cap E)$$

for $A \in \mathcal{A}$.

XVII.1.6.11. It is easily verified that $\nu \uparrow X$ is a measure on $(X, \mathcal{A})$, that $(\nu \uparrow X)$ is an extension of ν to X, that $\nu \uparrow X = (\nu \uparrow X)\llcorner_E$, and that $\nu \uparrow X$ is the unique measure on (X, A) with these properties and the smallest extension of ν to X (in the sense that if μ is any extension, then $\mu \geq (\nu \uparrow X)$.) In fact, if μ is any measure on $(X, \mathcal{A})$, then $\mu = ([\mu|_E] \uparrow X) + ([\mu|_{X\setminus E}] \uparrow X) = (\mu\llcorner_E) + (\mu\llcorner_{X\setminus E})$.

Restriction of Measures to Sub-σ-Algebras

There is another type of restriction of measures:

XVII.1.6.12. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$. Then the *restriction* of μ to $\mathcal{B}$ is the measure $\mu|_{\mathcal{B}}$ on $(X, \mathcal{B})$ defined by

$$[\mu|_{\mathcal{B}}](B) = \mu(B)$$

for $B \in \mathcal{B}$.

Note that $\mu|_{\mathcal{B}}$ is technically a different measure than μ since the specification of the σ-algebra of measurable subsets is part of the definition of a measure. There are occasions, however, where making an explicit distinction between μ and $\mu|_{\mathcal{B}}$ seems overly pedantic, and we will occasionally overlook the distinction when no confusion can occur. The most important example is restriction of Lebesgue measure to the Borel sets.

XVII.1.6.13. We can also regard the measure μ as an *extension* of $\mu|_{\mathcal{B}}$ to the larger σ-algebra $\mathcal{A}$. Existence of extensions of measures to larger σ-algebras is a very subtle matter and is not possible in general; see II.10.1.1., XVII.1.3.7., () for some instances of extensions and nonexistence of extensions.

Push-Forward of Measures

We can push measures forward by measurable functions:

XVII.1.6.14. DEFINITION. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measure spaces, and $f : X \to Y$ an $(\mathcal{A}, \mathcal{B})$-measurable function (). If μ is a measure on $(X, \mathcal{A})$, then the *push-forward* of μ by f is the measure $f_*(\mu)$ on $(Y, \mathcal{B})$ defined by

$$[f_*(\mu)](B) = \mu(f^{-1}(B))$$

for $B \in \mathcal{B}$.

As usual, it is routine to show that $f_*(\mu)$ is a measure on $(Y, \mathcal{B})$.

XVII.1.6.15. EXAMPLES. (i) If $(X, \mathcal{A})$ is a measurable space and $E \in \mathcal{A}$, and $f : E \to X$ is the inclusion map, and ν is a measure on $(E, \mathcal{A}_E)$, then $f_*(\nu) = \nu \uparrow X$.

(ii) Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$. Let f be the identity map on X, regarded as a measurable function from $(X, \mathcal{A})$ to $(X, \mathcal{B})$. Then $f_*(\mu) = \mu|_{\mathcal{B}}$.

XVII.1.6.16. There is no reasonable way to pull back measures from $(Y, \mathcal{B})$ to $(X, \mathcal{A})$ in general, but this can be done under some special circumstances (e.g. ()).

XVII.1.7. Measures Defined on Rings and σ-Rings

XVII.1.8. Egorov's Theorem

XVII.1.8.1. THEOREM. [EGOROV'S THEOREM] Let $(X, \mathcal{A}, \mu)$ be a finite measure space. Let (f_n) be a sequence of real-valued measurable functions on X converging pointwise a.e. to a real-valued (measurable) function f. Then for every $\epsilon > 0$ there is a set $A \in \mathcal{A}$ with $\mu(X \setminus A) < \epsilon$ and $f_n \to f$ uniformly on A.

PROOF: Fix $\epsilon > 0$. Let $E \in \mathcal{A}$ with $f_n(x) \to f(x)$ for all $x \in E$ and $\mu(X \setminus E) = 0$. For each n and m, set

$$A_{n,m} = \left\{ x \in X : |f_k(x) - f(x)| < \frac{1}{m} \text{ for all } k \geq n \right\} .$$

We have $A_{n,m} \in \mathcal{A}$ for all n, m. For fixed m, we have $A_{n,m} \subseteq A_{n+1,m}$ for all n and $\cup_n A_{n,m} \supseteq E$, i.e. $\cap_n [X \setminus A_{n,m}] \subseteq X \setminus E$. Thus by ()

$$0 \leq \inf_n \mu(X \setminus A_{n,m}) \leq \mu(X \setminus E) = 0$$

and thus there is an n_m such that $\mu(X \setminus A_{n_m,m}) < \frac{\epsilon}{2^m}$.

Let $A = \cap_{m=1}^{\infty} A_{n_m,m}$. Then

$$\mu(X \setminus A) = \mu\left(\bigcup_{m=1}^{\infty} [X \setminus A_{n_m,m}] \right) \leq \sum_{m=1}^{\infty} \mu(X \setminus A_{n_m,m}) < \sum_{m=1}^{\infty} \frac{\epsilon}{2^m} = \epsilon$$

and $f_n \to f$ uniformly of A: if $\delta > 0$, fix m with $\frac{1}{m} \leq \delta$; then, for all $k \geq n_m$ and all $x \in A \subseteq A_{n_m,m}$, we have

$$|f_k(x) - f(x)| < \frac{1}{m} \leq \delta \ .$$



XVII.1.8.2. The first version of this theorem was stated and proved by C. Severini in 1910 and by Egorov independently shortly afterwards. The general version above is due to F. Riesz and W. Sierpiński. Further generalizations were given by Korovkin () and Mokobodski ().

XVII.1.9. Measures on δ-Rings

Measures can be defined on σ-rings, and more generally on δ-rings, with virtually no modification in the theory, and some authors, e.g. [Hal50], take this approach. There is little advantage to this increased generality for ordinary measures, but it is very useful for signed and complex measures ().

XVII.1.10. Exercises

XVII.1.10.1. Let μ be a finitely additive measure on a measurable space $(X, \mathcal{A})$. Show that the following are equivalent:

(i) μ is countably subadditive (XVII.1.2.1.).

(ii) μ is upward continuous (XVII.1.2.7.).

(iii) μ is countably additive, i.e. a measure.

XVII.1.10.2. Let μ be counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Show that for any $n \in \{0\} \cup \mathbb{N} \cup \{\infty\}$, there are subsets A, B of $\mathbb{N}$ with $B \subseteq A$, $\mu(A) = \mu(B) = \infty$, and $\mu(A \setminus B) = n$.

XVII.1.10.3. Let X be an uncountable set, and x_0 a fixed element of X. Let $\mathcal{A}$ be the collection of all countable subsets of X not containing x_0, and their complements.

(a) Show that $\mathcal{A}$ is a σ-algebra on X.

(b) Let μ be the restriction of counting measure to $\mathcal{A}$. Show that μ is semifinite.

(c) The point x_0 is not contained in any measurable set of finite measure. Conclude that μ cannot be decomposable.

(d) If the cardinality of X is κ, then the cardinality of $\mathcal{A}$ is $\kappa^{\aleph_0}$. In particular, if $card(X) = 2^{\aleph_0}$, then $card(\mathcal{A}) = 2^{\aleph_0}$.

XVII.1.10.4. [Bog07,] Let X be a set of cardinality $\aleph_1$, Y a set of cardinality $\aleph_2$, and $Z = X \times Y$. If $E \subseteq Z$, $x \in X$, and $y \in Y$, write E_x and E^y for the vertical and horizontal cross sections of E at x and y respectively ().

(a) Let $\mathcal{M}$ be the set of all subsets E of Z such that for every $x \in X$, E_x is either countable or cocountable, and for every $y \in Y$, E^y is either countable of cocountable. Show that $\mathcal{M}$ is a σ-algebra.

(b) If $E \in \mathcal{M}$, define

$$\mu(E) = card\{x \in X : E_x \text{ is cocountable}\} + card\{y \in Y : E^y \text{ is cocountable}\}$$

(where $\mu(E) = \infty$ if either set on the right is infinite). Show that μ is a measure on $(Z, \mathcal{M})$. Every countable subset (in particular, every singleton subset) of X is in $\mathcal{M}$ and has measure 0.

(c) Show that μ is semifinite. [If $\mu(E) > 0$, then E contains a set of the form $A \times \{y\}$ for a cocountable subset A of X and some $y \in Y$, or a set of the form $\{x\} \times B$ for cocountable $B \subseteq Y$.]

(d) Suppose μ is decomposable, and let $\{Z_i : i \in I\}$ be a decomposition. For each $y \in Y$, show that the set $(Z_i)^y$ is cocountable for exactly one i.

(e) If $y \in Y$, let $\phi(y)$ be the unique i as in (d). Then ϕ is a function from Y to I. For any $i \in I$, $\phi^{-1}(\{i\})$ is finite. Conclude that $card(I) \geq \aleph_2$.

(f) On the other hand, for each $x \in X$ there is an i such that $(Z_i)_x$ is cocountable. Thus there are only countably many i such that $(Z_i)_x$ is nonempty. Conclude that $card(I) \leq \aleph_1$.

(g) Conclude from (d)–(f) that μ is not decomposable.

XVII.1.10.5. Let I be an uncountable set, and for each $i \in I$ let X_i be a copy of $[0, 1]$. Let X be the disjoint union of the X_i. Let $\mathcal{M}$ be the σ-algebra of Lebesgue measurable sets in $[0, 1]$, and let

$$\mathcal{B} = \{B \subseteq X : B \cap X_i \in \mathcal{M} \text{ for all } i \in I\}$$

and $\mathcal{A}$ the collection of all $A \in \mathcal{B}$ such that $A \cap X_i$ is either $\emptyset$ or X_i for all but countably many i.

(a) Show that $\mathcal{B}$ and $\mathcal{A}$ are σ-algebras.

(b) For $A \in \mathcal{A}$, set $\mu(A) = \sum_{i \in I} \lambda(A \cap X_i)$, where λ is Lebesgue measure on X_i. Show that μ is a complete measure on $(X, \mathcal{A})$.

(c) Show that $\{X_i : i \in I\}$ is a decomposition of $(X, \mathcal{A}, \mu)$, so μ is decomposable and, in particular, semifinite.

(d) If $B \in \mathcal{B}$, set $\nu(B) = \mu(B)$ if $B \in \mathcal{A}$ and $\nu(B) = \infty$ if $B \in \mathcal{B} \setminus \mathcal{A}$. Show that ν is a complete measure on $(X, \mathcal{B})$ extending μ.

(e) If B is a subset of X such that $B \cap X_i$ is a singleton for all i, then $\nu(B) = \infty$, but B contains no subsets of finite positive ν-measure. Thus ν is not semifinite.

Note that this problem, including (e), can be done without using the Axiom of Choice by taking X to be the unit square and the X_i to be (say) horizontal cross sections. The set B of (e) can be taken to be a vertical line segment.

XVII.1.10.6. Let $(X, \mathcal{A}, \mu)$ be a semifinite measure space.

(a) If $\mathcal{A}$ contains only countably many sets of finite measure, show that μ is σ-finite.

(b) [R. Israel] If $\mathcal{A}$ contains only $\aleph_1$ sets of finite measure, and every point of X is contained in a measurable subset of finite measure, then A is decomposable. [Well order the sets of finite measure as $(A_\alpha : \alpha < \omega_1)$, and for each α let $X_\alpha = A_\alpha \setminus (\cup_{\beta<\alpha} A_\beta)$.]

The last condition in (b) is not automatic, at least if the continuum hypothesis is assumed (XVII.1.10.3.).

XVII.1.10.7. Let Y be an uncountable set and ν the measure on $(Y, \mathcal{P}(Y))$ with $\nu(A) = 0$ if A is countable and $\nu(A) = \infty$ if A is uncountable. Let $X = Y \times \{0, 1\}$, and for $A, B \subseteq Y$ set

$$\mu([A \times \{0\}] \cup [B \times \{1\}]) = \nu(A) + |B|$$

where $|B|$ is the number of elements in B (counting measure of B).

(a) Show that μ is a measure on $(X, \mathcal{P}(X))$.

(b) For each $y \in Y$ set $X_y = \{(y, 0), (y, 1)\}$. Then X is the disjoint union of the X_y and $\mu(X_y) = 1$ for all y. In particular, every point of X is contained in a measurable subset of measure 1.

(c) Nonetheless, $(X, \mathcal{P}(X), \mu)$ is not decomposable or even semifinite.

XVII.2. The Metric Space of a Measure Space

If $(X, \mathcal{A}, \mu)$ is a measure space, there is a natural pseudometric on $\mathcal{A}$ for which the metric quotient is complete.

Recall that if A and B are sets, the *symmetric difference* of A and B is

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B) \ .$$

$A \triangle B$ consists of elements which are in either A or B, but not both. The size of $A \triangle B$ is a measure of how much the sets A and B differ.
If $\mathcal{A}$ is an algebra of subsets of a set X, and $A, B \in \mathcal{A}$, then $(A \triangle B) \in \mathcal{A}$.

Here are some important simple properties of symmetric difference:
PROPOSITION. Let A, B, C be sets. Then

(i) $(A \triangle B) \triangle C = A \triangle (B \triangle C)$.

(ii) $A \triangle C \subseteq (A \triangle B) \cup (B \triangle C)$.

(iii) $A \triangle B = B \triangle A$.

(iv) $A \triangle A = \emptyset$.

(v) $A \triangle \emptyset = A$.

DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space. If $A, B \in \mathcal{A}$, set

$$\rho_\mu(A, B) = \min(\mu(A \triangle B), 1) \ .$$

PROPOSITION. If $(X, \mathcal{A}, \mu)$ is a measure space, then ρ_μ is a pseudometric on $\mathcal{A}$.

PROOF: It is obvious from XVII.2.(iii) that $\rho_\mu(A, B) = \rho_\mu(B, A) \geq 0$ for any $A, B \in \mathcal{A}$, and the triangle inequality follows from XVII.2.(ii) and monotonicity of μ. 👍

There is inconsistency in the exact way ρ_μ is defined in the literature, but the differences are not important since all standard definitions give uniformly equivalent metrics. The "natural" definition is just $\rho_\mu(A, B) = \mu(A \triangle B)$, but this can take the value ∞, which is not allowed for a pseudometric.

As usual, we can take a quotient of $\mathcal{A}$ on which ρ_μ defines a metric. The equivalence class of a set $A \in \mathcal{A}$ is

$$[A] = \{B \in \mathcal{A} : \mu(A \triangle B) = 0\} \ .$$

The set of equivalence classes is often written $\mathcal{A}/\mathcal{N}$, where $\mathcal{N}$ is the σ-ideal of μ-null sets. We will denote the induced metric on $\mathcal{A}/\mathcal{N}$ also by ρ_μ.
If (A_n) is a sequence of sets, recall the definition of $\limsup_{n\to\infty} A_n$ and $\liminf_{n\to\infty} A_n$ from ().

THEOREM. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\mathcal{N}$ the set of μ-null sets. Let (A_n) be a Cauchy sequence in $\mathcal{A}$ for ρ_μ. Then

$$\left[\left(\limsup_{n\to\infty} A_n\right) \setminus \left(\liminf_{n\to\infty} A_n\right)\right] \in \mathcal{N}$$

and $[\limsup A_n] = [\liminf A_n] = \lim_{n\to\infty} [A_n]$ in $\mathcal{A}/\mathcal{N}$. Thus $(\mathcal{A}/\mathcal{N}, \rho_\mu)$ is a complete metric space.

Note that this result has nothing to do with completeness of μ. In fact, if $(X, \bar{\mathcal{A}}, \bar{\mu})$ is the completion of $(X, \mathcal{A}, \mu)$, and $\bar{\mathcal{N}}$ is the set of $\bar{\mu}$-null sets, there is a natural bijection between $\bar{\mathcal{A}}/\bar{\mathcal{N}}$ and $\mathcal{A}/\mathcal{N}$, and $\rho_{\bar{\mu}}$ agrees with ρ_μ on this space, so the metric space of $\bar{\mu}$ is the "same" as for μ.

The usual set operations are continuous with respect to the topology on $\mathcal{A}$:

THEOREM. Let $(X, \mathcal{A}, \mu)$ be a measure space and $(\mathcal{A}/\mathcal{N}, \rho_\mu)$ the corresponding metric space. Then

(i) The functions $([A], [B]) \mapsto [A \cup B]$, $([A], [B]) \mapsto [A \cap B]$, $([A], [B]) \mapsto [A \setminus B]$, and $([A], [B]) \mapsto [A \triangle B]$ are well defined and jointly continuous, i.e. continuous functions from $(\mathcal{A}/\mathcal{N}) \times (\mathcal{A}/\mathcal{N})$ to $\mathcal{A}/\mathcal{N}$.

(ii) The function $[A] \mapsto [X \setminus A]$ is a well-defined continuous function from $\mathcal{A}/\mathcal{N}$ to $\mathcal{A}/\mathcal{N}$.

(iii) If $\mathcal{B}$ is an algebra of sets containing $\mathcal{N}$, the closure of $\mathcal{B}/\mathcal{N}$ in $\mathcal{A}/\mathcal{N}$ is $\sigma(B)/\mathcal{N}$, where $\sigma(B)$ is the σ-algebra generated by $\mathcal{B}$.

XVII.3. Probability

"How can we venture to speak of the laws of chance? Is not chance the antithesis of all law?"

J. Bertrand[1]

XVII.3.1. Introduction

XVII.3.1.1. From its first beginnings, probability theory had some of the flavor of primitive measure theory. In 1933, A. KOLMOGOROV [**?**] cemented the connection by giving an axiomatization of probability which serves as the foundation of all modern probability theory. (Of course, KOLMOGOROV did not come up with this axiomatization out of thin air; the ideas were implicit and in fact increasingly explicit in other work in the field.)

XVII.3.1.2. The two basic ingredients of probability are:

(i) A *sample space* Ω, the set of all possible outcomes of a specific process or experiment. A subset of Ω is called an *event*.

(ii) A *probability* $\mathbf{P}$, an assignment to each event E a numerical probability or likelihood $\mathbf{P}(E)$ that E occurs (i.e. that the outcome of the process lies in E). We should have $0 \leq \mathbf{P}(E) \leq 1$ for any E, and $\mathbf{P}(\Omega) = 1$.

XVII.3.1.3. It is virtually a no-brainer that $\mathbf{P}$ should be finitely additive. KOLMOGOROV realized that in order to get a satisfactory theory, one needs $\mathbf{P}$ to actually be countably additive.

XVII.3.1.4. Examples show that it is not possible in general to assign a probability to every subset of Ω in a countably additive way (the process described in I.3.1.17., along with XVII.4.10.5. and II.10.3.6., is an example where it is not reasonably possible). Thus probability can only be assigned to some collection of well-behaved events. This turns out not to be a serious drawback; indeed, experience shows that it is frequently useful to deliberately restrict the set of events to which a probability is assigned.

Nevertheless, it is important to be able to apply sequences of set-theoretic operations to allowable events and stay within the class of allowable events. This leads naturally to the assumption that the set of allowable events forms a σ-algebra of subsets of Ω.

Thus KOLMOGOROV was led to make the following definition:

XVII.3.1.5. DEFINITION. A *probability space* is a triple $(\Omega, \mathcal{E}, \mathbf{P})$, where

Ω is a set, called the *sample space*.

$\mathcal{E}$ is a σ-algebra of subsets of Ω, called the set of *(allowable) events*.

$\mathbf{P}$ is a function from $\mathcal{E}$ to $[0, 1]$ satisfying

(i) $\mathbf{P}(\Omega) = 1$.

[1] [Ber72]; cf. [Poi96, p. 64].

(ii) $\mathbf{P}$ is *countably additive*, i.e. if (E_n) is a sequence of pairwise disjoint sets in $\mathcal{E}$, then

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbf{P}(E_n) \ .$$

Such a $\mathbf{P}$ is called a *probability* on $(\Omega, \mathcal{E})$. If $E \in \mathcal{E}$, $\mathbf{P}(E)$ is called the *probability* of the event E.

XVII.3.1.6. If $(\Omega, \mathcal{E}, \mathbf{P})$ is a probability space, then $(\Omega, \mathcal{E})$ is a measurable space (). We have

$$\mathbf{P}(\emptyset) = \mathbf{P}(\emptyset \cup \emptyset \cup \cdots) = \sum_{n=1}^{\infty} \mathbf{P}(\emptyset) = \infty \cdot \mathbf{P}(\emptyset)$$

and thus $\mathbf{P}(\emptyset) = 0$ (since it is in $[0, 1]$). Therefore $(\Omega, \mathcal{E}, \mathbf{P})$ is a measure space (XVII.1.1.1.), with $\mathbf{P}$ a probability measure. Conversely, any probability measure space is a probability space. Thus probability spaces are precisely the same thing as probability measure spaces, with slightly different notation and terminology.

Therefore, the results of XVII.1.2. apply in probability spaces:

XVII.3.1.7. Proposition. Let $(\Omega, \mathcal{E}, \mathbf{P})$ be a probability space. Then

(i) $\mathbf{P}$ is *finitely additive*: if $E_1, \ldots, E_n \in \mathcal{E}$ are pairwise disjoint, then

$$\mathbf{P}\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{k=1}^{n} \mathbf{P}(E_k) \ .$$

In particular, if $E, F \in \mathcal{E}$ are disjoint, then $\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F)$.

(ii) $\mathbf{P}$ is *monotone*: if $E, F \in \mathcal{E}$, $E \subseteq F$, then $\mathbf{P}(E) \leq \mathbf{P}(F)$.

(iii) $\mathbf{P}$ is *countably subadditive*: if (E_n) is a sequence in $\mathcal{E}$, then

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} \mathbf{P}(E_n) \ .$$

$\mathbf{P}$ is thus also finitely subadditive; in particular, if $E, F \in \mathcal{E}$, then $\mathbf{P}(E \cup F) \leq \mathbf{P}(E) + \mathbf{P}(F)$. (In fact, we have $\mathbf{P}(E \cup F) = \mathbf{P}(E) + \mathbf{P}(F) - \mathbf{P}(E \cap F)$.)

(iv) $\mathbf{P}$ is *upward continuous*: If (E_n) is an increasing sequence in $\mathcal{E}$, then

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sup_n \mathbf{P}(E_n) = \lim_{n\to\infty} \mathbf{P}(E_n) \ .$$

(v) $\mathbf{P}$ is *downward continuous*: If (E_n) is a decreasing sequence in $\mathcal{E}$, then

$$\mathbf{P}\left(\bigcap_{n=1}^{\infty} E_n\right) = \inf_n \mathbf{P}(E_n) = \lim_{n\to\infty} \mathbf{P}(E_n) \ .$$

XVII.3.2. The Dictionary

The mathematical setup of probability is thus measure theory, using somewhat different language. Here is a dictionary of the correspondence of basic terminology between measure theory and probability. Some other terminology such as "distribution" is common to both.

Measure Theory	Probability
Measurable space	Sample space
Measurable set	Event
Measure	Probability
Real-valued measurable function	Random variable
Integral	Expectation
Almost everywhere (a.e.)	Almost surely (a.s.)
–	Independent

In fact, it is sometimes said that probability is "measure theory plus independence" (although this is clearly somewhat of an oversimplification). It is unfair and not really accurate to say that probability is *part* of measure theory, although modern probability is certainly built on measure theory.

XVII.3.3. Exercises

XVII.3.3.1. A Conditional Probability Paradox?

(a) You have n red balls, n white balls, and n blue balls ($n \geq 3$) in an urn. Your assistant chooses three balls at random from the urn and puts them in a box. What is the probability p that all three balls in the box have the same color?

(b) While the assistant is choosing the balls, your second assistant comes into the room and sees some of what the first assistant is doing. He tells you that he saw one of the balls the first assistant put in the box and it was red. What now is the conditional probability q that all the balls in the box are the same color, given that at least one of the balls is red? Explain why $q < p$, especially since if the second assistant had said that the ball he saw was white, or blue, the conditional probability is also q, and one of these situations must occur, with equal probability since the colors are completely symmetric.

(c) Suppose your second assistant doesn't speak much English, and you know not a word of his native language. He reports to you that the color of the ball he saw was *gloop*, but can't remember the English word for this color. What is the conditional probability that the three balls are the same color, given that at least one of the balls is *gloop*? On the one hand, no matter what *gloop* means, the conditional probability is q; on the other hand, he has given you no information, so the conditional probability should just be the probability p. Explain and resolve the paradox.

(d) The first assistant speaks the second assistant's language fluently, but is temporarily unavailable to translate. Anyway, the first assistant can provide more information than just the meaning of *gloop*; in fact, he knows what the colors of the three balls are. How does this affect the conditional probabilities, if at all?

(e) Do p and/or q depend on n? Explain why or why not.

XVII.3.3.2. ([Bar00], [Sav94]) Suppose we have three fair (cubical) dice whose faces are marked as follows:

$$\#1: \ 18, \ 9, \ 8, \ 7, \ 6, \ 5$$

$$\#2: \ 17, \ 16, \ 15, \ 4, \ 3, \ 2$$

$$\#3: \ 14, \ 13, \ 12, \ 11, \ 10, \ 1$$

We toss the dice, and let A, B, and C be the numbers on dice 1, 2, and 3 respectively. Then A, B, and C are random variables. Note that A, B, and C always have different values.

(a) Show that the probabilities $P(A > B)$, $P(B > C)$, and $P(C > A)$ are all greater than $\frac{1}{2}$.

(b) Show that $P(B > A \text{ and } B > C)$ (i.e. B is the largest of the three) is greater than $P(A > B \text{ and } A > C)$ and $P(C > A \text{ and } C > B)$.

(c) Show that the expected values satisfy $\mathbf{E}(A) < \mathbf{E}(B) < \mathbf{E}(C)$.

XVII.4. Measures Defined by Extensions

The most important construction of measures is by the Carathéodory extension process. One begins with a "premeasure" defined on a collection $\mathcal{C}$ of subsets of a set X, and extends the premeasure to a measure on a σ-algebra containing $\mathcal{C}$. The extension is made via an "outer measure" defined on all subsets of X.

Let us informally outline the process and try to give some motivation for the procedure, which can otherwise seem strange and nonintuitive at first glance. First consider how to define the "area" of a region in the plane. If A is a bounded subset of the plane $\mathbb{R}^2$, a reasonable and commonly-used way to define the area of A is to cover it with a rectangular grid $\mathcal{G}$, and let $\overline{V}(A, \mathcal{G})$ be the sum of the areas of the rectangles in the grid which contain points of A, and $\underline{V}(A, \mathcal{G})$ the sum of the areas of the rectangles in the grid entirely contained in A. $\overline{V}(A, \mathcal{G})$ and $\underline{V}(A, \mathcal{G})$ can be regarded as an "upper" or "outer" estimate, and a "lower" or "inner" estimate, respectively, for the area of A. As $\mathcal{G}$ becomes finer, the outer estimates decrease and the inner estimates increase, and for "nice" A they converge to the same value, which is a reasonable definition of the area of A. See IX.3.2.1. and IX.3.2.10..

It turns out that not enough subsets of the plane are "nice" enough to have an area defined in this way to give a satisfactory theory of measure and integration (i.e. such sets do not form a σ-algebra). But the idea of approximating from the outside and inside is one which, with suitable care, gives a good theory. We begin with a collection $\mathcal{C}$ of sets which have a reasonable "size" already specified, and, to get an "outer estimate" of the "size" of a general set A, consider collections (C_n) of sets in $\mathcal{C}$ which cover A (whose union contains A); the sum of the "sizes" of the C_n is an upper estimate for the "size" of A. There are two differences we allow from the grid covering example, which together give us a notion of "size" with considerably more flexibility:

(1) We allow countable collections (C_n), not just finite collections.

(2) We do not insist that the C_n are disjoint or even "nonoverlapping."

The infimum of all such upper estimates will be called the "outer measure" of the set A, written $\mu^*(A)$.

For the "inner" estimate, it turns out that the allowance of countably infinite covers is incompatible with truly estimating size from the "inside" as in the grid method. But in the grid example, there is a simple change of point of view which does work in the new setting: if R is a large rectangle which is subdivided to make the grid $\mathcal{G}$, it is easily seen that if $V(R)$ is the area of R, then $\underline{V}(A, \mathcal{G}) = V(R) - \overline{V}(R \setminus A, \mathcal{G})$. So, if we are working within subsets of a fixed set Y of finite measure, we can replace inner estimates for the size of our set A by outer estimates for the size of $Y \setminus A$. The criterion that the inner and outer estimates for the size of A should converge to the same limit as the covers get "finer" translates into the requirement that the sum of the outer measure $\mu^*(A)$ and the outer measure $\mu^*(Y \setminus A)$ should be the measure of Y. If the whole set X has finite measure, it turns out that for A to be "measurable" (to have a well-behaved notion of "size") we need only require this condition for $Y = X$; however, in the general setting we must have a technical strengthening: for every $Y \subseteq X$, we require

$$\mu^*(Y \cap A) + \mu^*(Y \cap A^c) = \mu^*(Y) \ .$$

Roughly speaking, the set A is measurable if, for every $Y \subseteq X$, the sets $Y \cap A$ and $Y \cap A^c$ can be finely covered by "almost disjoint" countable families from $\mathcal{C}$.

It turns out then that under reasonable conditions on the size function ("premeasure") on $\mathcal{C}$ we can prove that the collection of "measurable" subsets of X is a σ-algebra containing $\mathcal{C}$, and that μ^* restricts to a measure on this σ-algebra which agrees with the premeasure on $\mathcal{C}$, and in particular the premeasure on $\mathcal{C}$ extends to a measure on a σ-algebra containing $\mathcal{C}$.

We will do most of the construction for "outer measures" defined abstractly, which have the crucial properties of outer measures induced by premeasures. This is essential since there are important outer measures which are not induced (at least not in any useful way) from premeasures.

XVII.4.1. Outer Measures

Outer measures are a crucial technical tool in measure theory. Almost all interesting measures are constructed (at least indirectly) from outer measures.

XVII.4.1.1. DEFINITION. Let X be a set. An *outer measure* on X is a function $\mu^* : \mathcal{P}(X) \to [0, \infty]$ such that

(i) $\mu^*(\emptyset) = 0$

(ii) If $A \subseteq B$, then $\mu^*(A) \leq \mu^*(B)$ (*monotonicity*)

(iii) If $A = \cup_{n=1}^{\infty} A_n$, then $\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$ (*countable subadditivity*).

There are two important differences between measures and outer measures. First, an outer measure is defined on *all* subsets of X; second, an outer measure is not even finitely additive, let alone countably additive, in general.

XVII.4.1.2. In the case of a measure, monotonicity (condition (ii)) follows from the other properties (), but (ii) does not follow from (i) and (iii) for outer measures in general (take Example XVII.4.5.28. with $\alpha < 1$). In the presence of (i), both (ii) and (iii) can be replaced by the single condition

(iii′) If $A \subseteq \cup_{n=1}^{\infty} A_n$, then $\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$.

[(iii′) obviously follows from (ii) and (iii), and (iii′) trivially implies (iii); for (iii′) $\Rightarrow$ (ii), if $A \subseteq B$, set $A_1 = B$ and $A_n = \emptyset$ for $n > 1$, and use (i).]

XVII.4.1.3. As with measures, countable subadditivity for outer measures implies *finite subadditivity*: if $A \subseteq \cup_{k=1}^{n} A_k$, then $\mu^*(A) \leq \sum_{k=1}^{n} \mu^*(A_k)$. This follows from countable subadditivity by taking $A_k = \emptyset$ for $k > n$. In particular, if $A, B \subseteq X$, then $\mu^*(A \cup B) \leq \mu^*(A) + \mu^*(B)$.

XVII.4.1.4. **Note:** It must be noted that a substantial number of references, particularly ones concerned with geometric measure theory, call an outer measure a "measure"; in these references, a measure on a set X is defined on all subsets of X, but is only countably additive on a σ-algebra of "measurable" subsets of X. The distinction is largely a semantic one: every measure has a canonical extension to an outer measure, and conversely every outer measure naturally defines a σ-algebra of measurable subsets on which it is countably additive (a measure in our sense). The two processes are almost, but not quite, inverses of each other, however. The terminology we use is more flexible and superior for most purposes: for example, it is often crucial to keep track of the σ-algebra on which a measure is defined, particularly in probability where it is often necessary to consider and distinguish between two or more σ-algebras.

XVII.4.1.5. Most outer measures are constructed beginning with a "measure" defined on a collection of subsets which is too small to be a σ-algebra, usually a semiring. Suppose $\mathcal{C}$ is a collection of subsets of a set X, with the property that $\emptyset \in \mathcal{C}$. Let ι be a function from $\mathcal{C}$ to $[0, \infty]$. We will eventually impose some reasonable measure-like conditions on $\mathcal{C}$ and ι, but to get an outer measure the only requirement on ι is that $\iota(\emptyset) = 0$.

Define a function μ^* from $\mathcal{P}(X)$ to $[0, \infty]$ as follows. If $A \subseteq X$, set

$$\mu^*(A) = \inf \left\{ \sum_{n=1}^{\infty} \iota(C_n) \ : \ C_n \in \mathcal{C}, \ A \subseteq \bigcup_{n=1}^{\infty} C_n \right\} .$$

If A has no countable cover by sets in $\mathcal{C}$, we have a choice: it is standard to set $\mu^*(A) = \infty$, which we will do, but there is another possibility (XVII.4.5.16.).

If $A \subseteq \bigcup_{n=1}^{\infty} C_n$, we say the sequence $\{C_n\}$ *covers* A.

Note that since $\emptyset \in \mathcal{C}$, finite as well as countably infinite covers are allowed in computing $\mu^*(A)$, since a finite cover $\{C_1, \ldots, C_m\}$ of A can be expanded to a countably infinite cover by setting $C_n = \emptyset$ for $n > m$ and using $\iota(\emptyset) = 0$.

XVII.4.1.6. Some authors do not include the requirement that $\emptyset \in \mathcal{C}$ and/or that $\iota(\emptyset) = 0$, obtaining that $\mu^*(\emptyset) = 0$ since $\emptyset$ is covered by the "empty cover" and hence $\mu^*(\emptyset)$ is less than or equal to the "empty sum," which is by definition 0. But there is no advantage in this approach ($\emptyset$ can be added to $\mathcal{C}$ if necessary and $\iota(\emptyset)$ can be defined or redefined to be 0 with no penalty) and it causes technical problems, e.g. when one wants to use finite covers as well as countably infinite ones. The same convention also justifies the definition of $\mu^*(A)$ if A does not have a countable cover from $\mathcal{C}$: $\mu^*(A)$ is an infimum over $\emptyset$, which is by definition $+\infty$.

XVII.4.1.7. PROPOSITION. This function μ^* is an outer measure on X.

PROOF: Since $\emptyset \in \mathcal{C}$, $\{\emptyset\}$ is a cover of $\emptyset$, so $\mu^*(\emptyset) = 0$. If $A \subseteq B$, any countable cover of B is a countable cover of A, so the infimum in calculating $\mu^*(A)$ is over a larger collection of covers, and hence $\mu^*(A) \leq \mu^*(B)$. Finally, suppose $A = \cup_{n=1}^{\infty} A_n$. If $\sum_{n=1}^{\infty} \mu^*(A_n) = \infty$, then clearly $\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$. In the case that $\sum_{n=1}^{\infty} \mu^*(A_n) < \infty$, so $\mu^*(A_n) < \infty$ for all n, let $\epsilon > 0$, and for each n let $\{C_{nk}\}$ be a cover of A_n with

$$\sum_{k=1}^{\infty} \iota(C_{nk}) \leq \mu^*(A_n) + 2^{-n}\epsilon \ .$$

Then $\{C_{nk} : n, k \in \mathbb{N}\}$ is a cover of A, so

$$\mu^*(A) \leq \sum_{n,k} \iota(C_{nk}) \leq \sum_{n=1}^{\infty} \mu^*(A_n) + \epsilon \ .$$

Since ϵ is arbitrary, $\mu^*(A) \leq \sum_{n=1}^{\infty} \mu^*(A_n)$. 👍

XVII.4.1.8. DEFINITION. The outer measure μ^* is the outer measure *induced* by ι.

This construction of an outer measure from a set function is often called *Method I*. See XVII.6.2.5. for "Method II," and XVII.4.5.16. for another method.

Note that $\mu^*(C) \leq \iota(C)$ for any $C \in \mathcal{C}$, since $\{C\}$ is a cover of C by $\mathcal{C}$; but $\mu^*(C)$ does not equal $\iota(C)$ in general:

XVII.4.1.9. EXAMPLE. Let $X = \mathbb{N}$, $\mathcal{C} = \mathcal{P}(\mathbb{N})$. Set $\iota(A) = 0$ if A is finite and $\iota(A) = 1$ if A is infinite. Then ι is monotone and (finitely) subadditive, but $\mu^*(A) = 0$ for all $A \subseteq \mathbb{N}$.

An outer measure induced by such a set function has an outer regularity property. If $\mathcal{C}$ is a collection of subsets of X containing $\emptyset$, write $\mathcal{C}_\sigma$ for the set of countable unions of sets in $\mathcal{C}$, and $\mathcal{C}_{\sigma\delta}$ for the set of countable intersections of sets in $\mathcal{C}_\sigma$. $\mathcal{C}_\sigma$ and $\mathcal{C}_{\sigma\delta}$ are contained in the σ-algebra generated by $\mathcal{C}$.

XVII.4.1.10. PROPOSITION. Let $\mathcal{C}$ be a set of subsets of X with $\emptyset \in \mathcal{C}$, $\iota : \mathcal{C} \to [0, \infty]$ with $\iota(\emptyset) = 0$, and μ^* the induced outer measure on X. Let $E \subseteq X$ with $\mu^*(E) < \infty$. Then

(i) For every $\epsilon > 0$, there is an $A \in \mathcal{C}_\sigma$ with $E \subseteq A$ and $\mu^*(A) \leq \mu^*(E) + \epsilon$.

(ii) There is a $B \in \mathcal{C}_{\sigma\delta}$ with $E \subseteq B$ and $\mu^*(E) = \mu^*(B)$.

PROOF: (i) Let $\epsilon > 0$. Take a sequence (C_n) in $\mathcal{C}$ with $E \subseteq \cup_{n=1}^\infty C_n$ and $\sum_{n=1}^\infty \iota(C_n) \leq \mu^*(E) + \epsilon$. If $A = \cup_{n=1}^\infty C_n$, then $A \in \mathcal{C}_\sigma$, $E \subseteq A$, and

$$\mu^*(A) \leq \sum_{n=1}^\infty \mu^*(C_n) \leq \sum_{n=1}^\infty \iota(C_n) \leq \mu^*(E) + \epsilon \ .$$

(ii): For each n let $A_n \in \mathcal{C}_\sigma$ with $E \subset A_n$ and $\mu^*(A_n) \leq \mu^*(E) + \frac{1}{n}$. Take $B = \cap_{n=1}^\infty A_n$. Then $B \in \mathcal{C}_{\sigma\delta}$, $E \subseteq B$, and

$$\mu^*(E) \leq \mu^*(B) \leq \mu^*(A_n) \leq \mu^*(E) + \frac{1}{n}$$

for every n, so $\mu^*(B) = \mu^*(E)$. 👍

XVII.4.2. Lebesgue Outer Measure on $\mathbb{R}$

By far the most important example of the process of inducing an outer measure from a function on subsets is Lebesgue outer measure, used to construct Lebesgue measure.

XVII.4.2.1. DEFINITION. Let $\mathcal{C}$ be the set of all open intervals in $\mathbb{R}$ (including $\emptyset$). For an open interval I, let $\ell(I)$ be the length of I. The induced outer measure from ℓ is *Lebesgue outer measure* λ^*. Specifically: for $A \subseteq \mathbb{R}$,

$$\lambda^*(A) = \inf \left\{ \sum_{n=1}^\infty \ell(I_n) \ : \ I_n \in \mathcal{C}, A \subseteq \bigcup_{n=1}^\infty I_n \right\} \ .$$

The use of open intervals is a matter of taste. It is useful to know that the same outer measure is obtained if arbitrary intervals are used:

XVII.4.2.2. PROPOSITION. Let $\mathcal{I}$ be the set of all intervals in $\mathbb{R}$, including $\emptyset$ and singleton subsets. For $I \in \mathcal{I}$, let $\ell(I)$ be the length of I ($\ell(\{a\}) = 0$ for a singleton $\{a\}$). Then the outer measure on $\mathbb{R}$ induced by ℓ on $\mathcal{I}$ is Lebesgue outer measure.

PROOF: Let μ^* be the outer measure on $\mathbb{R}$ induced by ℓ on $\mathcal{I}$. Since $\mathcal{C} \subseteq \mathcal{I}$, any $\mathcal{C}$-cover of a set A is also an $\mathcal{I}$-cover, so the infimum in computing $\mu^*(A)$ is over more covers than for computing $\lambda^*(A)$ and $\mu^*(A) \leq \lambda^*(A)$. To show the opposite inequality, we may assume $\mu^*(A) < \infty$. Let $\epsilon > 0$, and let $\{I_n\}$ be an $\mathcal{I}$-cover of A with $\sum_{n=1}^{\infty} \ell(I_n) \leq \mu^*(A) + \epsilon/2$. Each I_n is an interval (possibly $\emptyset$ or a singleton); for each n, let I'_n be an open interval containing I_n of length $\leq \ell(I_n) + 2^{-n-1}\epsilon$. Then $\{I'_n\}$ is a cover of A by open intervals, so we have

$$\lambda^*(A) \leq \sum_{n=1}^{\infty} \ell(I'_n) \leq \sum_{n=1}^{\infty} \ell(I_n) + \frac{\epsilon}{2} \leq \mu^*(A) + \epsilon \ .$$

Since ϵ is arbitrary, $\lambda^*(A) \leq \mu^*(A)$. 👍

The next corollary can be proved directly in the same way, or by just applying XVII.4.2.2..

XVII.4.2.3. COROLLARY. Let $\mathcal{H}$ be the set of intervals in $\mathbb{R}$ which are open on the left and closed on the right (i.e. of the form $(a, b]$, $(-\infty, b]$, or $(a, +\infty)$), along with the empty set. For $I \in \mathcal{H}$, let $\ell(I)$ be the length of I. Then the outer measure on $\mathbb{R}$ induced by ℓ on $\mathcal{H}$ is Lebesgue outer measure.

PROOF: Let ν^* be the outer measure induced by ℓ on $\mathcal{H}$. Since $\mathcal{H} \subseteq \mathcal{I}$, we have $\mu^*(A) \leq \nu^*(A)$ for any $A \subseteq \mathbb{R}$. If $A \subseteq \mathbb{R}$ and $\{I_n\}$ is a $\mathcal{C}$-cover of A, let I'_n be I_n with right endpoint added (if there is one); then $\{I'_n\}$ is an $\mathcal{H}$-cover of A, and $\sum_{n=1}^{\infty} \ell(I'_n) = \sum_{n=1}^{\infty} \ell(I_n)$; thus $\nu^*(A) \leq \lambda^*(A)$. 👍

It is easily shown that the class $\mathcal{C}_f$ of bounded open intervals also induces Lebesgue outer measure, as do the classes $\mathcal{I}_f$ and $\mathcal{H}_f$; for any $\delta > 0$, the classes $\mathcal{C}_\delta$ of intervals of length $\leq \delta$ and $\mathcal{I}_\delta$, $\mathcal{H}_\delta$ also induce Lebesgue outer measure on $\mathbb{R}$ (Exercise ().)

The next result is crucial, and shows that Lebesgue outer measure has the expected property of extending the interval length function. (In fact, without this theorem it is not even obvious that Lebesgue outer measure is not identically zero.) This result may seem "obvious" but is not, and the proof is somewhat delicate. Some references omit proving this entirely; the proof in [Fol99] (and in some other references) is described as "fussy." We give a slick proof from [GP10], attributed to VON NEUMANN.

XVII.4.2.4. THEOREM. If I is an interval in $\mathbb{R}$, then $\lambda^*(I) = \ell(I)$.

PROOF: We have $\lambda^*(I) \leq \ell(I)$, since $\{I\}$ is a cover of I. To show the reverse inequality, first suppose $I = [a, b]$. Let $\epsilon > 0$, and let (I_k) be a sequence of open intervals which cover $[a, b]$ with $\sum_{k=1}^{\infty} \ell(I_k) \leq \lambda^*(I) + \epsilon$ (so the I_k are necessarily bounded). Since $[a, b]$ is compact, $\{I_1, \ldots, I_n\}$ covers I for some n. Fix such an n.

If J is a bounded interval in $\mathbb{R}$, write $\#(J)$ for the number of integers in J. We have

$$\ell(J) - 1 \leq \#(J) \leq \ell(J) + 1 \ .$$

If $\alpha > 0$, then $\{\alpha I_1, \ldots, \alpha I_n\}$ covers αI, so we have

$$\alpha\ell(I) - 1 = \ell(\alpha I) - 1 \leq \#(\alpha I) \leq \sum_{k=1}^{n} \#(\alpha I_k) \leq \sum_{k=1}^{n} [\ell(\alpha I_k) + 1] = \alpha \left[\sum_{k=1}^{n} \ell(I_k) \right] + n \ .$$

Dividing by α, we obtain

$$\ell(I) - \frac{1}{\alpha} \leq \left[\sum_{k=1}^{n} \ell(I_k) \right] + \frac{n}{\alpha} \ .$$

This is true for all $\alpha > 0$, so letting $\alpha \to +\infty$ we obtain

$$\ell(I) \leq \sum_{k=1}^{n} \ell(I_k) \leq \sum_{k=1}^{\infty} \ell(I_k) < \lambda^*(I) + \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $\ell(I) \leq \lambda^*(I)$.

Now suppose $I = (a, b)$. By subadditivity,

$$\ell(I) = b - a = \lambda^*([a, b]) \leq \lambda^*(\{a\}) + \lambda^*(I) + \lambda^*(\{b\}) = 0 + \lambda^*(I) + 0 \ .$$

If $I = (a, b]$ or $I = [a, b)$, then $(a, b) \subseteq I \subseteq [a, b]$, so by monotonicity $\lambda^*(I) = \ell(I)$.

If I has infinite length, then either $(a, \infty) \subseteq I$ or $(-\infty, a) \subseteq I$ for some a. If $(a, \infty) \subseteq I$, then $(a, a+n) \subseteq I$ for all n, so by monotonicity

$$n = \lambda^*((a, a + n)) \leq \lambda^*(I)$$

for all n, and $\lambda^*(I) = \infty = \ell(I)$. The argument if $(-\infty, a) \subseteq I$ is similar. 👍

XVII.4.3. The Measure Induced by an Outer Measure

XVII.4.3.1. Definition. Let μ^* be an outer measure on a set X. A subset A of X is *μ^*-measurable* if, for every $Y \subseteq X$,

$$\mu^*(Y) = \mu^*(Y \cap A) + \mu^*(Y \cap A^c) = \mu^*(Y \cap A) + \mu^*(Y \setminus A) \ .$$

Intuitively in the case where μ^* comes from a set function by Method I, A splits any set Y into two pieces $Y \cap A$ and $Y \cap A^c$ which have "almost disjoint" covers.

Note that the inequality $\mu^*(Y) \leq \mu^*(Y \cap A) + \mu^*(Y \cap A^c)$ holds for any Y and A by subadditivity, so only the reverse inequality is needed for measurability of A. This reverse inequality is automatic if $\mu^*(Y) = \infty$, so we obtain:

XVII.4.3.2. Proposition. Let μ^* be an outer measure on a set X, and $A \subseteq X$.

(i) A is μ^*-measurable if and only if

$$\mu^*(Y) \geq \mu^*(Y \cap A) + \mu^*(Y \setminus A)$$

for every $Y \subseteq X$ with $\mu^*(Y) < \infty$.

(ii) A is μ^*-measurable if and only if, for every $Y \subseteq A$ and $Z \subseteq X \setminus A$ with $\mu^*(Y)$, $\mu^*(Z)$ finite,

$$\mu^*(Y \cup Z) \geq \mu^*(Y) + \mu^*(Z) \ .$$

XVII.4.3.3. COROLLARY. Let μ^* be an outer measure on a set X. If $A \subseteq X$ and $\mu^*(A) = 0$, then A is μ^*-measurable.

PROOF: If $Y \subseteq X$, then by monotonicity we have $\mu^*(Y \cap A) = 0$, and

$$\mu^*(Y) \geq \mu^*(Y \cap A^c) = \mu^*(Y \cap A) + \mu^*(Y \cap A^c) \ .$$

The name "measurable" for such sets is explained in the introduction, and justified by the following theorem, which is the main result of this section and one of the fundamental theorems of measure theory:

XVII.4.3.4. THEOREM. [CARATHÉODORY] Let μ^* be an outer measure on a set X, and $\mathcal{M}$ the set of μ^*-measurable subsets of X. Then

(i) $\mathcal{M}$ is a σ-algebra of subsets of X.

(ii) The restriction μ of μ^* to $\mathcal{M}$ is a measure (i.e. countably additive).

(iii) The measure space $(X, \mathcal{M}, \mu)$ is complete.

PROOF: It is obvious from the definition of measurability that $\emptyset, X \in \mathcal{M}$ and that $A \in \mathcal{M}$ implies $A^c \in \mathcal{M}$. Let $A, B \in \mathcal{M}$. If $Y \subseteq X$, then

$$Y \cap (A \cup B) = [Y \cap (A \cap B)] \cup [Y \cap (A \cap B^c)] \cup [Y \cap (A^c \cap B)]$$

and so, by subadditivity,

$$\mu^*(Y \cap (A \cup B)) \leq \mu^*(Y \cap (A \cap B)) + \mu^*(Y \cap (A \cap B^c)) + \mu^*(Y \cap (A^c \cap B)) \ .$$

Thus we have

$$\begin{aligned} \mu^*(Y) &= \mu^*(Y \cap A) + \mu^*(Y \cap A^c) \\ &= \mu^*(Y \cap A \cap B) + \mu^*(Y \cap A \cap B^c) + \mu^*(Y \cap A^c \cap B) + \mu^*(Y \cap A^c \cap B^c) \\ &\geq \mu^*(Y \cap (A \cup B)) + \mu^*(Y \cap A^c \cap B^c) = \mu^*(Y \cap (A \cup B)) + \mu^*(Y \cap (A \cup B)^c) \end{aligned}$$

so $A \cup B$ is μ^*-measurable. Therefore $\mathcal{M}$ is an algebra.

Next suppose $A, B \in \mathcal{M}$ are disjoint. We have

$$\mu^*(A \cup B) = \mu^*((A \cup B) \cap A) + \mu^*((A \cup B) \cap A^c) = \mu^*(A) + \mu^*(B)$$

so μ^* is finitely additive on $\mathcal{M}$.

Now suppose (A_n) is a pairwise disjoint sequence in $\mathcal{M}$, and set $B_n = \cup_{k=1}^n A_k$. Then $B_n \in \mathcal{M}$ for all n. Set $B = \cup_{n=1}^\infty B_n = \cup_{n=1}^\infty A_n$. If $Y \subseteq X$, we have

$$\mu^*(Y) = \mu^*(Y \cap B_n) + \mu^*(Y \cap B_n^c) = \mu^*(Y \cap B_n \cap A_n) + \mu^*(Y \cap B_n \cap A_n^c) + \mu^*(Y \cap B_n^c)$$

$$= \mu^*(Y \cap A_n) + \mu^*(Y \cap B_{n-1}) + \mu^*(Y \cap B_n^c) = \cdots = \sum_{k=1}^n \mu^*(Y \cap A_k) + \mu^*(Y \cap B_n^c) \geq \sum_{k=1}^n \mu^*(Y \cap A_k) + \mu^*(Y \cap B^c)$$

for every n, and hence

$$\mu^*(Y) \geq \sum_{k=1}^{\infty} \mu^*(Y \cap A_k) + \mu^*(Y \cap B^c) \geq \mu^*(Y \cap B) + \mu^*(Y \cap B^c) \geq \mu^*(Y)$$

and thus all the inequalities are equalities. So $B \in \mathcal{M}$ and $\mathcal{M}$ is a σ-algebra (). Applying this last result to $Y = B$, we obtain $\mu^*(B) = \sum_{k=1}^{\infty} \mu^*(A_k)$, so μ^* is countably additive on $\mathcal{M}$.

The fact that $(X, \mathcal{M}, \mu)$ is complete follows immediately from XVII.4.3.3.. 👍

XVII.4.3.5. The set $\mathcal{M}$ of μ^*-measurable subsets is more than just a σ-algebra; for example, unlike a general σ-algebra, it is also closed under the Suslin operation (XVII.4.6.7.).

XVII.4.3.6. The definition of measurable set we have used is due to Carathéodory, and is now standard. Lebesgue's definition of a measurable set (within a bounded interval I of $\mathbb{R}$ with Lebesgue outer measure) was different and apparently less restrictive: $A \subseteq I$ is measurable in Lebesgue's sense if $\lambda^*(A) + \lambda^*(I \setminus A) = \lambda^*(I) = \ell(I)$. The two notions are the same in good cases (XVII.4.5.38.), but not in general. Carathéodory's definition is more generally useful, and even for finite measure spaces gives simpler proofs.

Restrictions and Extensions

XVII.4.3.7. If μ^* is an outer measure on a set X, and $Y \subseteq X$, then the restriction μ^*_Y of μ^* to $\mathcal{P}(Y)$ is an outer measure on Y. How do the μ^*_Y-measurable subsets of Y relate to the subsets of Y which are μ^*-measurable? If Y is not μ^*-measurable, there is clearly a big difference (e.g. Y is μ^*_Y-measurable). But if Y is μ^*-measurable, there is no difference:

XVII.4.3.8. Proposition. Let μ^* be an outer measure on a set X, and Y a μ^*-measurable subset of X. Let μ^*_Y be the restriction of μ^* to $\mathcal{P}(Y)$. Then a subset A of Y is μ^*_Y-measurable if and only if it is μ^*-measurable.

Proof: If $A \subseteq Y$ is μ^*-measurable, then for any $Z \subseteq Y$ we have

$$\mu^*_Y(Z) = \mu^*(Z) = \mu^*(Z \cap A) + \mu^*(Z \cap A^c) = \mu^*(Z \cap A) + \mu^*(Z \cap (Y \setminus A)) = \mu^*_Y(Z \cap A) + \mu^*_Y(Z \cap (Y \setminus A))$$

so A is μ^*_Y-measurable. (This direction does not use that Y is μ^*-measurable.) Conversely, suppose A is μ^*_Y-measurable. Then, for any $Z \subseteq Y$,

$$\mu^*(Z) = \mu^*_Y(Z) = \mu^*_Y(Z \cap A) + \mu^*_Y(Z \cap (Y \setminus A)) = \mu^*(Z \cap A) + \mu^*(Z \cap (Y \cap A^c))$$

So if $W \subset X$, we have

$$\mu^*(W) = \mu^*(W \cap Y) + \mu^*(W \cap Y^c) = \mu^*((W \cap Y) \cap A) + \mu^*((W \cap Y) \cap (Y \cap A^c)) + \mu^*(W \cap Y^c)$$

$$= \mu^*(W \cap A) + \mu^*((W \cap Y) \cap A^c) + \mu^*(W \cap Y^c) \geq \mu^*(W \cap A) + \mu^*(W \cap A^c)$$

by subadditivity, so A is μ^*-measurable. 👍

XVII.4.4. Lebesgue Measure on $\mathbb{R}$

[**Prerequisite:** ()]

XVII.4.4.1. DEFINITION. Let $\mathcal{L}$ be the σ-algebra of λ^*-measurable subsets of $\mathbb{R}$. $\mathcal{L}$ is called the set of *Lebesgue measurable sets* in $\mathbb{R}$. Write λ for the restriction of λ^* to $\mathcal{L}$. Then λ is a measure on $(\mathbb{R}, \mathcal{L})$, called *Lebesgue measure* on $\mathbb{R}$.

XVII.4.4.2. PROPOSITION. $(\mathbb{R}, \mathcal{L}, \lambda)$ is a complete measure space.

This is an immediate consequence of ().

XVII.4.4.3. PROPOSITION. Every Borel set is λ^*-measurable, i.e. $\mathcal{B} \subseteq \mathcal{L}$.

PROOF: This follows immediately from (). However, there is a quicker argument. It suffices to show that (a, ∞) is measurable for any $a \in \mathbb{R}$ (actually just for a in a dense subset of $\mathbb{R}$), since such sets generate $\mathcal{B}$. If $Y \subseteq \mathbb{R}$ with $\lambda^*(Y) < \infty$, we need to show that

$$\lambda^*(Y) \geq \lambda^*(Y \cap (a, \infty)) + \lambda^*(Y \cap (-\infty, a]) \ .$$

Let $\epsilon > 0$, and let (I_n) be a sequence of intervals covering Y with $\sum_{n=1}^{\infty} \ell(I_n) \leq \lambda^*(Y) + \epsilon$. For each n, let $I_n' = I_n \cap (a, \infty)$ and $I_n'' = I_n \cap (-\infty, a]$. Then the I_n' and I_n'' are intervals (some may be empty), $\ell(I_n') + \ell(I_n'') = \ell(I_n)$, (I_n') is a cover of $Y \cap (a, \infty)$, and (I_n'') is a cover of $Y \cap (-\infty, a])$, so

$$\lambda^*(Y \cap (a, \infty)) + \lambda^*(Y \cap (-\infty, a]) \leq \sum_{n=1}^{\infty} \ell(I_n') + \sum_{n=1}^{\infty} \ell(I_n'') = \sum_{n=1}^{\infty} \ell(I_n) \leq \lambda^*(Y) + \epsilon \ .$$

Since ϵ is arbitrary, we conclude the desired inequality. 👍

XVII.4.4.4. COROLLARY. $(\mathbb{R}, \mathcal{L}, \lambda)$ is a σ-finite measure space.

PROOF: For every n, $[-n, n]$ is Lebesgue measurable and $\lambda([-n, n]) = 2n < \infty$, and $\mathbb{R} = \cup_n [-n, n]$. 👍

It is not true that every Lebesgue measurable set is a Borel set. In fact, the cardinality of $\mathcal{B}$ is $2^{\aleph_0}$ (); but every subset of the Cantor set K is Lebesgue measurable (), and $card(K) = 2^{\aleph_0}$, so it follows that $card(\mathcal{L}) = 2^{2^{\aleph_0}}$. For an explicit example of a Lebesgue measurable set which is not a Borel set, see (). However:

XVII.4.4.5. THEOREM. Let $A \in \mathcal{L}$. Then

(i) For any $\epsilon > 0$, there is an open set U containing A with $\lambda(U \setminus A) < \epsilon$.

(ii) For any $\epsilon > 0$, there is a closed set F contained in A with $\lambda(A \setminus F) < \epsilon$. If $\lambda(A) < \infty$, then F may be chosen to be compact.

(iii) There is a G_δ set D containing A with $\lambda(D \setminus A) = 0$.

(iv) There is an F_σ set E contained in A with $\lambda(A \setminus E) = 0$.

In particular, every Lebesgue measurable set differs from a Borel set by a set of measure 0.

PROOF: (i) For each $n \in \mathbb{Z}$, let $A_n = A \cap (n-1, n]$. Then $A_n \in \mathcal{L}$, the A_n are disjoint, $A = \cup_n A_n$, and $\lambda(A_n) \leq 1$. For each n there is a sequence (I_{nm}) of open intervals covering A_n with

$$\sum_{m=1}^{\infty} \ell(I_{nm}) < \lambda(A_n) + 2^{-(|n|+2)}\epsilon \ .$$

If $U_n = \cup_{m=1}^{\infty} I_{nm}$, then U_n is open, $A_n \subseteq U_n$, and

$$\lambda(A_n) \leq \lambda(U_n) \leq \sum_{m=1}^{\infty} \ell(I_{nm}) < \lambda(A_n) + 2^{-(|n|+2)}\epsilon$$

and hence $\lambda(U_n \setminus A_n) < 2^{-(|n|+2)}\epsilon$ since $\lambda(A_n) < \infty$.

Now let $U = \cup_{n\in\mathbb{Z}} U_n$. Then U is open, $A \subseteq U$, and $U \setminus A \subseteq \cup_n (U_n \setminus A_n)$, so

$$\lambda(U \setminus A) \leq \sum_{n\in\mathbb{Z}} \lambda(U_n \setminus A_n) < \sum_{n\in\mathbb{Z}} 2^{-(|n|+2)}\epsilon < \epsilon \ .$$

(ii): Let U be an open set containing A^c such that $\lambda(U \setminus A^c) < \epsilon/2$, and set $E = U^c$. Then E is closed, $E \subseteq A$, and $A \setminus E = U \setminus A^c$, so $\lambda(A \setminus E) < \epsilon/2$. If $\lambda(A) = \infty$, set $F = E$. Suppose $\lambda(A) < \infty$. For $n \in \mathbb{N}$, set $F_n = E \cap [-n, n]$. Then F_n is compact, and $E = \cup_n F_n$, so $A \setminus E = \cap_n A \setminus F_n$. This is a decreasing intersection of sets of finite measure, so $\lambda(A \setminus E) = \inf_n\{\lambda(A \setminus F_n)\} < \epsilon/2$ (). There is thus an n so that $\lambda(A \setminus F_n) < \epsilon$. Set $F = F_n$ for this n.
(iii): For each n let U_n be an open set containing A with $\lambda(U_n \setminus A) < 1/n$, and set $D = \cap_n U_n$.
(iv): For each n let F_n be a closed set contained in A with $\lambda(A \setminus F_n) < 1/n$, and set $E = \cup_n F_n$. 👍

XVII.4.4.6. Thus every Lebesgue measurable set A is "trapped between" a G_δ set D and an F_σ set E (i.e. $E \subseteq A \subseteq D$) with $\lambda(D \setminus E) = 0$, and conversely by completeness any set which is trapped between such D and E is Lebesgue measurable. (Note, however, that for a given A there is nothing unique about D and E.)

XVII.4.5. Extending Premeasures from Semirings

XVII.4.5.1. We will now show that under some reasonable restrictions, a set function $\iota : \mathcal{C} \to [0, \infty]$ extends to a measure μ on a σ-algebra $\mathcal{A}$ containing $\mathcal{C}$. To motivate the restrictions on ι, suppose an extension μ exists. Then a few moment's thought shows that the following must be true:

(i) ι must be monotone on $\mathcal{C}$.

(ii) If $C_1, \ldots, C_n$ are disjoint sets in $\mathcal{C}$, and $C := \cup_{k=1}^n C_k \in \mathcal{C}$, then $\iota(C) = \sum_{k=1}^n \iota(C_k)$.

(ii′) If $C_1, C_2, \ldots$ are disjoint sets in $\mathcal{C}$, and $C := \cup_{k=1}^\infty C_k \in \mathcal{C}$, then $\iota(C) = \sum_{k=1}^\infty \iota(C_k)$.

(iii) If $C, C_1, C_2, \cdots \in \mathcal{C}$ and $C \subseteq \cup_{k=1}^{\infty} C_k$, then $\iota(C) \leq \sum_{k=1}^{\infty} \iota(C_k)$.

(Note that (ii′) does not necessarily imply (ii) unless $\emptyset \in \mathcal{C}$ and $\iota(\emptyset) = 0$.)

It is also reasonable to require that $\mathcal{C}$ be closed under enough set operations to make these restrictions on ι have some teeth. It turns out to be sufficient that $\mathcal{C}$ be a semiring; in fact, (i) will follow automatically from (ii) and (iii) in this case, and in the presence of (i), (ii′) is equivalent to (ii) and (iii) combined.

XVII.4.5.2. DEFINITION. Let $\mathcal{S}$ be a semiring of subsets of a set X. A *premeasure* on $\mathcal{S}$ is a set function $\iota : \mathcal{S} \to [0, \infty]$ with the properties that

(i) $\iota(\emptyset) = 0$.

(ii) If $C_1, \ldots, C_n$ are pairwise disjoint sets in $\mathcal{S}$, and $C := \cup_{k=1}^{n} C_k \in \mathcal{S}$, then $\iota(C) = \sum_{k=1}^{n} \iota(C_k)$ (*finite additivity*).

(iii) If $C, C_1, C_2, \cdots \in \mathcal{S}$ and $C \subseteq \cup_{k=1}^{\infty} C_k$, then $\iota(C) \leq \sum_{k=1}^{\infty} \iota(C_k)$ (*countable subadditivity*).

The premeasure ι is *σ-finite* if there are $C_1, C_2, \cdots \in \mathcal{S}$ with $\iota(C_n) < \infty$ for all n and $\cup_{n=1}^{\infty} C_n = X$. If ι is σ-finite, then $\mathcal{S}$ is necessarily of countable type.

XVII.4.5.3. EXAMPLES. (i) Let $(X, \mathcal{A}, \mu)$ be any measure space, and $\mathcal{S}$ a semiring contained in $\mathcal{A}$. Then $\mu|_{\mathcal{S}}$ is a premeasure on $\mathcal{S}$.
(ii) Let $\mathcal{S}$ be any of the following semirings on $\mathbb{R}$:

All intervals

All bounded intervals

For fixed $\delta > 0$, all intervals of length $\leq \delta$

(in all cases, include the degenerate intervals $\emptyset$ and singletons). Then ℓ is a σ-finite premeasure on $\mathcal{S}$ (cf. XVII.4.2.4., Exercise ()).

XVII.4.5.4. **Note:** There is substantial nonuniformity of use of the term "premeasure," and no consensus on the proper definition of the term. Some authors use it much more broadly than we have, and others more narrowly. I believe our definition of "premeasure" is the most useful one, since it is directly used not only in defining Lebesgue measure and Radon measures (finite Borel measures on $\mathbb{R}$ defined via cumulative distribution functions), but also naturally in defining product measures.

Measure extensions can be defined by more primitive data than a premeasure on a semiring; see [Kön97] for a study.

XVII.4.5.5. PROPOSITION. A premeasure on a semiring is monotone.

PROOF: Let ι be a premeasure on a semiring $\mathcal{S}$, and $C, D \in \mathcal{S}$, $C \subseteq D$. Then $D \setminus C$ is a disjoint union of $C_1, \ldots, C_n \in \mathcal{S}$, and

$$\iota(D) = \iota(C) + \sum_{k=1}^{n} \iota(C_k) \geq \iota(C)$$

since $\iota(C_k) \geq 0$ for all k. 👍

Premeasures can be characterized alternatively by countable additivity, which is sometimes easier to check:

XVII.4.5.6. PROPOSITION. Let $\mathcal{S}$ be a semiring of subsets of a set X, and $\iota : \mathcal{S} \to [0, \infty]$. Then ι is a premeasure if and only if ι satisfies

(i) $\iota(\emptyset) = 0$.

(ii′) If $C_1, C_2, \ldots$ are pairwise disjoint sets in $\mathcal{S}$, and $C := \cup_{k=1}^{\infty} C_k \in \mathcal{S}$, then $\iota(C) = \sum_{k=1}^{\infty} \iota(C_k)$ (*countable additivity*).

Note that (ii′) (countable additivity) does not trivially imply (iii) (countable subadditivity) on semirings, even on algebras, since countable additivity only applies to sequences whose union is in the semiring. The proposition is a consequence of XVI.1.2.12. and XVI.1.2.13., which insure that "sufficiently many" sets in a semiring are finite or countable disjoint unions of sets in the semiring.
PROOF: Suppose ι satisfies (i) and (ii′). Then ι obviously satisfies (ii). Let $C, C_1, C_2, \cdots \in \mathcal{S}$ with $C \subseteq \cup_n C_n$. Then by XVI.1.2.13. there are pairwise disjoint $B_k \in \mathcal{S}$, each contained in a unique C_n, only finitely many in each C_n, such that $C = \cup_k B_k$. For each n, let D_n be the union of the B_k which are in C_n. Then $C_n \setminus D_n$ is a finite disjoint union $A_{n1}, \ldots, A_{nm_n}$ of sets in $\mathcal{S}$ by XVI.1.2.12.; hence we have

$$\iota(C_n) = \sum_{B_k \subseteq C_n} \iota(B_k) + \sum_{j=1}^{m_n} \iota(A_{nj}) \geq \sum_{B_k \subseteq C_n} \iota(B_k)$$

by (ii). Therefore we have by (ii′)

$$\iota(C) = \sum_{k=1}^{\infty} \iota(B_k) = \sum_{n=1}^{\infty} \sum_{B_k \subseteq C_n} \iota(B_k) \leq \sum_{n=1}^{\infty} \iota(C_n)$$

and thus ι satisfies (iii).

Conversely, if ι satisfies (i), (ii), and (iii), let $C_1, C_2, \ldots$ be pairwise disjoint sets in $\mathcal{S}$ with $C = \cup_{n=1}^{\infty} C_n \in \mathcal{S}$. We have $\iota(C) \leq \sum_{n=1}^{\infty} \iota(C_n)$ by (iii). On the other hand, if $D_n = \cup_{k=1}^{n} C_k$, we have that $C \setminus D_n$ is a finite disjoint union of sets $A_{n1}, \ldots, A_{nm_n}$ in $\mathcal{S}$, so

$$\iota(C) = \sum_{k=1}^{n} \iota(C_k) + \sum_{j=1}^{m_n} \iota(A_{nj}) \geq \sum_{k=1}^{n} \iota(C_k)$$

for any n, i.e. $\iota(C) = \sum_{k=1}^{\infty} \iota(C_k)$. 👍

Premeasures are technically simpler to work with if the domain is a ring instead of just a semiring. (In fact, some authors such as [Hal50] define a measure to be what we call a premeasure on a ring.) We do not make this restriction in the definition since many naturally occurring premeasures are only naturally defined on semirings. However, a premeasure on a semiring uniquely extends to a premeasure on the ring it generates, so there is no loss of generality for extension theory in considering just premeasures defined on rings (as some authors do):

XVII.4.5.7. THEOREM. Let ι be a premeasure on a semiring $\mathcal{S}$ of subsets of X, and let $\mathcal{R}$ be the ring of subsets of X generated by $\mathcal{S}$. Then ι extends uniquely to a premeasure on $\mathcal{R}$.

PROOF: Recall () that $\mathcal{R}$ consists of all finite disjoint unions of sets in $\mathcal{S}$. If ι extends to a premeasure κ on $\mathcal{R}$, and $R \in \mathcal{R}$ is a disjoint union of sets $C_1, \ldots, C_n \in \mathcal{S}$, we must have $\kappa(R) = \sum_{k=1}^{n} \iota(C_k)$. We first need to show that this formula gives a well-defined function on $\mathcal{R}$.

Suppose $R \in \mathcal{R}$ has two representations $R = \cup_{k=1}^{n} C_k = \cup_{j=1}^{m} D_j$ as disjoint finite unions of sets in $\mathcal{S}$. For each k and j set $E_{kj} = C_k \cap D_j$. Then each $E_{kj} \in \mathcal{S}$ since $\mathcal{S}$ is closed under finite intersections, and the E_{kj} are pairwise disjoint. Thus

$$\sum_{k=1}^{n} \iota(C_k) = \sum_{k=1}^{n} \sum_{j=1}^{m} \iota(E_{kj}) = \sum_{j=1}^{m} \sum_{k=1}^{n} \iota(E_{kj}) = \sum_{j=1}^{m} \iota(D_j)$$

and so κ is well defined.

To show that κ is a premeasure, we must show (ii) and (iii). For (ii), suppose $R_1, \ldots, R_n \in \mathcal{R}$ are pairwise disjoint and suppose $R := \cup_{k=1}^{n} R_k \in \mathcal{R}$. For each k, write R_k as a disjoint union $\cup_{j=1}^{m_k} C_{kj}$ with $C_{kj} \in \mathcal{S}$. Then

$$\kappa(R) = \sum_{k=1}^{n} \sum_{j=1}^{m_k} \iota(C_{kj}) = \sum_{k=1}^{n} \kappa(R_k) \ .$$

To show (iii), Suppose $R, R_1, R_2, \cdots \in \mathcal{R}$ and $R \subseteq \cup_k R_k$. Write $R_k = \cup_{j=1}^{m_k} C_{kj}$ and $R = \cup_{i=1}^{r} D_i$ as disjoint unions with $C_{kj}, D_i \in \mathcal{S}$. For each i, j, k, set $E_{ijk} = C_{kj} \cap D_i$. Then each $E_{ijk} \in \mathcal{S}$ and, for each i, $D_i \subseteq \cup_{j,k} E_{i,j,k}$, so

$$\iota(D_i) \leq \sum_{k,j} \iota(E_{ijk}) \ .$$

Then

$$\kappa(R) = \sum_{i=1}^{r} \iota(D_i) \leq \sum_{i,j,k} \iota(E_{ijk}) = \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} \sum_{i=1}^{r} \iota(E_{ijk}) \leq \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} \iota(C_{kj}) = \sum_{k=1}^{\infty} \kappa(R_k) \ .$$

👍

The model for constructing an extension of a premeasure to a measure is the construction of Lebesgue measure: begin with the premeasure ι, form the induced outer measure μ^* and the σ-algebra $\mathcal{M}$ of μ^*-measurable subsets; the restriction of μ^* to $\mathcal{M}$ is the desired extension. The precise statement is the next theorem, which is the main result of this section, and another of the most important facts in measure theory:

XVII.4.5.8. THEOREM. [EXTENSION THEOREM] Let $\mathcal{S}$ be a semiring on a set X, $\mathcal{A}$ the σ-algebra generated by $\mathcal{S}$, and ι a premeasure on $\mathcal{S}$. Let μ^* be the outer measure on X induced by ι, $\mathcal{M}$ the σ-algebra of μ^*-measurable subsets of X, and $\mu = \mu^*|_{\mathcal{M}}$. Then

(i) $\mathcal{S} \subseteq \mathcal{M}$, i.e. every set in $\mathcal{S}$ is μ^*-measurable.

(ii) The restriction of μ^* to $\mathcal{S}$ is ι, i.e. μ extends ι.

Thus ι extends to a complete measure μ on $\mathcal{M}$, and hence to a measure $\mu|_{\mathcal{A}}$ on $\mathcal{A}$. If $\mathcal{B}$ is a σ-algebra with $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{M}$, the extension $\mu|_{\mathcal{B}}$ is the largest extension of ι to a measure on $\mathcal{B}$, i.e. any extension ν satisfies $\nu \leq \mu$, and ν agrees with μ on sets of σ-finite μ-measure. If ι is σ-finite, then μ is the unique extension of ι to a measure on $\mathcal{B}$, and $(X, \mathcal{M}, \mu)$ is the completion of $(X, \mathcal{A}, \mu|_{\mathcal{A}})$ (hence also of $(X, \mathcal{B}, \mu|_{\mathcal{B}})$).

PROOF: We first prove (ii). Let $C \in \mathcal{S}$. Since $\{C\}$ is a cover of C, we have $\mu^*(C) \leq \iota(C)$. For the opposite inequality, we may assume $\mu^*(C) < \infty$. Let $\epsilon > 0$, and let (C_n) be a sequence of elements of $\mathcal{S}$ with $C \subseteq \cup_{n=1}^{\infty} C_n$ and $\sum_{n=1}^{\infty} \iota(C_n) \leq \mu^*(C) + \epsilon$. Then, by (iii),

$$\iota(C) \leq \sum_{n=1}^{\infty} \iota(C_n) \leq \mu^*(C) + \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $\iota(C) \leq \mu^*(C)$.

(i): Now let $C \in \mathcal{S}$, and let $A \subseteq X$ with $\mu^*(A) < \infty$. Fix $\epsilon > 0$, and let (C_n) be a sequence in $\mathcal{S}$ with $A \subseteq \cup_{n=1}^{\infty} C_n$ and $\sum_{n=1}^{\infty} \iota(C_n) \leq \mu^*(A) + \epsilon$. Then $A \cap C \subseteq \cup_{n=1}^{\infty}(C_n \cap C)$, and $C_n \cap C \in \mathcal{S}$, so

$$\mu^*(A \cap C) \leq \sum_{n=1}^{\infty} \mu^*(C_n \cap C) = \sum_{n=1}^{\infty} \iota(C_n \cap C) \ .$$

Also, $A \setminus C \subseteq \cup_{n=1}^{\infty}(C_n \setminus C)$. We have that $C_n \setminus C$ is a finite disjoint union $\cup_{k=1}^{m_n} D_{nk}$ for each n, with $D_{nk} \in \mathcal{S}$. So

$$\mu^*(A \setminus C) \leq \sum_{n=1}^{\infty} \sum_{k=1}^{m_n} \mu^*(D_{nk}) = \sum_{n=1}^{\infty} \sum_{k=1}^{m_n} \iota(D_{nk})$$

$$\mu^*(A \cap C) + \mu^*(A \setminus C) \leq \sum_{n=1}^{\infty} \iota(C_n \cap C) + \sum_{n=1}^{\infty} \sum_{k=1}^{m_n} \iota(D_{nk})$$

$$= \sum_{n=1}^{\infty} \left(\iota(C_n \cap C) + \sum_{k=1}^{m_n} \iota(D_{nk}) \right) \ .$$

We have that, for each n, C_n is the finite disjoint union of $C_n \cap C$ and the D_{nk}, so by (ii)

$$\iota(C_n) = \iota(C_n \cap C) + \sum_{k=1}^{m_n} \iota(D_{nk})$$

and thus

$$\mu^*(A \cap C) + \mu^*(A \setminus C) \leq \sum_{n=1}^{\infty} \iota(C_n) \leq \mu^*(A) + \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $\mu^*(A \cap C) + \mu^*(A \setminus C) \le \mu^*(A)$. Thus C is μ^*-measurable.

For the last part, suppose ν is a measure extending ι to a σ-algebra $\mathcal{B}$ with $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{M}$. If $B \in \mathcal{B}$, $\mu(B) < \infty$, let $\epsilon > 0$ and $B \subseteq \cup_{n=1}^{\infty} C_n$ with $C_n \in \mathcal{S}$ and $\sum_{n=1}^{\infty} \iota(C_n) \le \mu(B) + \epsilon$. Then

$$\nu(B) \le \sum_{n=1}^{\infty} \nu(C_n) = \sum_{n=1}^{\infty} \iota(C_n) \le \mu(B) + \epsilon \ .$$

Since ϵ is arbitrary, $\nu(B) \le \mu(B)$. On the other hand, let $A = \cup_{n=1}^{\infty} C_n \in \mathcal{A}$. Then $\mu(A) \le \sum_{n=1}^{\infty} \iota(C_n) < \infty$. Since $\mu(A \setminus B) < \infty$, we have $\nu(A \setminus B) \le \mu(A \setminus B)$. Let $D_1 = C_1$ and $D_n = C_n \setminus \cup_{k=1}^{n-1} C_k$. Then D_n is a finite disjoint union of sets in $\mathcal{S}$, so $\nu(D_n) = \mu(D_n)$; and A is the disjoint union of the D_n, so $\nu(A) = \mu(A)$. Then

$$\nu(B) = \nu(A) - \nu(A \setminus B) \ge \mu(A) - \mu(A \setminus B) = \mu(B) \ .$$

So $\nu(B) = \mu(B)$, and thus $\mu = \nu$ for sets in $\mathcal{B}$ of finite μ-measure. Then $\nu = \mu$ also for sets of σ-finite μ-measure, which can be written as countable disjoint unions of sets of finite μ-measure; so if ι is σ-finite, so is μ and $\nu = \mu$. In general, the only possible difference between μ and ν is that there could be a set B which is not of σ-finite μ-measure, where $\mu(B) = \infty$ but $\nu(B) < \infty$. Thus $\nu \le \mu$.

For the last statement, let (C_n) be a sequence in $\mathcal{S}$ with $\cup_{n=1}^{\infty} C_n = X$ and $\iota(C_n) < \infty$ for all n. Let $M \in \mathcal{M}$. Fix $n \in \mathbb{N}$; then for each $m \in \mathbb{N}$ there is a sequence (D_{nmk}) in $\mathcal{S}$ with

$$M \cap C_n \subseteq \bigcup_{k=1}^{\infty} D_{nmk}$$

and

$$\sum_{k=1}^{\infty} \iota(D_{nmk}) < \mu(M \cap C_n) + \frac{1}{m} \ .$$

Set

$$A_{nm} = \bigcup_{k=1}^{\infty} D_{nmk} \ .$$

Then $A_{nm} \in \mathcal{A}$, $M \cap C_n \subseteq A_{nm}$, and

$$\mu(A_{nm}) < \mu(M \cap C_n) + \frac{1}{m} \ .$$

Thus if

$$A_n = C_n \cap \left[\bigcap_{m=1}^{\infty} A_{nm} \right]$$

we have $A_n \in \mathcal{A}$, $M \cap C_n \subseteq A_n$, and $\mu(A_n) = \mu(M \cap C_n)$. Thus if $B_n = A_n \setminus (M \cap C_n)$, then $B_n \in \mathcal{M}$ and $\mu(B_n) = 0$. Set $A = \cup_{n=1}^{\infty} A_n$. Then $A \in \mathcal{A}$, and we have $M \subseteq A$ and

$$B = A \setminus M = \bigcup_{n=1}^{\infty} [A_n \setminus (M \cap C_n)] = \bigcup_{n=1}^{\infty} B_n$$

so $\mu(B) = 0$. Thus M is the union of a set in $\mathcal{A}$ and a set B of μ-measure 0. Repeating the above argument for $M \cap C_n$, there is an $E \in \mathcal{A}$ with $B \subseteq E$ and $\mu(E) = 0$. Thus M is the union of a set in $\mathcal{A}$ and a subset of a null set in $\mathcal{A}$. 👍

There is an extremely important corollary of the uniqueness part:

XVII.4.5.9. Corollary. Let $\mathcal{S}$ be a semialgebra on X, and $\mathcal{A}$ the σ-algebra generated by $\mathcal{S}$. If μ and ν are measures on $\mathcal{A}$ which agree on $\mathcal{S}$ and are σ-finite on $\mathcal{S}$, then $\mu = \nu$ on $\mathcal{A}$.

Of course, it is enough to check that one of μ, ν is σ-finite on $\mathcal{S}$. Note that it is *not* sufficient to check that $\mu = \nu$ on $\mathcal{S}$ and that μ, say, is σ-finite on $\mathcal{A}$, unless we know that μ is the measure on $\mathcal{A}$ induced by $\mu|_{\mathcal{S}}$: there is an $\mathcal{S}$ on $\mathbb{R}$ which generates the Borel sets as a σ-algebra, with the property that every nonempty set in $\mathcal{S}$ has infinite Lebesgue measure (Exercise ()); Lebesgue measure is σ-finite on $\mathcal{B}$ but not on $\mathcal{S}$, and agrees on $\mathcal{S}$ with counting measure and the measure μ which is infinite on all nonempty sets (μ is the measure on $\mathcal{B}$ induced by $\lambda|_{\mathcal{S}}$).

XVII.4.5.10. Theorems XVII.4.3.4. and XVII.4.5.8. are often combined into a single theorem, which is called the *Carathéodory Extension Theorem*, and the construction called the *Carathéodory extension process*, even though Carathéodory only proved XVII.4.3.4. (in addition to giving the definition of measurable set (XVII.4.3.1.)). Theorem XVII.4.5.8. was first proved by Fréchet []; the proof using XVII.4.3.4. is due independently to Hahn [] and Kolmogorov [].

XVII.4.5.11. Theorem. [Carathéodory Extension Theorem] If ι is a premeasure on a semiring $\mathcal{S}$ of subsets of a set X, then there is an extension of ι to a complete measure μ on a σ-algebra $\mathcal{M}$ containing $\mathcal{S}$. If ι is σ-finite on $\mathcal{S}$, then the extension μ is unique if $\mathcal{M}$ is taken to be the σ-algebra generated by $\mathcal{S}$ and the null sets.

It is interesting to see what this construction gives when applied to the case where $\mathcal{S}$ is a σ-algebra and ι is a measure on $\mathcal{S}$:

XVII.4.5.12. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and let μ^* be the outer measure induced by μ, $\mathcal{M}$ the σ-algebra of μ^*-measurable sets, and $\nu = \mu^*|_{\mathcal{M}}$. Then $(X, \mathcal{M}, \nu)$ is the saturation () of the completion (or the completion of the saturation) of $(X, \mathcal{A}, \mu)$. In particular, if μ is σ-finite, then $(X, \mathcal{M}, \nu)$ is the completion of $(X, \mathcal{A}, \mu)$.

XVII.4.5.13. There is an important extension of the uniqueness statement at the end of XVII.4.5.8. We first need a technical fact. Let $\mathcal{S}$ be a semiring of subsets of a set X, ι a premeasure on $\mathcal{S}$, and Y a union of countably many sets in $\mathcal{S}$. Let $\mathcal{S}_Y$ be the collection of all sets in $\mathcal{S}$ which are subsets of Y, and let ι_Y be the restriction of ι to Y.. Then $\mathcal{S}_Y$ is a semiring of subsets of Y and ι_Y is a premeasure on $\mathcal{S}_Y$. Let μ_Y^* and μ_Y° be the outer measures on Y defined by ι_Y.

XVII.4.5.14. Lemma. Using the above notation, μ_Y^* is equal to the restriction of μ^* to $\mathcal{P}(Y)$ (i.e. the notation is consistent with the notation of XVII.4.3.8.).

Proof:

XVII.4.5.15. PROPOSITION. Let $\mathcal{S}$ be a semiring of subsets of a set X and ι a premeasure on $\mathcal{S}$, and μ^* the outer measure on X induced by ι. Let $\mathcal{C}$ be the collection of subsets of X which can be covered by countably many sets in $\mathcal{S}$ of finite premeasure. Then any outer measure on X extending ι agrees with μ^* on sets in $\mathcal{C}$.

PROOF: If $C \in \mathcal{C}$, let Y be the union of a countable collection (S_n) of sets in $\mathcal{S}$ covering C with $\iota(S_n) = \mu^*(S_n) < \infty$ for all n. Then μ^*_Y is the restriction of μ^* to $\mathcal{P}(Y)$ by XVII.4.5.14.. The premeasure ι_Y is σ-finite, so there is a unique outer measure on Y extending ι_Y by XVII.4.5.8., namely μ^*_Y. Thus any outer measure ν^* on X, restricted to $\mathcal{P}(Y)$, agrees with μ^*_Y, and in particular $\nu^*(C) = \mu^*_Y(C) = \mu^*(C)$. 👍

Another Approach

XVII.4.5.16. In constructing an outer measure from a set function (e.g. a premeasure) ι on a collection $\mathcal{C}$ of subsets of a set X, we defined $\mu^*(A) = \infty$ if there is no countable cover of A by sets in $\mathcal{C}$. But we could have proceeded differently: we could have first defined μ^* in the usual way for sets with a countable cover by $\mathcal{C}$, and then for general A define

$$\mu^\circ(A) = \sup\{\mu^*(B) : B \subseteq A, B \text{ has a countable cover by } \mathcal{C}\} .$$

We then have $\mu^\circ(A) = \mu^*(A)$ if A has a countable cover by $\mathcal{C}$, and of course $\mu^\circ \leq \mu^*$ if μ^* is defined as in (). If $\mathcal{C}$ is of countable type, i.e. X is a countable union of sets in $\mathcal{C}$ (e.g. in the case of Lebesgue measure), then every subset of X has a countable cover by $\mathcal{C}$ and there is no difference between μ° and μ^* (so the following discussion is not relevant to most cases of greatest interest). In general, $\mu^\circ(A) = \mu^*(A)$ for any A with $\mu^*(A) < \infty$, and more generally for any countable union of such sets. The difference between μ° and μ^* is that there may be sets A with no countable cover by $\mathcal{C}$ for which $\mu^\circ(A) < \infty$ (for any such A, $\mu^*(A) = \infty$).

XVII.4.5.17. EXAMPLE. For a very simple example where $\mu^\circ \neq \mu^*$, let X be an uncountable set, and $\mathcal{C}$ the collection of countable subsets of X. Set $\iota(C) = 0$ for $C \in \mathcal{C}$. Then $\mu^\circ(A) = 0$ for all $A \subseteq X$, but $\mu^*(A) = \infty$ for A uncountable. Note that $\mathcal{C}$ is a σ-ring and ι is a premeasure (local measure) on $\mathcal{C}$. For an even simpler (trivial) example, let X be a nonempty set (say a one-element set) and $\mathcal{C} = \{\emptyset\}$. Then $\mu^\circ(X) = 0$ but $\mu^*(X) = \infty$.

It is slightly harder to prove that μ° is an outer measure (cf. XVII.4.1.7.):

XVII.4.5.18. PROPOSITION. This function μ° is an outer measure on X.

PROOF: Since $\emptyset \in \mathcal{C}$, $\{\emptyset\}$ is a cover of $\emptyset$, so $\mu^\circ(\emptyset) = \mu^*(\emptyset) = 0$. If $A_1 \subseteq A_2$, the supremum for calculating $\mu^\circ(A_2)$ is over a larger collection of sets than the supremum for calculating $\mu^\circ(A_1)$, so $\mu^\circ(A_1) \leq \mu^\circ(A_2)$. Finally, suppose $A = \cup_{n=1}^\infty A_n$. If $\sum_{n=1}^\infty \mu^\circ(A_n) = \infty$, then clearly $\mu^\circ(A) \leq \sum_{n=1}^\infty \mu^\circ(A_n)$. In the case that $\sum_{n=1}^\infty \mu^\circ(A_n) < \infty$, so $\mu^\circ(A_n) < \infty$ for all n, let B be a subset of A with a countable cover by $\mathcal{C}$, and for each n set $B_n = B \cap A_n$. Then B_n has a countable cover by $\mathcal{C}$, and $\mu^*(B_n) = \mu^\circ(B_n) \leq \mu^\circ(A_n) < \infty$ for each n. Fix $\epsilon > 0$, and for each n let $\{C_{nk}\}$ be a cover of B_n with

$$\sum_{k=1}^\infty \iota(C_{nk}) \leq \mu^\circ(B_n) + 2^{-n}\epsilon .$$

Then $\{C_{nk} : n, k \in \mathbb{N}\}$ is a cover of B, so

$$\mu^\circ(B) \leq \sum_{n,k} \iota(C_{nk}) \leq \sum_{n=1}^{\infty} \mu^\circ(B_n) + \epsilon \leq \sum_{n=1}^{\infty} \mu^\circ(A_n) + \epsilon \ .$$

Since ϵ is arbitrary, $\mu^\circ(B) \leq \sum_{n=1}^{\infty} \mu^\circ(A_n)$. Since $\mu^\circ(A)$ is the supremum of $\mu^\circ(B)$ over all such B, $\mu^\circ(A) \leq \sum_{n=1}^{\infty} \mu^\circ(A_n)$. 👍

The Extension Theorem holds for μ° with slight modification:

XVII.4.5.19. THEOREM. [EXTENSION THEOREM FOR μ°] Let $\mathcal{S}$ be a semiring on a set X, $\mathcal{A}$ the σ-algebra generated by $\mathcal{S}$, and ι a premeasure on $\mathcal{S}$. Let μ° be the outer measure on X induced by ι, $\mathcal{M}^\circ$ the σ-algebra of μ°-measurable subsets of X, and $\underline{\mu} = \mu^\circ|_{\mathcal{M}^\circ}$. Then

(i) $\mathcal{S} \subseteq \mathcal{M}^\circ$, i.e. every set in $\mathcal{S}$ is μ°-measurable.

(ii) The restriction of μ° to $\mathcal{S}$ is ι, i.e. $\underline{\mu}$ extends ι.

Thus ι extends to a complete measure $\underline{\mu}$ on $\mathcal{M}^\circ$, and hence to a measure $\underline{\mu}|_{\mathcal{A}}$ on $\mathcal{A}$. If ι is σ-finite, then $\underline{\mu}$ is the unique extension of ι to a measure on $\mathcal{B}$, and $(X, \mathcal{M}^\circ, \underline{\mu})$ is the completion of $(X, \mathcal{A}, \mu|_{\mathcal{A}})$.

PROOF: The proof of (ii) is identical to the proof for μ^*, since elements of $\mathcal{C}$ have countable $\mathcal{C}$-covers. We recall the argument for completeness. Let $C \in \mathcal{S}$. Since $\{C\}$ is a cover of C, we have $\mu^\circ(C) \leq \iota(C)$. For the opposite inequality, we may assume $\mu^\iota(C) < \infty$. Let $\epsilon > 0$, and let (C_n) be a sequence of elements of $\mathcal{S}$ with $C \subseteq \cup_{n=1}^{\infty} C_n$ and $\sum_{n=1}^{\infty} \iota(C_n) \leq \mu^\circ(C) + \epsilon$. Then, by (iii),

$$\iota(C) \leq \sum_{n=1}^{\infty} \iota(C_n) \leq \mu^\circ(C) + \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $\iota(C) \leq \mu^\circ(C)$.

(i): Now let $C \in \mathcal{S}$, and let $A \subseteq X$ with $\mu^\circ(A) < \infty$. We must show

$$\mu^\circ(A) \geq \mu^\circ(A \cap C) + \mu^\circ(A \setminus C) \ .$$

We have $\mu^\circ(A \cap C), \mu^\circ(A \setminus C) < \infty$. Fix $\epsilon > 0$, and let B_1 be a subset of $A \cap C$ which has a countable $\mathcal{C}$-cover and $\mu^\circ(A \cap C) < \mu^\circ(B) + \frac{\epsilon}{4}$, and B_2 a similar subset of $A \setminus C$ with $\mu^\circ(A \setminus C) < \mu^\circ(B_2) + \frac{\epsilon}{4}$, and set $B = B_1 \cup B_2$. Then $B \subseteq A$, so $\mu^\circ(B) \leq \mu^\circ(A)$, and B has a countable $\mathcal{C}$-cover. Let (C_n) be a sequence in $\mathcal{S}$ with $B \subseteq \cup_{n=1}^{\infty} C_n$ and $\sum_{n=1}^{\infty} \iota(C_n) \leq \mu^\circ(B) + \frac{\epsilon}{2}$. Then $B_1 = B \cap C \subseteq \cup_{n=1}^{\infty}(C_n \cap C)$, and $C_n \cap C \in \mathcal{S}$, so

$$\mu^\circ(B_1) \leq \sum_{n=1}^{\infty} \mu^\circ(C_n \cap C) = \sum_{n=1}^{\infty} \iota(C_n \cap C) \ .$$

Also, $B_2 \subseteq \cup_{n=1}^{\infty}(C_n \setminus C)$. We have that $C_n \setminus C$ is a finite disjoint union $\cup_{k=1}^{m_n} D_{nk}$ for each n, with $D_{nk} \in \mathcal{S}$. So

$$\mu^\circ(B_2) \leq \sum_{n=1}^{\infty} \sum_{k=1}^{m_n} \mu^\circ(D_{nk}) = \sum_{n=1}^{\infty} \sum_{k=1}^{m_n} \iota(D_{nk})$$

$$\mu^\circ(A\cap C)+\mu^\circ(A\setminus C)\le \mu^\circ(B_1)+\mu^\circ(B_2)+\frac{\epsilon}{2}\le \sum_{n=1}^{\infty}\iota(C_n\cap C)+\sum_{n=1}^{\infty}\sum_{k=1}^{m_n}\iota(D_{nk})+\frac{\epsilon}{2}$$
$$=\sum_{n=1}^{\infty}\left(\iota(C_n\cap C)+\sum_{k=1}^{m_n}\iota(D_{nk})\right)+\frac{\epsilon}{2}\ .$$

We have that, for each n, C_n is the finite disjoint union of $C_n\cap C$ and the D_{nk}, so by (ii)

$$\iota(C_n)=\iota(C_n\cap C)+\sum_{k=1}^{m_n}\iota(D_{nk})$$

and thus

$$\mu^\circ(A\cap C)+\mu^\circ(A\setminus C)\le\sum_{n=1}^{\infty}\iota(C_n)+\frac{\epsilon}{2}<\mu^\circ(B)+\epsilon\le\mu^\circ(A)+\epsilon\ .$$

Since $\epsilon>0$ is arbitrary, $\mu^*(A\cap C)+\mu^*(A\setminus C)\le\mu^*(A)$. Thus C is μ^*-measurable.
The last statement follows immediately from XVII.4.5.8., since $\mu^\circ=\mu^*$ in this case. 👍

XVII.4.5.20. The missing thing from XVII.4.5.8. in this result is the maximality of μ. It is tempting to expect that $\underline{\mu}$ is the smallest extension of ι to $\mathcal{A}$ or $\mathcal{M}^\circ$. This is true in many cases, but not in general for $\mathcal{M}^\circ$ (XVII.4.5.24.); it is true if ι takes only finite values on $\mathcal{S}$, and more generally for locally σ-finite premeasures:

XVII.4.5.21. Proposition. Let ι be a premeasure on a semiring $\mathcal{S}$ of subsets of a set X. Let $\mathcal{C}$ be the collection of subsets of X which can be covered by countably many elements of $\mathcal{S}$. Then the following are equivalent:

(i) Every set in $\mathcal{S}$ is a countable union of elements of $\mathcal{S}$ on which ι is finite.

(ii) Every set in $\mathcal{S}$ can be covered by countably many sets in $\mathcal{S}$ on which ι is finite.

(iii) Every set in $\mathcal{C}$ can be covered by countably many sets in $\mathcal{S}$ on which ι is finite.

The premeasure ι is *locally σ-finite* if these conditions are satisfied.

Proof: (i) $\Rightarrow$ (ii) $\Longleftrightarrow$ (iii) is trivial. For (ii) $\Rightarrow$ (i), if $B\in\mathcal{S}$ is covered by (S_n), then $B=\cup_n(B\cap S_n)$, and $B\cap S_n\in\mathcal{S}$ since $\mathcal{S}$ is closed under finite intersections, and $\iota(B\cap S_n)\le\iota(S_n)<\infty$. 👍

We then have the following corollary of XVII.4.5.15.:

XVII.4.5.22. Corollary. Let ι be a locally σ-finite premeasure on a semiring $\mathcal{S}$ of subsets of a set X, and μ° the outer measure of X defined in (). Then μ° is the smallest outer measure on X extending ι.

Proof: If ν^* is an outer measure on X extending ι, then $\nu^*(C) = \mu^*(C) = \mu^\circ(C)$ for $C \in \mathcal{C}$. If A is any subset of X, and C is any set in $\mathcal{C}$ with $C \subseteq A$, we have $\nu^*(A) \geq \nu^*(C)$ by monotonicity. Thus we have

$$\nu^*(A) \geq \sup_{C \in \mathcal{C}_A} \nu^*(C) = \sup_{C \in \mathcal{C}_A} \mu^*(C) = \mu^\circ(A)$$

where $\mathcal{C}_A$ is the collection of subsets of A which are in $\mathcal{C}$. 👍

XVII.4.5.23. Corollary. Let ι be a locally σ-finite premeasure on a semiring $\mathcal{S}$ of subsets of a set X, μ° the outer measure of X defined in (), and $\underline{\mu}$ the restriction of μ° to the σ-algebra $\mathcal{M}^\circ$. If $\mathcal{A}$ is the σ-algebra generated by $\mathcal{S}$, $\mathcal{B}$ any σ-algebra with $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{M}^\circ$, and ν a measure on $(X, \mathcal{B})$ extending ι, then $\nu \geq \underline{\mu}|_{\mathcal{B}}$.

Proof: The premeasure ν on $\mathcal{B}$ extends to an outer measure ν^* on X, which is an outer measure extending ι, and hence $\nu^* \geq \mu^\circ$. 👍

XVII.4.5.24. Example. Consider the following very simple example. Let $X = \{1, 2\}$, $\mathcal{S} = \{\emptyset, X\}$, and set $\iota(\emptyset) = 0$, $\iota(X) = \infty$. Then $\mu^\circ(A) = \mu^*(A) = \infty$ for any nonempty $A \subseteq X$. In this case $\mathcal{M}^\circ = \mathcal{M} = \mathcal{P}(X)$ and $\underline{\mu} = \mu = \mu^\circ = \mu^*$. But $\underline{\mu}$ is not the smallest extension of ι to a measure on $(X, \mathcal{P}(X))$; in fact, there is no smallest extension, since there are extensions μ_1 and μ_2 with $\mu_1(\{1\}) = \mu_2(\{2\}) = \infty$, $\mu_1(\{2\}) = \mu_2(\{1\}) = 0$. Note that in this case $\mathcal{S}$ is a σ-algebra and ι is a measure on $(X, \mathcal{S})$.

XVII.4.5.25. There is no obvious relation between the σ-algebras $\mathcal{M}$ and $\mathcal{M}^\circ$ in general, except that both contain $\mathcal{S}$ and hence $\mathcal{A}$. One indication that there may be a difference is that there are often many more $\underline{\mu}$-null sets than μ-null sets, e.g. in XVII.4.5.17. (although in this case $\mathcal{M}^\circ = \mathcal{M} = \mathcal{P}(X)$). But they cannot be too different:

XVII.4.5.26. Proposition. Let $\mathcal{S}$ be a semiring of subsets of a set X, ι a premeasure on $\mathcal{S}$, and μ^* and μ° the extensions of ι to outer measures on X. Let $\mathcal{A}$ be the σ-algebra generated by $\mathcal{S}$, and $\mathcal{M}$ and $\mathcal{M}^\circ$ the σ-algebras of μ^*-measurable and μ°-measurable sets respectively. Let E be a subset of X which can be covered by countably many sets in $\mathcal{S}$. Then E is in $\mathcal{M}^\circ$ if and only if it is in $\mathcal{M}$.

Proof: Let (S_n) be a sequence of sets in $\mathcal{S}$ covering E, and let $Y = \cup_n S_n$. Then $Y \in \mathcal{A} \subseteq \mathcal{M} \cap \mathcal{M}^\circ$, and $E \subseteq Y$. Every subset Z of Y can be covered by countably many sets in $\mathcal{S}$, and hence $\mu^\circ(Z) = \mu^*(Z)$, i.e. using the language of XVII.4.3.8., $\mu_Y^\circ = \mu_Y^*$. Thus by XVII.4.3.8., if $Z \subseteq Y$, we have

Z is μ°-measurable

$\iff$ Z is μ_Y°-measurable

$\iff$ Z is μ_Y^*-measurable

$\iff Z$ is μ^*-measurable.



Regular Outer Measures

Not every outer measure is induced by a premeasure. The ones that are can be characterized in a simple way:

XVII.4.5.27. DEFINITION. An outer measure μ^* on a set X is *regular* if, for every $E \subseteq X$, there is a μ^*-measurable set A with $E \subseteq A$ and $\mu^*(E) = \mu^*(A)$.

Not every outer measure is regular:

XVII.4.5.28. EXAMPLES. (i) If $X = \{0, 1\}$, set $\mu^*(\emptyset) = 0$, $\mu^*(\{0\}) = \mu^*(\{1\}) = 1$, and $\mu^*(X) = \alpha$. If $1 \leq \alpha \leq 2$, μ^* is an outer measure. If $\alpha < 2$, the only μ^*-measurable sets are $\emptyset$ and X; if $1 < \alpha < 2$, μ^* is not regular.

(ii) Let $X = \mathbb{N}$. Set $\mu^*(\emptyset) = 0$, $\mu^*(E) = 1$ if E is nonempty and finite, and $\mu^*(E) = 2$ if E is infinite. The only μ^*-measurable sets are $\emptyset$ and $\mathbb{N}$, and μ^* is not regular.

XVII.4.5.29. There is a related property. If μ^* is an outer measure on X and (E_n) is an increasing sequence of subsets of X, with $E = \cup_{n=1}^{\infty} E_n$, then $\mu^*(E_1) \leq \mu^*(E_2) \leq \cdots \leq \mu^*(E)$ by monotonicity; thus

$$\mu^*(E) \geq \lim_{n\to\infty} \mu^*(E_n) = \sup_n \mu^*(E_n) \ .$$

We do not have equality in general (XVII.4.5.31.).

XVII.4.5.30. DEFINITION. Let μ^* be an outer measure on a set X. Then μ^* is *monotone regular* if, whenever (E_n) is an increasing sequence of subsets of X, we have

$$\mu^*(E) = \lim_{n\to\infty} \mu^*(E_n) = \sup_n \mu^*(E_n) \ .$$

XVII.4.5.31. EXAMPLES. The example of XVII.4.5.28.(ii) is not monotone regular: let $E_n = \{1, \ldots, n\}$. The example of XVII.4.5.28.(i) is monotone regular (as is any outer measure on a finite set) but not regular.

XVII.4.5.32. PROPOSITION. A regular outer measure is monotone regular.

PROOF: Let μ^* be a regular outer measure on X, and (E_n) an increasing sequence of subsets of X, with $E = \cup_{n=1}^{\infty} E_n$. It suffices to show that $\mu^*(E) \leq \sup_n \mu^*(E_n)$. Let A_n be a μ^*-measurable set with $E_n \subseteq A_n$ and $\mu^*(A_n) = \mu^*(E_n)$ for all n. Set $B_n = \cap_{k=n}^{\infty} A_k$. We have that $E_n \subseteq A_k$ whenever $n \leq k$, and hence $E_n \subseteq B_n \subseteq B_{n+1}$ for all n. Since $B_n \subseteq A_n$, we have $\mu^*(E_n) \leq \mu^*(B_n) \leq \mu^*(A_n) = \mu^*(E_n)$, so we have equality throughout. If $B = \cup_{n=1}^{\infty} B_n$, then $E \subseteq B$, so by () we have

$$\mu^*(E) \leq \mu^*(B) = \sup_n \mu^*(B_n) = \sup_n \mu^*(E_n) \ .$$



XVII.4.5.33. PROPOSITION. An outer measure induced by a premeasure is regular.

PROOF: Suppose the outer measure μ^* on X is induced by $\iota : \mathcal{S} \to [0, \infty]$, and let $E \subseteq X$. If $\mu^*(E) = \infty$, take $A = X$. If $\mu^*(E) < \infty$, by XVII.4.1.10. let $A \in \mathcal{S}_{\sigma\delta}$ with $E \subseteq A$ and $\mu^*(A) = \mu^*(E)$.

XVII.4.5.34. COROLLARY. Lebesgue outer measure is a regular outer measure.

For the converse, we prove a more precise statement:

XVII.4.5.35. PROPOSITION. Let μ^* be an outer measure on X, $\mathcal{M}$ the set of μ^*-measurable sets, and $\mu^\star$ the outer measure induced by $\mu^*|_{\mathcal{M}}$. Then $\mu^\star$ is a regular outer measure, $\mu^\star \geq \mu^*$, every set in $\mathcal{M}$ is $\mu^\star$-measurable, every $\mu^\star$-measurable set of σ-finite $\mu^\star$-measure is in $\mathcal{M}$, and $\mu^\star|_{\mathcal{M}} = \mu^*|_{\mathcal{M}}$; $\mu^\star = \mu^*$ if and only if μ^* is regular.

PROOF: Most of this follows immediately from previous results. By XVII.4.5.8., setting $\mathcal{S} = \mathcal{M}$ and $\iota = \mu^*|_{\mathcal{M}}$, every set in $\mathcal{M}$ is $\mu^\star$-measurable, and $\mu^\star|_{\mathcal{M}} = \mu^*|_{\mathcal{M}}$. The outer measure $\mu^\star$ is regular by XVII.4.5.33.. We have that, for any $E \subseteq X$,

$$\mu^\star(E) = \inf\{\mu^*(A) : A \in \mathcal{M}, E \subseteq A\}$$

and the infimum is attained () at some $A \in \mathcal{M}$; for this A we have $\mu^*(E) \leq \mu^*(A) = \mu^\star(E)$. Thus $\mu^\star \geq \mu^*$.

If E is $\mu^\star$-measurable and $\mu^\star(E) < \infty$, then there is a μ^*-measurable set A with $E \subseteq A$ and $\mu^*(A) = \mu^\star(E)$; then $\mu^\star(A) = \mu^*(A) = \mu^\star(E)$, so $\mu^\star(A \setminus E) = 0$. Since $\mu^*(A \setminus E) \leq \mu^\star(A \setminus E)$, $\mu^*(A \setminus E) = 0$ and $A \setminus E \in \mathcal{M}$ (). Thus $E = A \setminus (A \setminus E) \in \mathcal{M}$. The case where E has σ-finite $\mu^\star$-measure follows immediately.

Finally, if μ^* is regular, let $E \subseteq X$, and $A \in \mathcal{M}$ with $E \subseteq A$ and $\mu^*(E) = \mu^*(A)$. Then $\{A\}$ is an $\mathcal{M}$-cover of E, so $\mu^\star(E) \leq \mu^*(A) = \mu^*(E)$. Thus $\mu^\star \leq \mu^*$. So $\mu^\star = \mu^*$. 

XVII.4.5.36. EXAMPLE. [Rog70, p. 14] The σ-algebra of $\mu^\star$-measurable sets can be strictly larger than $\mathcal{M}$. Let X be a set of cardinality $\aleph_2$, and define μ^* on X by setting $\mu^*(E) = 0$ if E is countable, $\mu^*(E) = 1$ if $card(E) = \aleph_1$, and $\mu^*(E) = \infty$ if $card(E) = \aleph_2$. Then μ^* is an outer measure on X which is not regular. The σ-algebra $\mathcal{M}$ of μ^*-measurable sets is the σ-algebra of countable and cocountable sets. Construct $\mu^\star$ as above from $\mu^*|_{\mathcal{M}}$. Then, for any $E \subseteq X$, $\mu^\star(E) = 0$ if E is countable and $\mu^\star(E) = \infty$ if E is uncountable, and every subset of X is $\mu^\star$-measurable.

Putting the last two results together, we obtain:

XVII.4.5.37. PROPOSITION. An outer measure is induced by a premeasure if and only if it is regular.

As a corollary of XVII.4.5.35., we can show that Lebesgue's definition of measurability agrees with Carathéodory's for finite regular outer measures:

XVII.4.5.38. THEOREM. Let μ^* be a regular outer measure on a set X, with $\mu^*(X) < \infty$. Then $A \subseteq X$ is μ^*-measurable if and only if $\mu^*(A) + \mu^*(X \setminus A) = \mu^*(X)$.

(Of course, $\mu^*(A) + \mu^*(X \setminus A) \geq \mu^*(X)$ is automatic, so only the reverse inequality needs to be checked.) PROOF: If A is μ^*-measurable, the equality holds by definition. Conversely, suppose $A \subseteq X$ and $\mu^*(A) + \mu^*(X \setminus A) = \mu^*(X)$. Let B and C be μ^*-measurable subsets of X with $A \subseteq B$, $\mu^*(B) = \mu^*(A)$, $(X \setminus A) \subseteq C$, $\mu^*(C) = \mu^*(X \setminus A)$. We have

$$\mu^*(B \setminus C) + \mu^*(B \cap C) + \mu^*(C \setminus B) = \mu^*(B \cup C) = \mu^*(X) \ .$$

We also have

$$\mu^*(X) = \mu^*(A) + \mu^*(X \setminus A) = \mu^*(B) + \mu^*(C) = \mu^*(B \setminus C) + 2\mu^*(B \cap C) + \mu^*(C \setminus B)$$

and subtracting we get $\mu^*(B \cap C) = 0$. If $D = B \setminus A$, then $D = B \cap (X \setminus A) \subseteq (B \cap C)$, so $\mu^*(D) = 0$ and D is μ^*-measurable. Hence so is $A = B \setminus D$. 👍

This result can fail for finite outer measures which are not regular (Exercise ()). It also usually fails for outer measures which are not finite: for example, there is a nonmeasurable subset A of $\mathbb{R}$ with $\lambda^*(A) = \lambda^*(\mathbb{R} \setminus A) = \infty$ ().

Inner Measure

XVII.4.5.39. If μ^* is a finite regular outer measure on a set X, we can define the *inner measure* of a subset A of X by

$$\mu_*(A) = \mu^*(X) - \mu^*(X \setminus A) \ .$$

Then the result of XVII.4.5.38. can be rephrased:

XVII.4.5.40. COROLLARY. Let μ^* be a finite regular outer measure on a set X, and μ_* the corresponding inner measure. Then $A \subseteq X$ is μ^*-measurable if and only if $\mu^*(A) = \mu_*(A)$.

Generalizations of Regularity

XVII.4.5.41. The notion of regularity can be generalized. If $\mathcal{C}$ is a collection of subsets of a set X, and μ^* is an outer measure on X, then μ^* is *$\mathcal{C}$-regular* if every subset Y of X is contained in some $C \in \mathcal{C}$ with $\mu^*(C) = \mu^*(Y)$ (this implies that $X \in \mathcal{C}$). Normally this definition is applied only when $\mathcal{C} \subseteq \mathcal{M}$, where $\mathcal{M}$ is the set of μ^*-measurable sets. Thus μ^* is regular if and only if it is $\mathcal{M}$-regular. If μ^* is a σ-finite outer measure induced from $\iota : \mathcal{C} \to [0, \infty]$, then μ^* is $\mathcal{C}_{\sigma\delta}$-regular (XVII.4.1.10.).

XVII.4.5.42. Proposition. Let μ^* be an outer measure on a set X, $\mathcal{M}$ the set of μ^*-measurable subsets of X, and $\mathcal{A}$ a sub-σ-algebra of $\mathcal{M}$ such that $\mu^*|_{\mathcal{A}}$ is σ-finite. If μ^* is $\mathcal{A}$-regular, then $\mathcal{M}$ is the σ-algebra generated by $\mathcal{A}$ and the $\mathcal{A}$-null sets (subsets of sets in $\mathcal{A}$ of outer measure 0). In particular, if $(X, \mathcal{A}, \mu^*|_{\mathcal{A}})$ is complete and σ-finite, then $\mathcal{A} = \mathcal{M}$.

XVII.4.5.43. This does not mean that if $(X, \mathcal{A}, \mu)$ is a σ-finite complete measure space, then it is not possible to extend μ to a measure on a strictly larger σ-algebra; there are many examples where this is possible. See Exercise () for a simple example, and (), (), ... for more natural and important ones. What XVII.4.5.42. does say is that if such a μ is extended to a measure ν on a larger σ-algebra $\mathcal{B}$, and ν^* is the outer measure induced by ν, then ν^* is not $\mathcal{A}$-regular.

XVII.4.5.44. An important special case of regularity is when μ^* is an outer measure on a topological space X. Then μ^* is *Borel regular* if it is $\mathcal{B}$-regular, where $\mathcal{B}$ is the class of Borel sets in X. This definition is usually reserved for Borel outer measures, i.e. ones where all Borel sets are μ^*-measurable. The most important example is that Lebesgue outer measure is Borel regular (XVII.4.4.5., XVII.4.5.34.).

XVII.4.6. The Suslin Operation

XVII.4.6.1. There is a curious operation on sets called the *Suslin operation* (sometimes called the *A-operation*), which is reminiscent of the construction in the proof of XIII.10.4.7.. For each finite sequence $(s_1, \ldots, s_n)$ of natural numbers, let $A_{s_1,\ldots,s_n}$ be a subset of a set X. The collection of finite sequences in $\mathbb{N}$ is countable, so there are countably many such sets in all. Define

$$S(\{A_{s_1,\ldots,s_n}\}) = \bigcup_{s \in \mathcal{N}} \bigcap_{n=1}^{\infty} A_{s_1,\ldots,s_n}$$

where $\mathcal{N} = \mathbb{N}^\infty$ is the set of all sequences in $\mathbb{N}$ (). Although the operation begins with a countable collection of sets, the union is over all sequences (uncountably many) in $\mathcal{N}$.

If $s = (s_1, s_2, \ldots) \in \mathcal{N}$, write $s|n = (s_1, \ldots, s_n) \in \mathbb{N}^n$. These are called *initial segments* of s. Take $s|0$ to be the "empty string." Thus

$$S(\{A_{s_1,\ldots,s_n}\}) = \bigcup_{s \in \mathcal{N}} \bigcap_{n=1}^{\infty} A_{s|n} \ .$$

XVII.4.6.2. More generally, if $\mathcal{E}$ is a nonempty collection of subsets of a set X, and all the $A_{s_1,\dots,s_n}$ are in $\mathcal{E}$, the indexed collection $\{A_{s_1,\dots,s_n}\}$ is called a *Suslin scheme in* $\mathcal{E}$. Denote by $\mathcal{S}(\mathcal{E})$ the collection $\{S(\{A_{s_1,\dots,s_n}\})\}$, for all Suslin schemes $\{A_{s_1,\dots,s_n}\}$ in $\mathcal{E}$, along with $\emptyset$ (we will automatically have $\emptyset \in \mathcal{S}(\mathcal{E})$ in many cases, e.g. if $\emptyset$ is a countable intersection of sets in $\mathcal{E}$). The sets in $\mathcal{S}(\mathcal{E})$ are called the *$\mathcal{E}$-Suslin sets* (sometimes called the *$\mathcal{E}$-analytic sets*, but we reserve this term for sets defined differently in (); the two classes of sets coincide in most natural settings). It is obvious that if $\mathcal{E} \subseteq \mathcal{F}$, then $\mathcal{S}(\mathcal{E}) \subseteq \mathcal{S}(\mathcal{F})$.

We typically take $\mathcal{E}$ to be the collection of closed or compact sets in $\mathbb{R}^n$ (or in a topological space), or a semialgebra (algebra, σ-algebra) of sets such as the Borel sets in $\mathbb{R}^n$.

The sets in $\mathcal{S}(\mathcal{E})$ have a simpler explicit description than the sets in $\sigma(\mathcal{E})$ (cf. ()). The collection $\mathcal{S}(\mathcal{E})$ is usually different from $\sigma(\mathcal{E})$ although it has some of the same properties; it frequently happens that $\mathcal{S}(\mathcal{E})$ contains $\sigma(\mathcal{E})$ (XVII.4.6.4.(iv)).

XVII.4.6.3. A Suslin scheme $\{A_{s_1,\dots,s_n}\}$ is *monotone* (or *regular*) if $A_{s_1,\dots,s_n,s_{n+1}} \subseteq A_{s_1,\dots,s_n}$ for all $(s_1,\dots,s_{n+1})$. If $\mathcal{E}$ is closed under finite intersections (e.g. in one of the common situations described above), then for any Suslin scheme $\{A_{s_1,\dots,s_n}\}$ in $\mathcal{E}$ there is a monotone Suslin scheme $\{B_{s_1,\dots,s_n}\}$ with $S(\{B_{s_1,\dots,s_n}\}) = S(\{A_{s_1,\dots,s_n}\})$: just set

$$B_{s_1,\dots,s_n} = A_{s_1} \cap A_{s_1,s_2} \cap \cdots \cap A_{s_1,\dots,s_n}$$

for each $(s_1,\dots,s_n)$.

XVII.4.6.4. PROPOSITION. Let $\mathcal{E}$ be a collection of subsets of a set X. Then

(i) $\mathcal{E} \subseteq \mathcal{S}(\mathcal{E})$.

(ii) $\mathcal{S}(\mathcal{S}(\mathcal{E})) = \mathcal{S}(\mathcal{E})$. Thus if $\mathcal{E} \subseteq \mathcal{F} \subseteq \mathcal{S}(\mathcal{E})$, then $\mathcal{S}(\mathcal{F}) = \mathcal{S}(\mathcal{E})$.

(iii) $\mathcal{S}(\mathcal{E})$ is closed under countable unions and countable intersections.

(iv) If the complement of every set in $\mathcal{E}$ is in $\mathcal{S}(\mathcal{E})$, then $\sigma(\mathcal{E}) \subseteq \mathcal{S}(\mathcal{E})$.

PROOF: (i): If $A \in \mathcal{E}$, set $A_{s_1,\dots,s_n} = A$ for all $(s_1,\dots,s_n)$. Then $\{A_{s_1,\dots,s_n}\}$ is a Suslin scheme in $\mathcal{E}$, and $S(\{A_{s_1,\dots,s_n}\}) = A$.
(ii): Let $\phi : \mathbb{N}^2 \to \mathbb{N}$ be the bijection given by $\phi(n,m) = 2^{n-1}(2m-1)$. Let α and β be the coordinate functions of the inverse map, i.e. $\alpha(2^{n-1}(2m-1)) = n$, $\beta(2^{n-1}(2m-1)) = m$. Define a bijection $\psi : \mathcal{N} \times \mathcal{N}^\infty \to \mathcal{N}$ by

$$\psi(s,(t^{(k)}))(r) = \phi(s_r, t^{(\alpha(r))}_{\beta(r)})$$

where $t^{(k)}_j$ is the j'th term of the sequence $t^{(k)}$. Note that if $r = \phi(n,m)$, then $\psi(s,(t^{(k)}))|r$ completely determines $s|n$ and $t^{(n)}|m$, i.e. if $(p_1,\dots,p_r) \in \mathbb{N}^r$, then there are $(s_1,\dots,s_n)$ and $(t_1,\dots,t_m)$ such that whenever $\psi(s,(t^{(k)}))|r = (p_1,\dots,p_r)$, $s|n = (s_1,\dots,s_n)$ and $t^{(n)}|m = (t_1,\dots,t_m)$.

Now let $B = \cup_{s\in\mathcal{N}} \cap_n B_{s|n} \in \mathcal{S}(\mathcal{S}(\mathcal{E}))$, where $\{B_{s_1,\dots,s_n}\}$ is a Suslin scheme in $\mathcal{S}(\mathcal{E})$, and suppose for each $(s_1,\dots,s_n)$ we have $B_{s_1,\dots,s_n} = \cup_{t\in\mathcal{N}} \cap_m A^{t|m}_{s_1,\dots,s_m}$ with $A^{t|m}_{s_1,\dots,s_n} \in \mathcal{E}$. For a finite sequence $(p_1,\dots,p_r)$, let $(s_1,\dots,s_n)$ and $(t^{(n)}_1,\dots,t^{(n)}_m)$ be the initial segments determined by $(p_1,\dots,p_r)$ as above, and define

$$E_{p_1,\dots,p_r} = A^{t^{(n)}_1,\dots,t^{(n)}_m}_{s_1,\dots,s_n} .$$

Then $\{E_{p_1,\dots,p_r}\}$ is a Suslin scheme in $\mathcal{E}$; set $E = \cup_{p\in\mathcal{N}} \cap_r E_{p|r} \in \mathcal{S}(\mathcal{E})$.

We have that $x \in B$ if and only if there is an $s \in \mathcal{N}$ such that $x \in B_{s|n}$ for all n, i.e. if and only if for each n there is $(t^{(n)}) \in \mathcal{N}^\infty$ such that $x \in A_{s|n}^{t^{(n)}|m}$ for all m. Set $p = \psi(s, (t^{(n)}))$. Then, for any r, $E_{p|r} = A_{s|n}^{t^{(n)}|m}$ where $r = \phi(n, m)$, so $x \in B$ if and only if $x \in E_{p|r}$ for all r, for some $p \in \mathcal{N}$, i.e. if and only if $x \in E$. Thus $B = E \in \mathcal{S}(\mathcal{E})$. So $\mathcal{S}(\mathcal{S}(\mathcal{E})) \subseteq \mathcal{S}(\mathcal{E})$, and the opposite inclusion follows from (i).

(iii): By (ii), it suffices to show that $\mathcal{S}(\mathcal{E})$ contains all countable unions and countable intersections of sets in $\mathcal{E}$. Suppose (E_n) is a sequence of sets in $\mathcal{E}$. For each $(s_1, \dots, s_n)$, set $A_{s_1,\dots,s_n} = E_{s_1}$. Then $\{A_{s_1,\dots,s_n}\}$ is a Suslin scheme in $\mathcal{E}$, and $S(\{A_{s_1,\dots,s_n}\}) = \cup_{n=1}^\infty E_n$. Similarly, set $A_{s_1,\dots,s_n} = E_n$ for all $(s_1, \dots, s_n)$; then $S(\{A_{s_1,\dots,s_n}\}) = \cap_{n=1}^\infty E_n$.

(iv): Follows immediately from XVI.1.7.2.(b) (cf. XVII.5.3.12.), since $\mathcal{E} \subseteq \mathcal{S}(\mathcal{E}) \cap \widetilde{\mathcal{S}(\mathcal{E})}$ which is a σ-algebra. ☝

XVII.4.6.5. $\mathcal{S}(\mathcal{E})$ is not closed under complements in general, even if $\mathcal{E}$ is a σ-algebra. For example, if $\mathcal{B}$ is the σ-algebra of Borel sets in $\mathbb{R}$, then $\mathcal{S}(\mathcal{B})$ is not closed under complements ().

XVII.4.6.6. Let $\mathcal{K}$ be the collection of compact subsets of $\mathbb{R}^n$, $\mathcal{F}$ the collection of closed subsets, and $\mathcal{B}$ the collection of Borel sets. The complement of every compact set is open, hence a countable union of compact sets; thus by XVII.4.6.4.(iv) $\sigma(\mathcal{K}) = \mathcal{B} \subseteq \mathcal{S}(\mathcal{K})$. Thus by (ii) $\mathcal{S}(\mathcal{K}) = \mathcal{S}(\mathcal{F}) = \mathcal{S}(\mathcal{B})$, i.e. every $\mathcal{B}$-Suslin set is $\mathcal{K}$-Suslin. These sets are called the *Suslin sets* in $\mathbb{R}^n$. See also Exercise XVII.5.3.9..

The next theorem is the main result of this subsection.

XVII.4.6.7. Theorem. Let μ^* be an outer measure on a set X, $\mathcal{M}$ the σ-algebra of μ^*-measurable sets, and $\mu = \mu^*|_{\mathcal{M}}$. Then

(i) $\mathcal{M}$ is closed under the Suslin operation, i.e. $\mathcal{S}(\mathcal{M}) = \mathcal{M}$.

(ii) Suppose μ^* is monotone regular. Let $\mathcal{E}$ be a subset of $\mathcal{M}$ which is closed under finite unions and countable intersections, and such that $\mu(E) < \infty$ for all $E \in \mathcal{E}$. Then, for every $A \in \mathcal{S}(\mathcal{E})$ with $\mu(A) < \infty$, we have
$$\mu(A) = \sup\{\mu(E) : E \in \mathcal{E}, E \subseteq A\} .$$

Combining XVII.4.6.7. with XVII.4.5.12., we obtain:

XVII.4.6.8. Corollary. Let $(X, \mathcal{A}, \mu)$ be a complete σ-finite measure space. Then $\mathcal{A}$ is closed under the Suslin operation. ("σ-finite" can be replaced by "saturated.")

XVII.4.6.9. The conclusion of this corollary can be false if the measure is not complete. In fact, the σ-algebra of Borel sets on $\mathbb{R}$ is not closed under the Suslin operation ().

We now prove Theorem XVII.4.6.7..

PROOF: ([Rog70], [Bog07]) We first prove (i) and (ii) in the case where μ^* is monotone regular. Let $A = S(\{A_{s_1,\dots,s_n}\}) \in \mathcal{S}(\mathcal{M})$, where $\{A_{s_1,\dots,s_n}\}$ is a Suslin scheme in $\mathcal{M}$ which we may assume is monotone since $\mathcal{M}$ is closed under finite intersections. For (ii), we take the $A_{s_1,\dots,s_n}$ to be in $\mathcal{E}$. To show that $A \in \mathcal{M}$, we use the criterion of XVII.4.3.2.(ii). So let $Y \subseteq A$ and $Z \subseteq X \setminus A$ with $\mu^*(Y), \mu^*(Z) < \infty$; in (ii) we take $Y = A$ (and in this case we can take $Z = \emptyset$). For (i) we must show that $\mu^*(Y \cup Z) \geq \mu^*(Y) + \mu^*(Z)$. We will show that for any $\epsilon > 0$ there is an $E \in \mathcal{M}$, $E \subseteq A$, with $\mu^*(Y \cap E) \geq \mu^*(Y) - \epsilon$, and $E \in \mathcal{E}$ in case (ii); (ii) will then follow immediately, and (i) will also follow since then $Z \subseteq X \setminus E$, so

$$\mu^*(Y \cup Z) \geq \mu^*((Y \cap E) \cup Z) = \mu^*(Y \cap E) + \mu^*(Z) \geq \mu^*(Y) - \epsilon + \mu^*(Z)$$

and since $\epsilon > 0$ is arbitrary, $\mu^*(Y \cup Z) \geq \mu^*(Y) + \mu^*(Z)$.

Fix $\epsilon > 0$. For $m \in \mathbb{N}$, set

$$B_m = Y \cap \bigcup_{s \in \mathcal{N}_m} \bigcap_k A_{s|k}$$

where $\mathcal{N}_m = \{s \in \mathcal{N} : s_1 \leq m\}$. Then the B_m increase with m, and $\cup_m B_m = Y$. So

$$\mu^*(Y) = \sup_m \mu^*(B_m)$$

since μ^* is monotone regular. Thus there is an m_1 such that

$$\mu^*(B_{m_1}) > \mu^*(Y) - \epsilon \ .$$

Proceed inductively. Suppose $m_1, \dots, m_n$ have been defined and

$$B_{m_1,\dots,m_n} = Y \cap \bigcup_{s \in \mathcal{N}_{m_1,\dots,m_n}} \bigcap_k A_{s|k}$$

where $\mathcal{N}_{m_1,\dots,m_n} = \{s \in \mathcal{N} : s_1 \leq m_1, \dots, s_n \leq m_n\}$ and

$$\mu^*(B_{m_1,\dots,m_n}) > \mu^*(Y) - \epsilon \ .$$

For $m \in \mathcal{N}$, set

$$B_{m_1,\dots,m_n,m} = Y \cap \bigcup_{s \in \mathcal{N}_{m_1,\dots,m_n,m}} \bigcap_k A_{s|k}$$

where $\mathcal{N}_{m_1,\dots,m_n,m} = \{s \in \mathcal{N} : s_1 \leq m_1, \dots, s_n \leq m_n, s_{n+1} \leq m\}$. Then the $B_{m_1,\dots,m_n,m}$ increase in m, and $\cup_m B_{m_1,\dots,m_n,m} = B_{m_1,\dots,m_n}$, so as before there is an m_{n+1} such that

$$\mu^*(B_{m_1,\dots,m_{n+1}}) > \mu^*(Y) - \epsilon \ .$$

Now for each n set

$$E_n = \bigcup_{s \in \mathcal{N}_{m_1,\dots,m_n}} A_{s|n} \ .$$

This is a finite union, so $E_n \in \mathcal{M}$ for all n, and $E_n \in \mathcal{E}$ in case (ii). We have $B_{m_1,\dots,m_n} \subseteq Y \cap E_n$ for all n. We also have $E_{n+1} \subseteq E_n$ for all n. Set $E = \cap_n E_n$.

We claim that $\mu^*(E) \geq \mu^*(Y) - \epsilon$. This is obvious from downward continuity () in case (ii) since μ is finite on $\mathcal{E}$ and $\mu^*(E_n) \geq \mu^*(B_{m_1,\dots,m_n}) > \mu^*(Y) - \epsilon$ for all n. For case (i), define a new outer measure μ_Y^* on X by

$$\mu_Y^*(S) = \mu^*(Y \cap S)$$

for $S \subseteq X$. Then every set in $\mathcal{M}$ is μ_Y^*-measurable by (), and we have

$$\mu^*(E) \geq \mu_Y^*(E) = \inf_n \mu_Y^*(E_n) = \inf_n \mu^*(Y \cap E_n) \geq \inf_n \mu^*(B_{m_1,\dots,m_n}) \geq \mu^*(Y) - \epsilon$$

where the first equality is downward continuity for the finite measure μ_Y.

Finally, we show $E \subseteq A$. Let $x \in E$. Write

$$\mathbb{N}^n_{m_1,\dots,m_n} = \{(s_1,\dots,s_n) : s_1 \leq m_1, \dots, s_n \leq m_n\} \subseteq \mathbb{N}^n$$

for each n. If $k > n$, say $(s_1,\dots,s_n) \in \mathbb{N}^n_{m_1,\dots,m_n}$ is *extendible to* k if there are $s_{n+1},\dots,s_k$ such that $(s_1,\dots,s_n,s_{n+1},\dots,s_k) \in \mathbb{N}^k_{m_1,\dots,m_k}$ and $x \in A_{s_1,\dots,s_k}$. Since $x \in E_k$ for all k, there is for every k an s_1 for which (s_1) is extendible to k. Since there are only finitely many $s_1 \leq m_1$, there is an s_1 such that (s_1) is extendible to arbitrarily large k. Similarly, for this s_1 there is an s_2 such that (s_1,s_2) is extendible to arbitrarily large k. Continuing inductively, a sequence $s = (s_1, s_2, \dots)$ is generated such that $x \in A_{s|n}$ for all n, i.e. $x \in \cap_n A_{s|n} \subseteq A$.

This completes the proof of (ii), and of (i) if μ^* is monotone regular.

For (i) in the case of general μ^*, let A, Y, Z be as above. Form the outer measure μ_W^*, where $W = Y \cup Z$. Then every set in $\mathcal{M}$ is μ_W^*-measurable. We also have that $\mu_W^*(X \setminus W) = 0$, hence $X \setminus W$ and also W are μ_W^*-measurable. Let ι be the restriction of μ_W^* to the μ_W^*-measurable sets, and ν^* the outer measure on X induced by ι. Then ν^* is a regular outer measure, hence monotone regular (XVII.4.5.32.), and every μ_W^*-measurable set is ν^*-measurable (XVII.4.5.35.); in particular, W and all sets in $\mathcal{M}$ are ν^*-measurable. So

$$\nu^*(W) = \iota(W) = \mu_W^*(W) = \mu(W) \ .$$

We also have $\nu^*(S) \geq \mu_W^*(S)$ for all $S \subseteq X$ (XVII.4.5.35.); in particular, $\nu^*(Y) \geq \mu_W^*(Y) = \mu^*(Y)$ and $\nu^*(Z) \geq \mu_W^*(Z) = \mu^*(Z)$. By the first part of the proof, A is ν^*-measurable, so

$$\mu^*(Y \cup Z) = \mu^*(W) = \nu^*(W) = \nu^*(W \cap A) + \nu^*(W \cap A^c) = \nu^*(Y) + \nu^*(Z) \geq \mu^*(Y) + \mu^*(Z)$$

and it follows that A is μ^*-measurable.

This completes the proof of XVII.4.6.7.. 👍

XVII.4.7. Local Measures on σ-Rings and δ-Rings

In some situations it is useful to consider "measures" defined on σ-rings or, more generally, δ-rings; in fact, some authors such as [Hal50] define measures using σ-rings instead of σ-algebras. The theory is largely identical with a few modest technical complications, especially after the extension theory of this chapter has been developed. The primary utility of the increased generality comes with considering signed and complex measures (); for ordinary measures there is little advantage. Although many authors call these more general functions "measures," we will reserve the term *measure* for ones defined on σ-algebras to avoid any confusion, and use the term "local measure" for the more general version (the term is nonstandard but will be justified in XVII.4.7.2.), but we will typically use the letters μ, ν, etc. for local measures as well as measures. ✋

XVII.4.7.1. DEFINITION. Let X be a set, and $\mathcal{D}$ a δ-ring of subsets of X. A premeasure (XVII.4.5.6.) on $\mathcal{D}$ is called a *local measure* on $\mathcal{D}$. A local measure is *finite* if it does not take the value $+\infty$.

A local measure on a δ-ring $\mathcal{D}$ is thus a function $\mu : \mathcal{D} \to [0, +\infty]$ with $\mu(\emptyset) = 0$ and which is countably additive on sequences of pairwise disjoint sets which are "uniformly bounded" by a set in $\mathcal{D}$. If $\mathcal{D}$ is a σ-ring, a local measure is unrestrictedly countably additive, thus is indistinguishable from a measure except that unrestricted complements are not allowed; in fact, the difference only occurs in situations where X is not a countable union of sets in $\mathcal{D}$, and thus essentially does not happen if μ is finite (XVII.4.7.3.).

XVII.4.7.2. If μ is a local measure on a δ-ring $\mathcal{D}$, The Extension Theorems XVII.4.5.8. and XVII.4.5.19. give two canonical extensions of μ to a measure on the σ-algebra $\mathcal{A}$ generated by $\mathcal{D}$ (and the proof of the Extension Theorem simplifies greatly in this case). It is better to work with the extension $\underline{\mu}$ of XVII.4.5.19.. Since $\mathcal{D}$ is a hereditary subset of $\mathcal{A}$ (XVI.1.4.8.), μ is thus the restriction of an honest measure to a hereditary subset of its σ-algebra, justifying the term "local measure." Here the canonical extension $\underline{\mu}$ can be replaced by any extension of μ to $\mathcal{A}$. Conversely, any such restriction of a measure to a hereditary subset of its σ-algebra is a local measure. In particular, if μ is a local measure on $\mathcal{D}$, and $Y \in \mathcal{D}$, then the restriction of μ to the σ-algebra $\mathcal{D}_Y$ of elements of $\mathcal{D}$ which are subsets of Y is a measure on $(Y, \mathcal{D}_Y)$, further justifying the term "local measure."

XVII.4.7.3. If μ is a local measure on a δ-ring $\mathcal{D}$, the canonical extension $\underline{\mu}$ to the generated σ-algebra $\mathcal{A}$ is often not finite even if μ is finite (consider counting measure on the δ-ring of finite subsets of an infinite set; see XVII.4.7.4. for a more interesting example). The local measure μ is not necessarily bounded ($\underline{\mu}$ will be finite if and only if μ is bounded), so there may be no finite measure on $\mathcal{A}$ extending μ. But if $\mathcal{D}$ is a σ-ring and μ is finite, then $\underline{\mu}$ is a finite measure on $\mathcal{A}$ extending μ and there is a $Y \in \mathcal{D}$ with $\underline{\mu}(X \setminus Y) = 0$ (Exercise XVII.5.3.12.), and thus μ is effectively just a measure on $\mathcal{D}_Y$.

XVII.4.7.4. EXAMPLE. Let $\mathcal{D}$ be the set of bounded Lebesgue measurable sets in $\mathcal{D}$. Then $\mathcal{D}$ is a δ-ring, and the restriction of Lebesgue measure on $\mathbb{R}$ to $\mathcal{D}$ is a finite local measure. This $\mathcal{D}$ could be replaced by the set of bounded Borel sets, or the collection of Lebesgue measurable sets (or Borel sets) of finite Lebesgue measure. In each case the (unique) extension is Lebesgue measure.

Canonical Local Measures Associated With a Measure

XVII.4.7.5. Let $(X, \mathcal{A}, \mu)$ be a measure space. There are two obvious local measures associated with μ. Let $\mathcal{F}_\mu = \{A \in \mathcal{A} : \mu(A) < +\infty\}$. Then $\mathcal{F}_\mu$ is a δ-ring, and the restriction of μ to $\mathcal{F}_\mu$ is finite. We can also let $\mathcal{S}_\mu$ be the set of sets in $\mathcal{A}$ of σ-finite measure; then $\mathcal{S}_\mu$ is a σ-ring, equal to $\mathcal{A}$ if and only if μ is σ-finite. We of course have $\mathcal{F}_\mu \subseteq \mathcal{S}_\mu$. The σ-algebra generated by $\mathcal{S}_\mu$ (and, *a fortiori*, $\mathcal{F}_\mu$) is not all of $\mathcal{A}$ in general (we could even have $\mathcal{S}_\mu = \{\emptyset\}$ in extreme cases). For integration purposes, the local measure $\mu|_{\mathcal{S}_\mu}$ carries all the information of μ. Note that if μ is a finite local measure on a δ-ring $\mathcal{D}$, and ν is an extension of μ (canonical or not) to the σ-algebra $\mathcal{A}$ generated by $\mathcal{D}$, then $\mathcal{F}_\nu$ may be strictly larger than $\mathcal{D}$: for example, if $\mathcal{D}$ is the set of bounded Borel sets in $\mathbb{R}$, so that $\mathcal{A}$ is the σ-algebra of Borel sets $\mathcal{B}$, and μ is the restriction of Lebesgue measure to $\mathcal{D}$, then λ is the unique extension of μ to $\mathcal{B}$, but $\mathcal{F}_\lambda$ is the set of Borel sets of finite Lebesgue measure, which is strictly larger than $\mathcal{D}$.

XVII.4.8. Radon Measures on $\mathbb{R}$

XVII.4.9. Lebesgue Measure and Radon Measures on $\mathbb{R}^n$

XVII.4.10. Nonmeasurable Subsets of $\mathbb{R}$

An interesting and rather subtle question is whether every subset of $\mathbb{R}$ is Lebesgue measurable. Using the Axiom of Choice, we construct a family of nonmeasurable subsets. This construction is due to Vitali.

XVII.4.10.1. Define a relation R on $\mathbb{R}$ by xRy if $y-x \in \mathbb{Q}$. It is easily checked that R is an equivalence relation. If $a \in \mathbb{R}$, then the equivalence class of a is

$$[a] = a + \mathbb{Q} = \{a + q : q \in \mathbb{Q}\} .$$

Thus the equivalence classes are countable and dense. (They are exactly the cosets of $\mathbb{Q}$ in the additive group $\mathbb{R}$.) As for any equivalence relation, the equivalence classes are pairwise disjoint and their union is $\mathbb{R}$.

XVII.4.10.2. Definition. (AC) Let P be a subset of $(0,1)$ which contains exactly one element of each equivalence class of R. If $q \in \mathbb{Q}$, set $P_q = P + q$.

Since each equivalence class is dense, it has nonempty intersection with $(0,1)$. The existence of such a set P is then guaranteed by the Axiom of Choice. (Any other actual interval could be used in place of $(0,1)$.)

XVII.4.10.3. Proposition. The sets $\{P_q : q \in \mathbb{Q}\}$ are pairwise disjoint, and $\cup_{q\in\mathbb{Q}} P_q = \mathbb{R}$.

Proof: If $x \in (P+q_1) \cap (P+q_2)$, then $x = p_1 + q_1 = p_2 + q_2$ for $p_1, p_2 \in P$. Then $p_2 - p_1 = q_1 - q_2 \in \mathbb{Q}$, so $p_1 R p_2$ and hence $p_1 = p_2$ since P contains only one element from any equivalence class. Thus also $q_1 = q_2$. If $x \in \mathbb{R}$, then xRp for some $p \in P$. Set $q = x - p \in \mathbb{Q}$; then $x = p + q \in P_q$. 👍

Now let $J = \mathbb{Q} \cap (-1,1)$, and set $N = \cup_{q\in J} P_q$.

XVII.4.10.4. Proposition. We have $(0,1) \subseteq N \subseteq (-1,2)$. Thus $1 \le \lambda^*(N) \le 3$.

Proof: Since $P \subseteq (0,1)$, it is obvious that $P_q \subseteq (-1,2)$ for any $q \in J$. For the other containment, let $x \in (0,1)$. Then $x = p + q$ for some $p \in P$, $q \in \mathbb{Q}$ by XVII.4.10.3.. Then $q = x - p \in (-1,1)$, so $x \in N$. 👍

XVII.4.10.5. Theorem. P is not Lebesgue measurable.

Proof: Suppose P is measurable. Then P_q is measurable for all q, and $\lambda(P_q) = \lambda(P)$; so N is measurable. If $\lambda(P) = 0$, then $\lambda(N) = \sum_{q\in J} \lambda(P_q) = 0$, contradicting XVII.4.10.4.. And if $\lambda(P) > 0$, then $\lambda(N) = \sum_{q\in J} \lambda(P_q) = \infty$ since J is infinite, again contradicting XVII.4.10.4.. 👍

In fact, we can do better:

XVII.4.10.6. PROPOSITION. If A is a Lebesgue measurable subset of P, then $\lambda(A) = 0$.

PROOF: Let $B = \cup_{q\in J}(A+q)$. The sets $A+q$ are pairwise disjoint since $A \subseteq P$. Since $\lambda(A+q) = \lambda(A)$ for all q, and J is infinite, we would have $\lambda(B) = \infty$ if $\lambda(A) > 0$. But $B \subseteq N \subseteq (-1, 2)$, so $\lambda(B) \leq 3$. Thus we must have $\lambda(A) = 0$ (and hence $\lambda(B) = 0$). 👍

XVII.4.10.7. THEOREM. (AC) Let $Y \subseteq \mathbb{R}$. If $\lambda^*(Y) > 0$, then Y contains a subset which is not Lebesgue measurable.

PROOF: For $q \in \mathbb{Q}$, let $Y_q = Y \cap P_q$. Then the Y_q are pairwise disjoint and $Y = \cup_{q\in\mathbb{Q}} Y_q$ by XVII.4.10.3.. If Y_q is measurable, then $\lambda(Y_q) = 0$ by XVII.4.10.6.. Thus, if all the Y_q are measurable, then $\lambda^*(Y) \leq \sum_{q\in\mathbb{Q}} \lambda^*(Y_q) = 0$. 👍

This result is only interesting if Y is Lebesgue measurable, since it is trivial otherwise. The important consequence is that every measurable set of positive measure contains a nonmeasurable subset.

XVII.4.10.8. Homeomorphisms of $\mathbb{R}$ do not necessarily preserve Lebesgue measurability (XVII.5.2.3.), i.e. Lebesgue measurability of subsets of $\mathbb{R}$ is not a topological property, although it is a metric property (in fact, see XIX.5.1.12..) Note that the property of being a Borel set is a topological property.

XVII.4.10.9. The Axiom of Choice was used in an essential way in the definition of P, so all results of this subsection depend on AC. In fact, R. SOLOVAY showed [] that (assuming that the existence of an inaccessible cardinal is consistent with ZFC; cf. ()) the existence of a nonmeasurable subset of $\mathbb{R}$ cannot be proved in ZF, i.e. it is consistent with ZF to assume that every subset of $\mathbb{R}$ is Lebesgue measurable. Specifically, he showed that if an inaccessible cardinal exists, then there is a model for ZF in which every subset of $\mathbb{R}$ is Lebesgue measurable; this model satisfies DC, but of course not AC. Thus even ZF+DC is not sufficient to prove the existence of a subset of $\mathbb{R}$ which is not Lebesgue measurable.

But the full Axiom of Choice is not needed to obtain a set which is not Lebesgue measurable. The argument only requires a choice function for the set of cosets of $\mathbb{Q}$ in $\mathbb{R}$, hence existence of a choice function for $\mathcal{P}(\mathbb{R})$ suffices, which will be automatic if $\mathbb{R}$ can be well ordered. Thus the Continuum Hypothesis is sufficient (but not necessary) to obtain Vitali's example (). For another example of a Lebesgue nonmeasurable set based on the CH, see XVIII.3.3.1..

Actually, assuming AC and some weak version of the CH, there is no finite continuous probability measure at all on $([0,1], \mathcal{P}([0,1]))$. See II.10.3. for a discussion. It is also known that ZF+(Hahn-Banach Theorem) suffices to prove existence of a Lebesgue nonmeasurable set [], and the Hahn-Banach Theorem is strictly weaker than the Axiom of Choice ().

Bernstein's Nonmeasurable Set

There is another construction of a nonmeasurable subset of $\mathbb{R}$, due to F. BERNSTEIN. This construction uses more machinery than VITALI'S, but works in great generality.

XVII.4.10.10. The general setting is an outer measure μ^* on a set X, for which there is a collection $\mathcal{C} = \{C_i : i \in I\}$ of subsets of X, with the following properties:

(i) $|I| \leq 2^{\aleph_0}$.

(ii) $|C_i| = 2^{\aleph_0}$ for all i.

(iii) $\mu^*(C_i) > 0$ for all i.

(iv) If A is any μ^*-measurable subset of X with $\mu^*(A) > 0$, then $C_i \subseteq A$ for some i.

Actually, $2^{\aleph_0}$ can be replaced by any other infinite cardinal here; the case of $2^{\aleph_0}$ is just the important one for applications.

XVII.4.10.11. Lebesgue outer measure on $\mathbb{R}$ has such a $\mathcal{C}$: the collection of closed subsets of $\mathbb{R}$ with positive Lebesgue measure. There are $2^{\aleph_0}$ closed subsets of $\mathbb{R}$ in all () (and $2^{\aleph_0}$ closed intervals, each of positive measure). Each closed set of positive measure is uncountable, hence has cardinality $2^{\aleph_0}$ by (). Condition (iv) is satisfied by XVII.4.4.5..

XVII.4.10.12. THEOREM. Let μ^* be an outer measure on a set X for which there is a $\mathcal{C}$ satisfying (i)–(iv). If Y is any subset of X with $\mu^*(Y) > 0$, there is a subset of Y which is not μ^*-measurable.

PROOF: Put a well ordering on I such that the set of predecessors of any $i \in I$ has cardinality strictly less than $2^{\aleph_0}$ (the Axiom of Choice guarantees that this can be done, by putting I in one-one correspondence with the cardinal $card(I)$). Inductively choose for each i two distinct elements x_i and y_i of C_i such that x_i and y_i are also distinct from the set $\{x_j, y_j : j < i\}$; this is possible since the cardinality of this set is strictly smaller than the cardinality of C_i. Let $A = \{x_i : i \in I\}$.

Let $Y \subseteq X$ with $\mu^*(Y) > 0$. To show that Y contains a subset which is not μ^*-measurable, we may assume that Y is μ^*-measurable (since otherwise Y itself is the desired subset). We claim $A \cap Y$ is not μ^*-measurable. If $A \cap Y$ (and hence also $Y \setminus A$) is μ^*-measurable, we cannot have $\mu^*(A \cap Y) > 0$, since then $C_i \subseteq (A \cap Y)$ for some i; but $y_i \in C_i \setminus A$, a contradiction. Thus $\mu^*(A \cap Y) = 0$. Similarly, we cannot have $\mu^*(Y \setminus A) > 0$, so $\mu^*(Y \setminus A) = 0$. But then $\mu^*(Y) = \mu^*(A \cap Y) + \mu^*(Y \setminus A) = 0$, contradicting the assumption that $\mu^*(Y) > 0$. 👍

Note that AC is used twice: well ordering I and choosing the x_i and y_i.

XVII.4.10.13. This result applies more generally to any purely nonatomic Borel regular Borel outer measure on a Polish space, whose restriction to the Borel sets is σ-finite: an infinite Polish space has $2^{\aleph_0}$ Borel sets (), and an uncountable Borel set in a Polish space has cardinality $2^{\aleph_0}$ (). Compare with XIII.13.9.14..

XVII.5. Measurable Functions

A crucial companion notion to measure spaces is measurable functions. Although they can be defined more generally (XVI.1.6.2.), we will restrict to real-valued or extended real-valued functions, which are by far the most important case.

XVII.5.1. Measurable Functions on a Measure Space

XVII.5.1.1. If $(X, \mathcal{A}, \mu)$ is a measure space, we say a real-valued function on X is *μ-measurable* if it is $\mathcal{A}$-measurable in the sense of XVI.2.2.2., i.e. $f^{-1}(B) \in \mathcal{A}$ for every Borel set B in $\mathbb{R}$. This definition has nothing specifically to do with μ, but only the σ-algebra $\mathcal{A}$ (which recall is part of the definition of μ). There is a useful slight variation:

XVII.5.1.2. Definition. Let μ^* be an outer measure on a set X, and $\mathcal{A}$ the σ-algebra of μ^*-measurable subsets of X. A real-valued function is *μ^*-measurable* if it is $\mathcal{A}$-measurable in the sense of XVI.2.2.2..

XVII.5.1.3. Proposition XVI.2.2.4. can be rephrased in this context, using XVII.4.3.2.: f is μ^*-measurable if and only if, for every $Y \subseteq X$ with $\mu^*(Y) < \infty$, and every $r \in \mathbb{R}$,

$$\mu^*(Y) \geq \mu^*(Y \cap \{x \in X : f(x) \leq r\}) + \mu^*(Y \cap \{x \in X : f(x) > r\}) \ .$$

A similar statement holds for the other criteria in XVI.2.2.4..

There is a useful variation (cf. [Fed69, 2.3.2(7)]):

XVII.5.1.4. Proposition. Let μ^* be an outer measure on a set X, and $f : X \to \mathbb{R}$. The following are equivalent:

(i) f is μ^*-measurable.

(ii) For every $Y \subseteq X$, $\mu^*(Y) < \infty$, and $a, b \in \mathbb{R}$, $a < b$,

$$\mu^*(Y) \geq \mu^*(Y \cap \{x \in X : f(x) \leq a\}) + \mu^*(Y \cap \{x \in X : f(x) \geq b\}) \ .$$

(iii) There is a dense subset D of $\mathbb{R}$ such that, for every $Y \subseteq X$, $\mu^*(Y) < \infty$, and $a, b \in D$, $a < b$,

$$\mu^*(Y) \geq \mu^*(Y \cap \{x \in X : f(x) \leq a\}) + \mu^*(Y \cap \{x \in X : f(x) \geq b\}) \ .$$

In (ii) and (iii), the inequalities can be replaced by strict inequalities.

Proof: (i) $\Rightarrow$ (ii) by XVI.2.2.4. and monotonicity, and (ii) $\Rightarrow$ (iii) is trivial.

(iii) $\Rightarrow$ (i): Fix D, Y, and $r \in \mathbb{R}$, and set $A = \{x \in X : f(x) \leq r\}$. For each $k \in \mathbb{N}$, set

$$B_k = Y \cap \left\{ x \in X : r + \frac{1}{k+1} \leq f(x) \leq r + \frac{1}{k} \right\} \ .$$

We claim that if K is any finite subset of $\mathbb{N}$ consisting entirely of even numbers or entirely of odd numbers, then

$$\mu^* \left(\bigcup_{k \in K} B_k \right) \geq \sum_{k \in K} \mu^*(B_k) \ .$$

The proof is by induction on $|K|$. If $|K| = 1$ the statement is trivial. Now suppose $|K| > 1$ and j is the largest number in K. Set $C_K = \cup_{k\in K} B_k$, $u = \sup\{f(x) : x \in B_j\}$, and $v = \inf\{f(x) : x \in \cup_{k\in K\setminus\{j\}} B_k\}$. Then $u < v$. Let $a, b \in D$, $u < a < b < v$. Then

$$\mu^*(C_K) \geq \mu^*(C_K \cap \{x \in X : f(x) \leq a\}) + \mu^*(C_K \cap \{x \in X : f(x) \geq b\})$$

$$\geq \mu^*(B_j) + \mu^*(\cup_{k\in K\setminus\{j\}} B_k) \geq \mu^*(B_j) + \sum_{k\in K\setminus\{j\}} \mu^*(B_k) \ .$$

Thus we have

$$\sum_{k=1}^{\infty} \mu^*(B_k) \leq \mu^*\left(\bigcup_{k\ odd} B_k\right) + \mu^*\left(\bigcup_{k\ even} B_k\right) \leq 2\mu^*(Y) < \infty$$

so, if $\epsilon > 0$, there is an n such that

$$\epsilon > \sum_{k=n}^{\infty} \mu^*(B_k) \geq \mu^*\left(\bigcup_{k=n}^{\infty} B_k\right) = \mu^*\left(Y \cap \left\{x \in X : r < f(x) \leq r + \frac{1}{n}\right\}\right)$$

so

$$\mu^*(Y \setminus A) \leq \mu^*\left(Y \cap \left\{x \in X : r < f(x) \leq r + \frac{1}{n}\right\}\right) + \mu^*\left(Y \cap \left\{x \in X : f(x) \geq r + \frac{1}{n}\right\}\right)$$

$$< \mu^*\left(Y \cap \left\{x \in X : f(x) \geq r + \frac{1}{n}\right\}\right) + \epsilon \ .$$

If $c, d \in D$, $r < c < d < r + \frac{1}{n}$,

$$\mu^*(Y \cap A) + \mu^*(Y \setminus A) - \epsilon < \mu^*(Y \cap A) + \mu^*\left(Y \cap \left\{x \in X : r < f(x) \leq r + \frac{1}{n}\right\}\right)$$

$$\leq \mu^*(Y \cap \{x \in X : f(x) \leq c\}) + \mu^*(Y \cap \{x \in X : f(x) \geq d\}) \leq \mu^*(Y) \ .$$

Since $\epsilon > 0$ is arbitrary, the result follows. 👍

XVII.5.1.5. If the outer measure μ^* is monotone regular (XVII.4.5.30.), the proof simplifies greatly, since then for any ϵ there is automatically an n with

$$\mu^*(Y \setminus A) < \mu^*\left(Y \cap \left\{x \in X : f(x) \geq r + \frac{1}{n}\right\}\right) + \epsilon \ .$$

XVII.5.2. Lebesgue Measurable Functions

The most important class of functions from $\mathbb{R}$ to $\mathbb{R}$ for advanced analysis purposes is the class of Lebesgue measurable functions:

XVII.5.2.1. Definition. A function $f : \mathbb{R} \to \mathbb{R}$ is *Lebesgue measurable* if it is λ^*-measurable in the sense of XVII.5.1.2., i.e. if $f^{-1}(B)$ is a Lebesgue measurable set for every Borel set $B \subseteq \mathbb{R}$.

XVII.5.2.2. There is a subtlety in this definition. Note that the criterion for a function to be Lebesgue measurable is that the preimage of every *Borel* set be Lebesgue measurable, not that the preimage of every Lebesgue measurable set is Lebesgue measurable; in fact the preimage of a Lebesgue measurable set under a Lebesgue measurable function is not necessarily Lebesgue measurable (XVII.5.2.3.). In other words, a Lebesgue measurable function from $\mathbb{R}$ to $\mathbb{R}$ is an $(\mathcal{L}, \mathcal{B})$-measurable function in the sense of XVI.1.6.2., not necessarily an $(\mathcal{L}, \mathcal{L})$-measurable function. This may seem like a curious definition, but experience has shown that it is the most useful and appropriate one; too small a class of functions is $(\mathcal{L}, \mathcal{L})$-measurable (not even every strictly increasing continuous function is $(\mathcal{L}, \mathcal{L})$-measurable! See XVII.5.2.3..) The definition is much more natural if it is regarded as a special case of XVI.2.2.2..

XVII.5.2.3. Proposition. (AC) The preimage of a Lebesgue measurable set under a Lebesgue measurable function is not necessarily Lebesgue measurable. In fact, there is a strictly increasing continuous function (homeomorphism) $f : \mathbb{R} \to \mathbb{R}$ and a measurable subset $A \subseteq \mathbb{R}$ such that $f^{-1}(A)$ is not Lebesgue measurable.

Proof: Let K be the Cantor set, and K_α the generalized Cantor set of positive measure from (). There is a strictly increasing continuous function $f : \mathbb{R} \to \mathbb{R}$ with $f(K_\alpha) = K$. By XVII.4.10.7., there is a nonmeasurable subset B of K_α. Let $A = f(B)$; since $A \subseteq K$ and $\lambda(K) = 0$, A is measurable and $\lambda(A) = 0$. But $B = f^{-1}(A)$ is nonmeasurable.

A Criterion for Lebesgue Measurability

We have the following special case of XVI.2.2.4.:

XVII.5.2.4. Proposition. Let $f : \mathbb{R} \to \mathbb{R}$ a function. Then f is Lebesgue measurable if and only if any of the following conditions are satisfied:

(i) $f^{-1}(U) \in \mathcal{L}$ for every open set $U \subseteq \mathbb{R}$.

(ii) $f^{-1}(E) \in \mathcal{L}$ for every closed set $E \subseteq \mathbb{R}$.

(iii) $f^{-1}(U) \in \mathcal{L}$ for every open interval $U \subseteq \mathbb{R}$.

(iv) $f^{-1}((-\infty, a)) \in \mathcal{L}$ for every $a \in \mathbb{R}$.

(v) $f^{-1}((-\infty, a]) \in \mathcal{L}$ for every $a \in \mathbb{R}$.

(vi) $f^{-1}((a, +\infty)) \in \mathcal{L}$ for every $a \in \mathbb{R}$.

(vii) $f^{-1}([a, +\infty)) \in \mathcal{L}$ for every $a \in \mathbb{R}$.

(viii) $f^{-1}((-\infty, a)) \in \mathcal{L}$ for every $a \in \mathbb{Q}$.

(ix) $f^{-1}((-\infty, a]) \in \mathcal{L}$ for every $a \in \mathbb{Q}$.

(x) $f^{-1}((a, +\infty)) \in \mathcal{L}$ for every $a \in \mathbb{Q}$.

(xi) $f^{-1}([a, +\infty)) \in \mathcal{L}$ for every $a \in \mathbb{Q}$.

This proposition is one of the reasons why Lebesgue measurability is defined as it was.

XVII.5.2.5. Corollary. Every continuous function from $\mathbb{R}$ to $\mathbb{R}$ is Lebesgue measurable (in fact, Borel measurable).

Pointwise Limits

Because $(\mathbb{R}, \mathcal{L}, \lambda)$ is complete, we have:

XVII.5.2.6. Proposition. Let $f, g : \mathbb{R} \to \mathbb{R}$. If $f = g$ a.e. and f is Lebesgue measurable, so is g.

Proof: Let $C = \{t \in \mathbb{R} : f(t) \neq g(t)\}$. Then by assumption C is a subset of a λ-null set, hence λ-null by completeness. Let $a \in \mathbb{R}$. Then $A := \{t : g(t) < a\}$ differs from $\{t : f(t) < a\} \in \mathcal{L}$ by a subset of C, which is in $\mathcal{L}$ by completeness, so $A \in \mathcal{L}$. 👍

We can then give a generalization of XVI.2.2.9.:

XVII.5.2.7. Proposition. Let (f_n) be a sequence of Lebesgue measurable functions, with $f_n \to f$ pointwise a.e. Then f is Lebesgue measurable.

Proof: Let $C = \{t : f_n(t) \not\to f(t)\}$. By assumption, C is λ-null. For each n, set

$$g_n(t) = \begin{cases} f_n(t) & \text{if} \quad t \notin C \\ 0 & \text{if} \quad t \in C \end{cases}.$$

Then $g_n = f_n$ a.e., so is Lebesgue measurable, and $g_n \to g$ pointwise, where

$$g(t) = \begin{cases} f(t) & \text{if} \quad t \notin C \\ 0 & \text{if} \quad t \in C \end{cases}.$$

By XVI.2.2.9., g is Lebesgue measurable, and $f = g$ a.e., so f is Lebesgue measurable by XVII.5.2.6.. 👍

Other Properties

XVII.5.2.8. We summarize the other important properties of Lebesgue measurable functions:

A sum or product of Lebesgue measurable functions is Lebesgue measurable (XVI.2.4.11.).

The supremum or infimum of a pointwise bounded sequence of Lebesgue measurable functions is Lebesgue measurable (XVI.2.2.7.).

If f is Lebesgue measurable and g is Borel measurable (e.g. continuous), then $g \circ f$ is Lebesgue measurable (XVI.2.4.8.). [**Caution:** $f \circ g$ is not necessarily Lebesgue measurable!] In particular, if f is Lebesgue measurable, so are f_+, f_-, and $|f|$. If $f \geq 0$, then $\sqrt{f}$ is also Lebesgue measurable. ✋

Lebesgue Measurability vs. Borel Measurability

XVII.5.2.9. Since $\mathcal{B} \subseteq \mathcal{L}$, every Borel measurable function () is Lebesgue measurable. In particular, every continuous function is Lebesgue measurable.

Of course, not every Lebesgue measurable function is Borel measurable (e.g. χ_A for $A \in \mathcal{L} \setminus \mathcal{B}$). However:

XVII.5.2.10. PROPOSITION. Let f be a Lebesgue measurable function. Then there is a Borel measurable function g with $f = g$ a.e. If f is bounded below [resp. bounded above], we may additionally require that $g \leq f$ [resp. $g \geq f$].

PROOF: First suppose $f = \sum_{k=1}^{n} \alpha_k \chi_{A_k}$ is a simple function. By adding a constant, we may assume f is nonnegative. For each k there is an F_σ $E_k \subseteq A_k$ with $\lambda(A_k \setminus E_k) = 0$ (). Then $g = \sum_{k=1}^{n} \alpha_k \chi_{E_k}$ is a nonnegative Borel measurable simple function with $g \leq f$ and $g = f$ a.e.

Now suppose f is nonnegative. Let (f_n) be a sequence of simple functions with $0 \leq f_n \leq f$ for all n and $f_n \to f$ pointwise, so $f = \sup f_n$ (). Let g_n be a nonnegative Borel measurable simple function with $g_n \leq f_n$ and $g_n = f_n$ on the complement of a null set C_n. Set $g = \sup g_n$. Then g is Borel measurable (), $g \leq f$ since $g_n \leq f_n \leq f$ for all n, and $g = f$ on the complement of the null set $C = \cup_n C_n$.

If f is bounded below, then by adding a constant f is nonnegative, so the previous argument proves the last statement; for the parenthetical part consider $-f$.

Finally, for a general f let f_+ and f_- be the positive and negative parts of f. Then f_+ and f_- are Lebesgue measurable (). Let g_+ and g_- be Borel measurable functions equal to f_+ and f_- a.e. respectively. Then $g = g_+ - g_-$ is Borel measurable and equal to $f = f_+ - f_-$ a.e. 👍

It is unclear whether the inequalities can be satisfied in general if f is unbounded (they can be easily satisfied if g is allowed to take the values $\pm\infty$).

See XVIII.6.7.3. for an alternate proof using the Radon-Nikodym Theorem.

XVII.5.3. Exercises

XVII.5.3.1. (a) There is a sequence (S_i) of pairwise disjoint Borel subsets of $[1, \infty)$, each of infinite Lebesgue measure. [Let (N_i) be a sequence of pairwise disjoint infinite subsets of $\mathbb{N}$, e.g. by identifying $\mathbb{N}$ with $\mathbb{N} \times \mathbb{N}$. Let $S_i = \cup_{n \in N_i} [n, n+1)$.]

(b) Let $\mathcal{F}$ be the set of finite subsets of $\mathbb{N}$. $\mathcal{F}$ is countable; let i be a bijection from $\mathcal{F}$ to $\mathbb{N}$.

(c) Arrange the bounded open intervals in $\mathbb{R}$ with rational endpoints into a sequence (J_n), with each interval appearing infinitely often in the sequence.

(d) For each n, let

$$A_n = J_n \cup \left[\bigcup_{n \in F} S_{i(F)} \cap [n, \infty) \right]$$

(i.e. the union runs over all sets $F \in \mathcal{F}$ with $n \in F$) and let $\mathcal{S}$ be the set of finite intersections of the A_n and their complements. Then $\mathcal{S}$ is a semialgebra (). Show that every nonempty set in $\mathcal{S}$ has infinite Lebesgue measure. [If $E = A_{n_1} \cap \cdots \cap A_{n_k} \cap A_{m_1}^c \cap \cdots \cap A_{m_r}^c$, if $n_i = m_j$ for some i, j, then $E = \emptyset$, and otherwise $S_{i(\{n_1, \ldots, n_k\})} \cap [p, \infty) \subseteq E$, where $p = \max\{n_1, \ldots, n_k\}$. Note that $S_i \cap [p, \infty)$ has infinite Lebesgue measure

for every i and p.] If $\mathcal{A}$ is the algebra generated by $\mathcal{S}$, then $\mathcal{A}$ consists of finite disjoint unions of sets in $\mathcal{S}$ (), and hence every nonempty set in $\mathcal{A}$ has infinite Lebesgue measure.

(e) If J is a bounded open interval with rational endpoints, then there is an infinite sequence (n_k) such that $J_{n_k} = J$ for all k. Then $\cap_{k=1}^{\infty} A_{n_k} = J$. Thus J is in the σ-algebra generated by $\mathcal{S}$ (or $\mathcal{A}$). Since such intervals generate $\mathcal{B}$ as a σ-algebra, the σ-algebra generated by $\mathcal{S}$ (or $\mathcal{A}$) is $\mathcal{B}$.

(f) The restriction of λ to $\mathcal{S}$ (or $\mathcal{A}$) is not σ-finite, and λ is not the measure induced on $\mathcal{B}$ from $\lambda|_{\mathcal{S}}$ or $\lambda|_{\mathcal{A}}$ by the Carathéodory extension process.

(g) If $(X, \mathcal{B}, \mu)$ is a measure space with $\mu(X) = \infty$, is there always an algebra $\mathcal{A}$ in $\mathcal{B}$ which generates $\mathcal{B}$ as a σ-algebra, with the property that $\mu(A) = \infty$ for every nonempty $A \in \mathcal{A}$?

XVII.5.3.2. Let $X = \{x_1, x_2, x_3\}$ be a set with three elements. If $A \subseteq X$, let $\mu^*(A) = 0$ if $A = \emptyset$, $\mu^*(A) = 1$ if $|A| = 1$ or if $|A| = 2$, and $\mu^*(A) = 2$ if $|A| = 3$.

(a) Show that μ^* is an outer measure on X which is not regular.

(b) If $A = \{x_1, x_2\}$, show that $\mu^*(A) + \mu^*(X \setminus A) = \mu^*(X)$, but A is not μ^*-measurable. [Consider $Y = \{x_1, x_3\}$.]

XVII.5.3.3. Prove the following generalization of (). Let μ^* be an outer measure on a set X, and $\mathcal{M}$ the σ-algebra of μ^*-measurable sets.

(a) If (A_n) is a pairwise disjoint sequence of sets in $\mathcal{M}$ with $A = \cup_{n=1}^{\infty} A_n$, and Y is any subset of X, then $\mu^*(Y \cap A) = \sum_{n=1}^{\infty} \mu^*(Y \cap A_n)$.

(b) If $B_1 \subseteq B_2 \subseteq \cdots$ is an increasing sequence in $\mathcal{M}$ with $B = \cup_{n=1}^{\infty} B_n$, and Y is any subset of X, then $\mu^*(Y \cap B) = \sup_n \mu^*(Y \cap B_n)$.

(c) If $C_1 \supseteq C_2 \supseteq \cdots$ is a decreasing sequence in $\mathcal{M}$ with $C = \cap_{n=1}^{\infty} C_n$, and Y is any subset of X with $\mu^*(Y \cap C_1) < \infty$, then $\mu^*(Y \cap C) = \inf_n \mu^*(Y \cap C_n)$.

XVII.5.3.4. If μ^* is an outer measure on X and $A, Y \subseteq X$ with A μ^*-measurable, show that

$$\mu^*(A \cup Y) + \mu^*(A \cap Y) = \mu^*(A) + \mu^*(Y) \ .$$

XVII.5.3.5. Let μ^* be a regular outer measure on a set X. If $A, B \subseteq X$ with $A \cup B$ μ^*-measurable and $\mu^*(A \cup B) = \mu^*(A) + \mu^*(B) < \infty$, show that A and B are μ^*-measurable.

XVII.5.3.6. Let X and Y be sets, $f : X \to Y$ a function, and μ^* an outer measure on X. For $B \subseteq Y$, set $\nu^*(B) = \mu^*(f^{-1}(B))$. Show that ν^* is an outer measure on Y. This outer measure is usually denoted $f_*(\mu^*)$.

XVII.5.3.7. Let $\{\mu_i^*\}$ be a collection of outer measures on a set X, and let $\mu^* = \sup \mu_i^*$. Show that μ^* is an outer measure.

XVII.5.3.8. Let $X = (0,2]$, $\mathcal{A} = \{\emptyset, (0,1], (1,2], (0,2]\}$. Let μ be the restriction of Lebesgue measure to $\mathcal{A}$, and μ^* the induced outer measure on X.

(a) Describe $\mu^*(Y)$ for a general $Y \subseteq X$, and compare it to $\lambda^*(Y)$. Find the σ-algebra of μ^*-measurable subsets of X.

(b) Show that $(X, \mathcal{A}, \mu)$ is a finite complete measure space, but that μ extends to a measure (e.g. λ) on a strictly larger σ-algebra. Show that λ^* is not $\mathcal{A}$-regular.

XVII.5.3.9. Let $\mathcal{U}$ be the collection of open sets in $\mathbb{R}^n$, and let $\mathcal{Q}$ be the set of bounded open intervals with rational endpoints. Show that $\mathcal{S}(\mathcal{U}) = \mathcal{S}(\mathcal{Q})$ and that this collection of sets is the same as $\mathcal{S}(\mathcal{K})$, $\mathcal{S}(\mathcal{F})$, and $\mathcal{S}(\mathcal{B})$ of XVII.4.6.6.. These sets are usually just called the *Suslin subsets* of $\mathbb{R}^n$.

XVII.5.3.10. Show that the collection $\mathcal{S}(\mathcal{Q})$ of Suslin subsets of $\mathbb{R}^n$ (Exercise XVII.5.3.9.) has cardinality $2^{\aleph_0}$. [Show that there are exactly $2^{\aleph_0}$ Suslin schemes in $\mathcal{Q}$.] Since $\mathcal{B} \subseteq \mathcal{S}(\mathcal{Q})$, this gives a cleaner proof than () that $\mathcal{B}$ has cardinality $2^{\aleph_0}$.

XVII.5.3.11. Let $\mathcal{E}$ be a nonempty collection of subsets of a set X.

(a) If $\mathcal{A} = \mathcal{S}(\mathcal{E}) \cap \widetilde{\mathcal{S}(\mathcal{E})}$ is nonempty, it is a σ-algebra and the largest σ-algebra contained in $\mathcal{S}(\mathcal{E})$ (cf. XVI.1.7.2.).

(b) Show that $\mathcal{S}(\mathcal{E}) \cap \widetilde{\mathcal{S}(\mathcal{E})}$ can be empty.

XVII.5.3.12. Let μ be a finite local measure on a σ-ring $\mathcal{D}$ of subsets of X, and let $\mathcal{A}$ be the σ-algebra generated by $\mathcal{D}$.

(a) Show that μ is bounded on $\mathcal{D}$.

(b) Let $M = \sup\{\mu(A) : A \in \mathcal{D}\}$. For each n let $A_n \in \mathcal{D}$ with $\mu(A_n) \geq M - \frac{1}{n}$, and let $Y = \cup_n A_n \in \mathcal{D}$. Show that $\mu(Y) = M$.

(c) For $A \in \mathcal{A}$, define $\nu(A) = \mu(A \cap Y)$. Show that ν is a (finite) measure on $(X, \mathcal{A})$ extending μ.

In probability, $\mathbf{P}$ is usually used for a measure instead of μ. Thus $\mathbf{P}(E)$ is the probability (measure) of E, and $\mathbf{E}(\mathbf{X}) = \int \mathbf{X}\, d\mathbf{P}$ is the expectation (integral) of $\mathbf{X}$.

XVII.6. Measures on Metric Spaces

Many important outer measures (e.g. Lebesgue outer measure) are defined on metric spaces, and behave nicely relative to the topology of the space.

XVII.6.1. Metric Outer Measures and Borel Measures

XVII.6.1.1. DEFINITION. Let (X, ρ) be a metric space, and let μ^* be an outer measure on X. The outer measure μ^* is a *metric outer measure* on X (with respect to ρ) if, whenever $A, B \subseteq X$ and $\rho(A, B) > 0$ (), we have $\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)$.

XVII.6.1.2. EXAMPLE. By Exercise (), Lebesgue outer measure is a metric outer measure on $\mathbb{R}$ with the usual metric.

XVII.6.1.3. PROPOSITION. Let (X, ρ) be a metric space, and μ^* an outer measure on X for which every closed set is μ^*-measurable. Then μ^* is a metric outer measure.

PROOF: If $A, B \subseteq X$ with $\rho(A, B) > 0$, then $B \cap \bar{A} = \emptyset$. Since $\bar{A}$ is closed, we have

$$\mu^*(A \cup B) = \mu^*((A \cup B) \cap \bar{A}) + \mu^*((A \cup B) \cap (\bar{A})^c) = \mu^*(A) + \mu^*(B) \ .$$

The main theorem of this subsection is the converse:

XVII.6.1.4. THEOREM. Let (X, ρ) be a metric space, and μ^* a metric outer measure on X with respect to ρ. Then every Borel set in X is μ^*-measurable.

PROOF: Since the collection of μ^*-measurable sets is a σ-algebra, it suffices to show that every closed set F in X is μ^*-measurable. Fix $E \subseteq X$ with $\mu^*(E) < \infty$, and let $A = E \cap F$, $B = E \cap F^c$. It suffices to show that $\mu^*(E) \geq \mu^*(A) + \mu^*(B)$.

For $n \in \mathbb{N}$, set

$$B_n = \left\{ x \in B : \rho(x, F) \geq \frac{1}{n} \right\} \ .$$

We have $B = \cup_{n=1}^{\infty} B_n$ since F is closed; and we have $\mu^*(A \cup B_n) = \mu^*(A) + \mu^*(B_n)$ for each n since μ^* is a metric outer measure.

If μ^* were known to be monotone regular, the proof would be trivial:

$$\mu^*(E) = \mu^*(A \cup B) = \sup_n \mu^*(A \cup B_n) = \sup_n [\mu^*(A) + \mu^*(B_n)] = \mu^*(A) + \sup_n \mu^*(B_n) = \mu^*(A) + \mu^*(B) \ .$$

In general, we must show that μ^* has enough vestiges of monotone regularity to get the conclusion.

For each n let $D_n = B_{n+1} \setminus B_n$. We have $\rho(D_{n+1}, B_n) \geq \frac{1}{n} - \frac{1}{n+1} = \frac{1}{n(n+1)}$ by the triangle inequality, and thus $\mu^*(B_n \cup D_{n+1}) = \mu^*(B_n) + \mu^*(D_{n+1})$ for all n; since $B_n \cup D_{n+1} \subseteq B_{n+2}$, we have $\mu^*(B_n) + \mu^*(D_{n+1}) \leq \mu^*(B_{n+2})$ for all n (this also holds for $n = 0$ if we set $B_0 = \emptyset$). Inductively, we get

$$\sum_{k=1}^{n} \mu^*(D_{2k-1}) \leq \mu^*(B_{2n}) \leq \mu^*(E) < \infty$$

for all n. Similarly,

$$\sum_{k=1}^{n} \mu^*(D_{2k}) \leq \mu^*(B_{2n+1}) \leq \mu^*(E) < \infty$$

for all n. Thus $\sum_{k=1}^{\infty} \mu^*(D_k)$ converges; hence, for any $\epsilon > 0$ there is an n such that $\sum_{k=n+1}^{\infty} \mu^*(D_k) < \epsilon$. By countable subadditivity we then have for this n that

$$\mu^*(B) \leq \mu^*(B_n) + \sum_{k=n+1}^{\infty} \mu^*(D_k) < \mu^*(B_n) + \epsilon \ .$$

Thus $\mu^*(B) = \sup_n \mu^*(B_n)$. We then have

$$\mu^*(E) \geq \sup_n \mu^*(A \cup B_n) = \sup_n [\mu^*(A) + \mu^*(B_n)] = \mu^*(A) + \sup_n \mu^*(B_n) = \mu^*(A) + \mu^*(B) \ .$$



Borel Measures

XVII.6.1.5. DEFINITION. Let X be a topological space. A *Borel measure* on X is a measure on $(X, \mathcal{B})$, where $\mathcal{B}$ is the σ-algebra of Borel subsets of X.

XVII.6.1.6. Suppose X is a metrizable topological space, and ρ is a metric on X giving the topology. If μ^* is an outer measure on X which is a metric outer measure with respect to ρ, then μ^* defines a Borel measure on X by XVII.6.1.4.. Conversely, a Borel measure μ on X defines a (regular) outer measure μ^* on X by

$$\mu^*(Y) = \inf\{\mu(B) : B \text{ Borel}, Y \subseteq B\} = \min\{\mu(B) : B \text{ Borel}, Y \subseteq B\}$$

for $Y \subseteq X$ (this makes no reference to ρ, and works whether or not X is metrizable) which is a metric outer measure for ρ by XVII.6.1.3.. Thus every Borel measure on a metric space arises this way from a (regular) metric outer measure.

Borel measures on metric spaces are not regular in general (XVII.6.1.13.). But they are under a mild hypothesis:

XVII.6.1.7. DEFINITION. Let X be a topological space and μ a Borel measure on X. Then μ is *locally finite* if every (Borel) set in X of finite measure is contained in an open set of finite measure.

XVII.6.1.8. Some references define a locally finite measure to be a (Borel) measure for which every point has a neighborhood of finite measure. We will call such a measure *neighborhood-finite*. In good cases the notions are the same (XVII.6.1.14.). Our definition turns out to be more useful.

XVII.6.1.9. EXAMPLES. (i) Any finite Borel measure is locally finite.

(ii) Lebesgue measure on $\mathbb{R}^n$ is locally finite.

(iii) Counting measure on $\mathbb{Q}$ with its usual metric is not locally finite (every nonempty open set has infinite measure). So a σ-finite Borel measure on a separable metric space need not be locally finite.

(iv) Let $X = \{p\}$ and μ the measure on X with $\mu(X) = \infty$. Then μ is locally finite but not neighborhood-finite: $\{p\}$ is an infinite atom.

(v) Let $\mathcal{H}$ be a nonseparable Hilbert space. Let $\{\xi_i : i \in I\}$ be an orthonormal basis for $\mathcal{H}$ and $Y = \{3\xi_i : i \in I\}$. Then Y is a closed subset of $\mathcal{H}$, and any two points of Y are distance $3\sqrt{2} > 4$ apart. Define a Borel measure on $\mathcal{H}$ by setting $\mu(A) = 0$ if A is a (Borel) set which is the union of a subset of Y and a set which can be covered by countably many open balls of radius 1, and $\mu(A) = \infty$ otherwise. It is easily checked that μ is a measure, and it is neighborhood-finite since every open ball of radius ≤ 1 has measure 0. Y is a Borel set of finite measure; but if U is an open set of finite measure containing Y, then $U = A \cup V$, where $A \subseteq Y$ and V can be covered by a sequence (B_n) of open balls of radius 1. We have that $W := U \setminus Y \subseteq V$, and since the interior of Y is empty $Y \subseteq \bar{W} \subseteq \bar{V}$. For each $i \in I$ let $n(i)$ be the first n such that the distance from $3\xi_i$ to B_n is less than 1. If $i \neq j$, $n(i) \neq n(j)$ since $\|3\xi_i - 3\xi_j\| > 4$, so $i \mapsto n(i)$ is an injective map from I to $\mathbb{N}$, contradicting that I is uncountable. Thus μ is not locally finite.

A locally finite Borel measure on a reasonable space is regular in the following sense:

XVII.6.1.10. THEOREM. Let X be a topological space in which every open set is an F_σ (e.g. a metrizable space), μ a locally finite Borel measure on X, and $\epsilon > 0$. Then for every Borel set A in X with $\mu(A) < \infty$, there is a closed set $E \subseteq A$ and an open set U containing A with $\mu(U \setminus E) < \epsilon$.

PROOF: First assume μ is finite. Let $\mathcal{A}$ be the collection of all subsets of X for which the conclusion holds for every $\epsilon > 0$. If V is an open set and $\epsilon > 0$, there is an increasing sequence (F_n) of closed sets with $\cup_n F_n = V$, so $\mu(U) = \sup_n \mu(F_n)$. So if $\mu(V) < \infty$, there is an n with $\mu(F_n) > \mu(V) - \epsilon$. Thus, taking $U = V$ and $E = F_n$ for sufficiently large n, we have $V \in \mathcal{A}$.

We now show that $\mathcal{A}$ is a σ-algebra. $\mathcal{A}$ is obviously closed under complements. Every clopen set Z is in $\mathcal{A}$ (with $E = U = Z$), and in particular $X, \emptyset \in \mathcal{A}$. If (A_n) is a sequence in $\mathcal{A}$ and $A = \cup_n \mathcal{A}_n$, for each k there are F_k and U_k with $F_k \subseteq A_k \subseteq U_k$, F_k closed, U_k open, and $\mu(U_k \setminus F_k) < 2^{-k}\epsilon$. Set $U = \cup_k U_k$ and, for each m,

$$E_m = \cup_{k=1}^m F_k \ .$$

We have $E_m \subseteq A \subseteq U$ for all m, U is open, E_m is closed, and

$$\mu\left(U \setminus \bigcup_{k=1}^{\infty} F_k\right) = \mu\left(\bigcup_{k=1}^{\infty} [U_k \setminus F_k]\right) \leq \sum_{k=1}^{\infty} \mu(U_k \setminus F_k) < \epsilon$$

and, since

$$U \setminus \bigcup_{k=1}^{\infty} F_k = \bigcap_{m=1}^{\infty} U \setminus E_m$$

we have $\mu(U \setminus E_m) < \epsilon$ for some sufficiently large m.

Thus all Borel sets are in $\mathcal{A}$, and the result is proved if μ is finite.

Now suppose μ is locally finite and B is a Borel set in X with $\mu(B) < \infty$. There is an open set W with $B \subseteq W$ and $\mu(W) < \infty$. Replacing X by W and μ by $\mu|_W$, if $\epsilon > 0$ there is a relatively open set U and a relatively closed set E in W with $E \subseteq B \subseteq U$ and $\mu(U \setminus E) < \epsilon$. Then U is open in X. W is an F_σ in X, so there is an increasing sequence (F_n) of closed sets in X with $W = \cup_n F_n$. Set $E_n = E \cap F_n$. We have $E = G \cap W$ for some closed set G in X; then

$$E_n = (E \cap F_n) = (G \cap W) \cap F_n = G \cap F_n$$

since $F_n \subseteq W$; thus each E_n is closed in X, and $E = \cup_n E_n$. And we have

$$U \setminus E = \bigcap_{n=1}^{\infty} E \setminus E_n$$

so we have

$$\mu(U \setminus E_n) < \epsilon$$

for some sufficiently large n. 👍

XVII.6.1.11. Conversely, a Borel measure which is regular in this sense is obviously locally finite (the U must have finite measure).

There is a consequence (of the finite part) for general Borel measures (cf. [Fed69, 2.2.2]):

XVII.6.1.12. Corollary. Let μ be a Borel measure on a metric space X and A a Borel set in X.

(i) If $\mu(A) < \infty$, then for any $\epsilon > 0$ there is a closed set $E \subseteq A$ with $\mu(A \setminus E) < \epsilon$.

(ii) If A can be covered by countably many open sets of finite measure, then for any $\epsilon > 0$ there is an open set U with $A \subseteq U$ and $\mu(U \setminus A) < \epsilon$.

In particular, if X can be covered by countably many open sets of finite measure, then μ is locally finite.

Proof: (i): Define a new Borel measure ν on X by $\nu(B) = \mu(A \cap B)$ for Borel sets B. Then ν is a finite Borel measure. Apply XVII.6.1.10. to ν to get E (and U). Then

$$\mu(A \setminus E) = \nu(A \setminus E) \leq \nu(U \setminus E) < \epsilon \ .$$

(ii): Let (V_n) be a sequence of open sets covering A with $\mu(V_n) < \infty$ for all n. By (i), for each n there is a closed set $E_n \subseteq V_n \setminus A$ with $\mu([V_n \setminus A] \setminus E_n) < 2^{-n}\epsilon$. Set $U = \cup_n (V_n \setminus E_n)$. Then U is open and, since $A \cap V_n \subseteq V_n \setminus E_n$, $A \subseteq U$. We have

$$U \setminus A = \bigcup_{n=1}^{\infty} [V_n \setminus E_n] \setminus A = \bigcup_{n=1}^{\infty} [V_n \setminus A] \setminus E_n$$

so

$$\mu(U \setminus A) \leq \sum_{n=1}^{\infty} \mu([V_n \setminus A] \setminus E_n) < \sum_{n=1}^{\infty} 2^{-n}\epsilon = \epsilon \ .$$

XVII.6.1.13. EXAMPLE. Here is an interesting example due to DIEUDONNÉ [**?**] which shows the result can fail even for a finite Borel measure on a compact Hausdorff space, if open sets are not always F_σ's.

Let X be the set of ordinals less than or equal to the first uncountable ordinal ω_1, and let $X_0 = X \setminus \{\omega_1\}$. Then X is a compact Hausdorff space in the order topology, and X_0 is a dense open set. Let $\mathcal{F}_0$ be the set of all uncountable (relatively) closed subsets of X_0. Define a measure μ on the Borel sets of X by

$$\mu(B) = \begin{cases} 1 & \text{if} \quad F \subseteq A \text{ for some } F \in \mathcal{F}_0 \\ 0 & \text{otherwise} \end{cases} \ .$$

Suppose (F_n) is a sequence in $\mathcal{F}_0$. If $F = \cap_n F_n$, we claim $F \in \mathcal{F}_0$. F is closed in X_0. If F is countable, then $F \subseteq [0, \alpha_0]$ for some $\alpha_0 < \omega_1$. There is a countable $\alpha_1 > \alpha_0$ such that $(\alpha_0, \alpha_1]$ contains infinitely many elements of F_n for each n: for each n there is a countable $\alpha_{1,n}$ such that $(\alpha_0, \alpha_{1,n}]$ contains infinitely many elements of F_n since F_n is uncountable and $[0, \alpha_0]$ is countable. Set $\alpha_1 = \sup_n \alpha_{1,n}$. Repeating the argument inductively, there is an increasing sequence (α_m) of countable ordinals such that there are infinitely many elements of each F_n in $(\alpha_{m-1}, \alpha_m]$ for each m. Set $\beta = \sup_m \alpha_m$. Then β is countable, and is in F_n for each n since F_n is closed in X_0. Thus $\beta \in F$, a contradiction. So F is uncountable.

Now let $\mathcal{A}$ be the collection of all subsets A of X for which either A or A^c contains a set in $\mathcal{F}_0$ (they cannot both since $\mathcal{F}_0$ is closed under countable intersections). Clearly $\mathcal{A}$ contains $\emptyset$ and X and is closed under complements. If (A_n) is a sequence in $\mathcal{A}$, then if some A_n contains a set in $\mathcal{F}_0$ so does $A = \cup_n A_n$, and otherwise each A_n^c contains an F_n, so $A^c = \cap_n A_n$ contains $\cap_n F_n \in \mathcal{F}_0$. If F is a closed set in X, then if F is uncountable it contains $F \cap X_0 \in \mathcal{F}_0$, and if F is countable F^c contains $[\alpha, \omega_1) \in \mathcal{F}_0$ for some α. It follows that every Borel set is in $\mathcal{A}$.

We claim μ is a Borel measure on X. We have $\mu(\emptyset) = 0$ and $\mu(X) = 1$. To show that μ is countably additive, let (B_n) be a sequence of pairwise disjoint Borel sets and set $B = \cup_n B_n$. At most one B_n can contain a set in $\mathcal{F}_0$. If one does, then so does B, and $\mu(B) = 1 = \sum_n \mu(B_n)$. If none contain one, then B_n^c contains an $F_n \in \mathcal{F}_0$ for each n, so B^c contains $\cap_n F_n$ and therefore B contains no set in $\mathcal{F}_0$ and $\mu(B) = 0 = \sum_n \mu(B_n)$. However, μ is not regular, since $\mu(\{\omega_1\}) = 0$, but $\mu(U) = 1$ for every open neighborhood of ω_1.

Finally, we fill in a loose end:

XVII.6.1.14. PROPOSITION. Let X be a topological space and μ a Borel measure on X.

(i) If X is T_1 (or, more generally, $T_{1/2}$ (XIII.7.10.26.)) and μ is locally finite and has no infinite atoms, then μ is neighborhood-finite.

(ii) If X is second countable and regular (i.e. separable and metrizable) and μ is neighborhood-finite, then μ is locally finite and σ-finite (and in particular has no infinite atoms).

PROOF: (i) If $p \in X$, then $\{p\}$ is closed, hence Borel, and has finite measure.

(ii) We show that X can be covered by countably many open sets of finite measure (so in particular μ is σ-finite). Let $\mathcal{U} = \{U_n : n \in \mathbb{N}\}$ be a countable base for the topology. For each $p \in X$, there is a neighborhood V of p with $\mu(V) < \infty$, and there is an n such that $p \in U_n \subseteq V$ since $\mathcal{U}$ is a base; so the U_n with $\mu(U_n) < \infty$ cover X. Then μ is locally finite by XVII.6.1.12.. 👍

XVII.6.1.15. We call a Borel measure on a metric space *bounded-finite* if every bounded Borel set has finite measure. Some references define "locally finite" to mean "bounded-finite." But bounded-finite is an unnecessarily strong restriction in many cases. A bounded-finite measure is obviously neighborhood-finite, hence locally finite in the separable case. But a neighborhood-finite measure need not be bounded-finite: for example, there is a bounded metric ρ on $\mathbb{R}$ equivalent to the usual metric, and Lebesgue measure is neighborhood-finite but not bounded-finite on $(\mathbb{R}, \rho)$. In a metric space in which every closed bounded set is compact (e.g. $\mathbb{R}^n$ with the usual metric), a neighborhood-finite measure is bounded-finite. In particular, on a compact metric space, a neighborhood-finite measure is finite. ✋

Tight and Radon Measures

Locally finite Borel measures are regular: the measure of a Borel set can be approximated by the measures of closed subsets. It is often better to be able to approximate by the measures of *compact* subsets, and this is possible under fairly mild hypotheses.

XVII.6.1.16. DEFINITION. A Borel measure μ on a (Hausdorff) topological space X is *tight* if for every $\epsilon > 0$ there is a compact subset K of X with $\mu(X \setminus K) < \epsilon$.

This definition is usually reserved for finite measures.

XVII.6.1.17. DEFINITION. Let X be a Hausdorff topological space. A finite Borel measure μ on X is a *Radon measure* if for every Borel set $A \subseteq X$ and every $\epsilon > 0$ there is a compact $K \subseteq A$ and an open set U with $A \subseteq U$ and $\mu(U \setminus K) < \epsilon$.

XVII.6.1.18. The same definition makes sense for general Borel measures if $\mu(A)$ is required to be finite, but Radon measures are conventionally assumed to be finite (so the finiteness assumption does not have to be repeatedly specified). We will have occasion to use the definition for locally finite measures, which we will call "locally finite Radon measures."

We summarize some obvious connections between the definitions:

XVII.6.1.19. PROPOSITION. Let μ be a finite Borel measure on a Hausdorff topological space X. Then

(i) The measure μ is Radon if and only if it is regular and tight.

(ii) If X is metrizable, then μ is Radon if and only if it is tight (XVII.6.1.10.).

(iii) If X is compact, then μ is tight.

(iv) If X is compact and metrizable, then μ is Radon.

PROOF: Only (i) needs proof. If μ is Radon, it is obviously regular and tight. Conversely, if μ is regular and tight, Let A be a Borel set in X and $\epsilon > 0$. Let K be a compact set with $\mu(X \setminus K) < \epsilon/2$, and E and U closed and open sets with $E \subseteq A \subseteq U$ and $\mu(U \setminus E) < \epsilon/2$. Then $E \cap K$ is compact, and

$$\mu(U \setminus (E \cap K)) = \mu(U \setminus E) + \mu(E \setminus (E \cap K)) \leq \mu(U \setminus E) + \mu(X \setminus K) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$



XVII.6.1.20. EXAMPLE. Every finite Borel measure on a metric space is regular (XVII.6.1.10.), but not necessarily tight. Let X be a nonmeasurable subset of $[0,1]$ with $\lambda^*(X) = \lambda^*([0,1] \setminus X) = 1$ (). Give X the induced metric. If A is a Borel set in X, then $A = B \cap X$ for some Borel set B in $[0,1]$. Set $\mu(A) = \lambda(B)$ (i.e. $\mu = \lambda|_X$ in the sense of ()). If K is a compact subset of X, then K is compact in $[0,1]$, hence measurable, and $\lambda([0,1] \setminus K) \geq \lambda^*([0,1] \setminus X) = 1$, and thus $\mu(K) = \lambda(K) = 0$. Since $\mu(X) = 1$, μ is not tight (hence not Radon).

However, finite Borel measures on complete separable metric spaces are always Radon. We first need a lemma which is an extension of XVII.6.1.10. in this case:

XVII.6.1.21. LEMMA. Let X be a separable metric space, and μ a locally finite Borel measure on X. Then, for every Borel set $A \subseteq X$ with $\mu(A) < \infty$, and $\epsilon > 0$, there is a totally bounded closed set $E \subseteq A$ and an open set U containing A with $\mu(U \setminus E) < \epsilon$.

PROOF: Given A and ϵ, by XVII.6.1.10. there is a closed set $F \subseteq A$ and open U containing A with $\mu(U \setminus F) < \epsilon/2$, so it suffices to show there is a totally bounded closed set $E \subseteq F$ with $\mu(F \setminus E) < \epsilon/2$. Since F is itself a separable metric space and $\mu|_F$ is finite, and a relatively closed subset of F is closed in X, by replacing X by F and μ by $\mu|_F$ (and $\epsilon/2$ by ϵ), it suffices to show:

> If X is a separable metric space and μ a finite Borel measure on X, then for every $\epsilon > 0$ there is a totally bounded closed subset E of X with $\mu(X \setminus E) < \epsilon$.

To prove this, fix a dense sequence (x_k) in X and $\epsilon > 0$. For fixed $n \in \mathbb{N}$, let B_{nk} be the open ball of radius $1/n$ centered at x_k. Then the B_{nk} cover X. For $m \in \mathbb{N}$ set $V_{nm} = \cup_{k=1}^{m} B_{nk}$. Then (V_{nm}) is an increasing sequence of open, hence Borel, sets with union X, so

$$\mu(X) = \mu\left(\bigcup_{m=1}^{\infty} V_{nm}\right) = \sup_m \mu(V_{nm})$$

so there is an m with $\mu(V_{nm}) > \mu(X) - 2^{-n}\epsilon$. Let V_n be V_{nm} for such an m (say the first such m). We then have $\mu(X \setminus V_n) < 2^{-n}\epsilon$.
Let $V = \cap_n V_n$. Then V is a Borel set, and is totally bounded; and

$$\mu(X \setminus V) \leq \sum_{n=1}^{\infty} \mu(X \setminus V_n) < \sum_{n=1}^{\infty} 2^{-n}\epsilon = \epsilon \ .$$

Set $E = \bar{V}$. Since the closure of a totally bounded set is totally bounded (XIII.11.14.32.), E has the desired properties.



XVII.6.1.22. THEOREM. Every locally finite Borel measure on a Polish space is a locally finite Radon measure.

PROOF: Let X be a Polish space, and μ a locally finite Borel measure on X. Fix a complete metric on X. If A is a Borel set in X with $\mu(A) < \infty$, and $\epsilon > 0$, there are a totally bounded closed set E and open set U with $E \subseteq A \subseteq U$ and $\mu(U \setminus E) < \epsilon$. Since X is complete and E is closed, E is also complete, and since it is totally bounded it is compact. 👍

XVII.6.2. Method II Construction of Outer Measures

The standard construction of an outer measure from a set function (e.g. a premeasure) described in XVII.4.1.5. is sometimes called the *Method I* construction, particularly in Geometric Measure Theory. On a metric space, there is another construction of outer measures which is more roundabout, but gives interesting examples such as Hausdorff measures.

It is customary in Geometric Measure Theory to use upper-case script letters such as $\mathscr{M}$ to denote outer measures. (The reader is also cautioned that in much of the literature for Geometric Measure Theory, an outer measure is called a "measure." The outer measures of Geometric Measure Theory are typically metric outer measures and are applied only to Borel sets, where they are indeed countably additive.)

Method II uses covers of sets which are "fine" in the sense that every point is contained in a set in the cover of arbitrarily small diameter:

XVII.6.2.1. Definition. Let (X, ρ) be a metric space, and $E \subseteq X$. A *Vitali cover* of E is a collection $\mathcal{V}$ of subsets of X such that, for every $x \in E$ and $\epsilon > 0$, there is a $V \in \mathcal{V}$ with $x \in V$ and $diam(V) < \epsilon$.

XVII.6.2.2. The notion of Vitali cover depends on ρ. It is easily checked that if σ is another metric on X uniformly equivalent to ρ, then a collection $\mathcal{V}$ of subsets of X is a Vitali cover for E with respect to ρ if and only if it is a Vitali cover with respect to σ. However, this is not necessarily true if ρ and σ are just equivalent (). To denote the dependence on ρ where necessary, we will use the terminology "ρ-Vitali cover."

XVII.6.2.3. The sets in a Vitali cover are not required to be of any particular form. But note that if $\mathcal{V}$ is a Vitali cover of E, then

$$\bar{\mathcal{V}} = \{\bar{V} : V \in \mathcal{V}\}$$

is a Vitali cover of E consisting of closed sets. Thus we may work with Vitali covers consisting of closed sets, which is often convenient.

XVII.6.2.4. Examples. (i) The power set of X is a Vitali cover of any subset of X.

(ii) The collection of open balls in (X, ρ) is a Vitali cover of any subset E of X.

(iii) The collection of closed balls in (X, ρ) is a Vitali cover of any subset E of X.

(iv) In (ii) and (iii), we may restrict to balls of rational radius and/or with centers in E.

(v) If $\mathcal{V}$ is a Vitali cover of E and $\delta > 0$, then

$$\mathcal{V}_\delta = \{V \in \mathcal{V} : diam(V) < \delta\}$$

$$\bar{\mathcal{V}}_\delta = \{V \in \mathcal{V} : diam(V) \leq \delta\}$$

are also Vitali covers of E.

(vi) A Vitali cover of E is also a Vitali cover of any subset of E. In particular, a Vitali cover of X is a Vitali cover of any subset of X.

XVII.6.2.5. Fix a metric space (X, ρ) and a Vitali cover $\mathcal{V}$ of X. We may assume $\emptyset \in \mathcal{V}$. Let ι be a function from $\mathcal{V}$ to $[0, +\infty]$ with $\iota(\emptyset) = 0$. For any $\delta > 0$, define $\mathscr{M}^\iota_\delta$ to be the outer measure on X defined by the restriction of ι to $\mathcal{V}_\delta$ (XVII.6.2.4.(v)) as in (), i.e. for $E \subseteq X$ define

$$\mathscr{M}^\iota_\delta(E) = \inf\left\{\sum_{k=1}^{\infty} \iota(V_k) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \in \mathcal{V}, diam(V_k) < \delta\right\}$$

(with $\mathscr{M}^\iota_\delta(E) = +\infty$ if there are no countable covers of E by sets in $\mathcal{V}_\delta$). As usual, since $\emptyset \in \mathcal{V}$ and $\iota(\emptyset) = 0$, finite as well as countably infinite covers are allowed in computing $\mathscr{M}^\iota_\delta$.

As δ decreases, there are fewer such covers, so $\mathscr{M}^\iota_\delta$ increases. Define

$$\mathscr{M}^\iota = \sup_{\delta > 0} \mathscr{M}^\iota_\delta$$

and note that $\mathscr{M}^\iota$ is an outer measure by (). $\mathscr{M}^\iota$ is called the *outer measure defined by* $\mathcal{V}$ *and* ι *by Method II.*

The outer measure $\mathscr{M}^\iota$ does depend on $\mathcal{V}$ in general, but the standard notation does not reflect this since in reasonable cases (e.g. ()) it is independent of $\mathcal{V}$. It is most common to take $\mathcal{V}$ to be the entire power set of X.

XVII.6.2.6. Use of a strict inequality in the diameters of the sets used is arbitrary. We could have also used the covers $\bar{\mathcal{V}}_\delta$ to obtain outer measures $\bar{\mathscr{M}}^\iota_\delta$, where

$$\bar{\mathscr{M}}^\iota_\delta(E) = \inf\left\{\sum_{k=1}^{\infty} \iota(V_k) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \in \mathcal{V}, diam(V_k) \leq \delta\right\}$$

and then obtained

$$\mathscr{M}^\iota = \sup_{\delta > 0} \bar{\mathscr{M}}^\iota_\delta$$

since we have

$$\mathscr{M}^\iota_{\delta'}(E) \leq \bar{\mathscr{M}}^\iota_\delta(E) \leq \mathscr{M}^\iota_\delta(E)$$

for any E and any $0 < \delta < \delta'$.

XVII.6.2.7. In the most commonly used special case where $\mathcal{V} = \mathcal{P}(X)$, we may also restrict to covers by closed sets:

$$\mathscr{M}^\iota_\delta(E) = \inf\left\{\sum_{k=1}^{\infty} \iota(V_k) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \text{ closed}, diam(V_k) < \delta\right\}$$

since $diam(\bar{A}) = diam(A)$ for any $A \subseteq X$. In some cases, one can also restrict to covers by open sets (cf. XVII.6.3.6.).

XVII.6.2.8. It is the passage to the limit as $\delta \to 0$ which gives Method II outer measures some nice properties making them useful. The idea is very similar to the way arc length for curves is defined: short chords and a passage to a limit are necessary to truly give an appropriate notion of length for curves. See () for more on this.

The first example of a nice property of Method II outer measures is:

XVII.6.2.9. PROPOSITION. For any $\mathcal{V}$ and ι, $\mathscr{M}^\iota$ is a metric outer measure.

PROOF: Suppose $A, B \subseteq X$ and $\rho(A,B) > 0$. If $\delta < \frac{\rho(A,B)}{2}$, then covers of A and B by sets in $\mathcal{V}_\delta$ are disjoint, so any cover of $A \cup B$ by sets in $\mathcal{V}_\delta$ breaks into a disjoint union of a cover of A and a cover of B, and every pair of covers of A and B arises this way. It follows that

$$\mathscr{M}^\iota_\delta(A \cup B) = \mathscr{M}^\iota_\delta(A) + \mathscr{M}^\iota_\delta(B)$$

by taking infima. Then taking supremum over such small δ gives

$$\mathscr{M}^\iota(A \cup B) = \mathscr{M}^\iota(A) + \mathscr{M}^\iota(B) \ .$$



XVII.6.2.10. So by XVII.6.1.4., $\mathscr{M}^\iota$ defines a Borel measure on X, also denoted $\mathscr{M}^\iota$.

XVII.6.3. Hausdorff Measure

Perhaps the most important example of the Method II construction of outer measures is Hausdorff measure. Hausdorff measure is most commonly considered on $\mathbb{R}^n$, but it can be equally well defined (with slight additional technicalities) on any metric space. It gives a metric notion of "dimension" for subsets of the space, which need not be an integer. Actually, we will get a more refined and flexible notion of metric dimension than just a number.

To make the construction work on a metric space (X, ρ), we need to be able to cover subsets of X with countable collections of balls of arbitrarily small radius. Thus it follows easily that X must be separable. So throughout this section all metric spaces will be assumed *separable*.

We first consider a class of functions:

XVII.6.3.1. DEFINITION. Let $\mathcal{H}$ be the set of all nondecreasing, right continuous functions from $[0, +\infty]$ to $[0, +\infty]$. Let $\mathcal{H}_0$ be the set of functions $h \in \mathcal{H}$ with $h(0) = 0$ (and hence $\lim_{t \to 0+} h(t) = 0$).

If (X, ρ) is any metric space, we will define a Hausdorff measure on (X, ρ) corresponding to each $h \in \mathcal{H}$. (Unfortunately, there are many h's in the standard notation for Hausdorff measure, which must be carefully distinguished; it will be impossible to stick to the recommendations of II.1.3.14. if standard notation is used.)

XVII.6.3.2. DEFINITION. Let (X, ρ) be a metric space, and $h \in \mathcal{H}$. Define the *Hausdorff measure* $\mathscr{H}^h$ on X by $\mathscr{H}^h = \mathscr{M}^\iota$, the Method II outer measure on X using $\mathcal{V} = \mathcal{P}(X)$ and $\iota(V) = h(diam(V))$ (set $\iota(\emptyset) = 0$). Thus $\mathscr{H}^h = \sup_{\delta > 0} \mathscr{H}^h_\delta$, where for $E \neq \emptyset$ we have

$$\mathscr{H}^h_\delta(E) = \inf \left\{ \sum_{k=1}^{\infty} h(diam((V_k)) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X, diam(V_k) < \delta \right\}$$

(with $\mathscr{H}^h_\delta(E) = +\infty$ if there are no countable covers of E by sets of diameter $< \delta$).

We will use the notation $\mathscr{H}^h$ for both the outer measure and the associated Borel measure. When the dependence on the metric ρ is relevant, we will write $\mathscr{H}^h_\rho$.

The most important case is when h is a power function:

XVII.6.3.3. DEFINITION. Let $s > 0$. Then *s-dimensional Hausdorff measure* $\mathscr{H}^{(s)}$ on a metric space (X, ρ) is $\mathscr{H}^h$, where $h(t) = t^s$. In other words, $\mathscr{H}^{(s)} = \sup_{\delta>0} \mathscr{H}^{(s)}_\delta$, where for $E \neq \emptyset$ we have

$$\mathscr{H}^{(s)}_\delta(E) = \inf\left\{\sum_{k=1}^{\infty}[diam((V_k)]^s : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X, diam(V_k) < \delta\right\}$$

(with $\mathscr{H}^{(s)}_\delta(E) = +\infty$ if there are no countable covers of E by sets of diameter $< \delta$).

We also define $\mathscr{H}^{(0)}$ to be $\mathscr{H}^h$, where h is the constant function 1.

XVII.6.3.4. PROPOSITION. The measure $\mathscr{H}^{(0)}$ is counting measure on X.

PROOF: Suppose E is a finite subset of X with n elements. If δ is smaller than the minimum distance between distinct points of E, any cover of E by sets of diameter $< \delta$ must have at least n elements, so $\mathscr{H}^{(0)}_\delta(E) = n$. A similar argument applies in the case where E is infinite: for any n, we have $\mathscr{H}^{(0)}_\delta(E) > n$ for sufficiently small δ. 👍

XVII.6.3.5. In the definition of Hausdorff measure, we can also use the outer measures $\bar{\mathscr{H}}^h_\delta$, where

$$\bar{\mathscr{H}}^h_\delta(E) = \inf\left\{\sum_{k=1}^{\infty} h(diam((V_k)) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X, diam(V_k) \leq \delta\right\}$$

(cf. XVII.6.2.6.).

In computing $\mathscr{H}^h_\delta(E)$, we may restrict to covers by closed sets by XVII.6.2.7.. We may also use open covers (this is the place where right continuity of h is used):

XVII.6.3.6. PROPOSITION. Let (X, ρ) be a metric space, $h \in \mathcal{H}$, and $\delta > 0$. Then

$$\mathscr{H}^h_\delta(E) = \inf\left\{\sum_{k=1}^{\infty} h(diam((V_k)) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X \text{ open}, diam(V_k) < \delta\right\}$$

for any $E \subseteq X$.

PROOF: We clearly have

$$\mathscr{H}^h_\delta(E) \leq \inf\left\{\sum_{k=1}^{\infty} h(diam((V_k)) : E \subseteq \cup_{k=1}^{\infty} V_k, V_k \subseteq X \text{ open}, diam(V_k) < \delta\right\} .$$

For the reverse inequality, there is nothing to prove if $\mathscr{H}^h_\delta(E) = +\infty$, so assume $\mathscr{H}^h_\delta(E) < +\infty$. Let $\epsilon > 0$, and fix a cover $\{V_k\}$ of E with $diam(V_k) < \delta$ for all k and

$$\sum_{k=1}^{\infty} h(diam(V_k)) < \mathscr{H}^h_\delta(E) + \frac{\epsilon}{2} .$$

For each k, let U_k be an open set containing V_k such that $diam(U_k) < \delta$ and $h(diam(U_k)) \le h(diam(V_k)) + \frac{\epsilon}{2^{k+1}}$; U_k can be taken to be

$$U_k = \{x \in X : \rho(x, V_k) < r\}$$

for some sufficiently small $r > 0$, using the right continuity of h. Then

$$\sum_{k=1}^{\infty} h(diam(U_k)) \le \sum_{k=1}^{\infty} h(diam(V_k)) + \frac{\epsilon}{2} < \mathscr{H}^h_\delta(E) + \epsilon$$

and since $\epsilon > 0$ is arbitrary, we obtain the reverse inequality. 👍

XVII.6.3.7. The Hausdorff measures $\mathscr{H}^h$, even the s-dimensional Hausdorff measures $\mathscr{H}^{(s)}$, depend strongly on the choice of metric. If σ is uniformly equivalent to ρ, ...

Spherical Measures

XVII.6.3.8. Many slight variations on Hausdorff measure are possible. One variation is to take $\mathcal{V}$ to be the set of balls in (X, ρ) (in metric spaces such as $\mathbb{R}^n$ where each closed ball is the closure of an open ball of the same diameter, it does not matter whether one uses open balls, closed balls, or arbitrary balls in between; note that in a general metric space the diameter of a ball of radius r may be strictly smaller than $2r$). The measures made in the same way are sometimes called *spherical measures* and denoted $\mathscr{S}^h$ or $\mathscr{S}^{(s)}$. The measure $\mathscr{S}^h$ is not the same as $\mathscr{H}^h$ since a *set* of diameter d is not necessarily contained in a *ball* of diameter d (cf. (); it is clearly contained in a ball of diameter $\le 2d$). We have $\mathcal{H}^h(E) \le \mathscr{S}^h(E)$ for any h and E. For power functions we have

$$\mathscr{H}^{(s)}(E) \le \mathscr{S}^{(s)}(E) \le 2^s \mathscr{H}^{(s)}(E)$$

for any s and E. The relation for general h is not so simple unless $\lim_{t \to 0+} \dfrac{h(2t)}{h(t)}$ exists.

XVII.6.3.9. There is also a question of what should be considered the "size" of a ball in a general metric space: note that in a general metric space the diameter of a ball of radius r may be strictly smaller than $2r$ (this is not an issue in Euclidean space, but is even in subsets of Euclidean space with the induced metric). It is more convenient in general metric spaces to take $\iota(B) = 2r$ instead of $diam(B)$ for a ball of radius r. This will give "radius-based spherical measure"; we could also consider diameter-based spherical measure, which gives an alternate theory not as nicely behaved in general. Similarly, for other measures such as packing measure we will only consider radius-based versions. Thus, to make $\mathscr{S}^h$, we take for $\delta > 0$

$$\mathscr{S}^h_\delta(E) = \sum_{k=1}^{\infty} h(2r_k)$$

where the infimum is over all countable covers (B_k) of E by balls of radius $r_k \le \delta$, and

$$\mathscr{S}^h(E) = \inf_{\delta > 0} \mathscr{S}^h_\delta(E) \ .$$

We will remain with diameter-based Hausdorff measure.

XVII.6.3.10. A variation of spherical measure is *center-based spherical measure*, requiring the centers of the balls to be contained in the subset E. Specifically for $\delta > 0$ let

$$\mathscr{C}_\delta^{(s)} = \inf \sum_{k=1}^{\infty} (2r_k)^s$$

where the infimum is over all countable covers of E with balls centered in E, of diamerer $\leq \delta$, and

$$\tilde{\mathscr{C}}^{(s)}(E) = \inf_{\delta>0} \mathscr{C}_\delta^{(s)}(E) \ .$$

There is a subtlety, however (cf. [Edg08, 1.4.1]): $\tilde{\mathscr{C}}^{(s)}$ is not necessarily monotone, hence not an outer measure. This is easily fixed, however: write

$$\mathscr{C}^{(s)}(E) = \sup_{A \subseteq E} \tilde{\mathscr{C}}^{(s)}(A) \ .$$

The relations between these outer measures for power functions are:

XVII.6.3.11. Proposition. Let (X, ρ) be a metric space, and $E \subseteq X$. If $s \geq 0$, then

$$\mathscr{H}^{(s)}(E) \leq \mathscr{S}^{(s)}(E) \leq \mathscr{C}^{(s)}(E) \leq 2^s \mathscr{S}^{(s)}(E) \ .$$

[Note that the inequality $\mathscr{S}^{(s)}(E) \leq 2^s \mathscr{H}^{(s)}(E)$ of XVII.6.3.8. is for diameter-based spherical measure and not radius-based spherical measure; this inequality holds for radius-based spherical measure in Euclidean space but not in general metric spaces.]

XVII.6.3.12. Hausdorff measures are rarely finite or even σ-finite. They are, however, often semifinite: $\mathscr{H}^{(s)}(X, \rho)$ is always semifinite if (X, ρ) is complete and separable, a surprisingly recent result [How95] (special cases have long been known, cf. [Bes52], [Dav52], [Lar67b]). The case of $\mathscr{H}^h$ for more general h has been shown to be semifinite if (X, ρ) is metrically finite-dimensional (XVII.6.7.20.) [Lar67b] or if h satisfies a growth condition [How95], but this is false in general [DR69].

Hausdorff Dimension

As s increases, the Hausdorff measure $\mathscr{H}^{(s)}(E)$ of a subset E of a metric space (X, ρ) jumps suddenly down from $+\infty$ to 0. The critical value where this occurs is called the *Hausdorff dimension* of E. More generally, there are typically functions $h \in \mathcal{H}$ whose rate of convergence to 0 describe the boundary between infinite and zero Hausdorff measure.

XVII.6.3.13. Proposition. Suppose $g, h \in \mathcal{H}$ and $\lim_{t \to 0+} \frac{g(t)}{h(t)} = 0$. Let (X, ρ) be a metric space, and $E \subseteq X$.

(i) If $\mathscr{H}^h(E) < +\infty$, then $\mathscr{H}^g(E) = 0$.

(ii) If $\mathscr{H}^g(E) > 0$, then $\mathscr{H}^h(E) = +\infty$.

PROOF: Statement (ii) is the contrapositive of (i). To prove (i), let $\epsilon > 0$, and choose $\eta > 0$ such that $g(t) \leq \frac{\epsilon}{\mathscr{H}^h(E)+1} h(t)$ for $t < \eta$. If $0 < \delta < \eta$, there is a cover (V_k) of E with $diam(V_k) < \delta$ for all k, such that

$$\sum_{k=1}^{\infty} h(diam(V_k)) < \mathscr{H}^h_\delta(E) + 1 \leq \mathscr{H}^h(E) + 1$$

and for this cover we have

$$\sum_{k=1}^{\infty} g(diam(V_k)) \leq \frac{\epsilon}{\mathscr{H}^h(E) + 1} \sum_{k=1}^{\infty} h(diam(V_k)) < \epsilon$$

so we have $\mathscr{H}^g_\delta(E) < \epsilon$. Taking the supremum over such δ, we have $\mathscr{H}^g(E) \leq \epsilon$. Since $\epsilon > 0$ is arbitrary, $\mathscr{H}^g(E) = 0$. 👍

XVII.6.3.14. COROLLARY. Let $0 \leq s < t$. Let (X, ρ) be a metric space, and $E \subseteq X$. Then
(i) If $\mathscr{H}^{(s)}(E) < +\infty$, then $\mathscr{H}^{(t)}(E) = 0$.
(ii) If $\mathscr{H}^{(t)}(E) > 0$, then $\mathscr{H}^{(s)}(E) = +\infty$.

XVII.6.3.15. COROLLARY. Let (X, ρ) be a metric space, and E an infinite subset of X. Then

$$\sup\{s \geq 0 : \mathscr{H}^{(s)}(E) = \infty\} = \inf\{t > 0 : \mathscr{H}^{(t)}(E) = 0\}$$

(set $\inf(\emptyset) = +\infty$).

XVII.6.3.16. DEFINITION. The common value in XVII.6.3.15. is called the *Hausdorff dimension* of E in (X, ρ), denoted $dim_H(E)$. We also define the Hausdorff dimension of a finite set to be 0.

XVII.6.3.17. So the Hausdorff dimension of E is the critical value s where the Hausdorff measure $\mathscr{H}^{(s)}(E)$ jumps from $+\infty$ to 0. Examples show that if $s = dim_H(E)$, then $\mathscr{H}^{(s)}(E)$ can be 0, $+\infty$, or any finite positive number.

XVII.6.3.18. If only power functions are considered, for a given E there is an exact boundary between the functions h for which $\mathscr{H}^h(E) = +\infty$ and the h for which $\mathscr{H}^h(E) = 0$. But if all functions in $\mathcal{H}$ are considered, there is not such a clean boundary. Still, if there is an $h \in \mathcal{H}$ for which $0 < \mathscr{H}^h(E) < +\infty$, it can be considered a reasonable boundary: for any $g \in \mathcal{H}$ with $\lim_{t \to 0+} \frac{g(t)}{h(t)} = 0$, we have $\mathscr{H}^g(E) = 0$, and for any $g \in \mathcal{H}$ with $\lim_{t \to 0+} \frac{g(t)}{h(t)} = +\infty$, we have $\mathscr{H}^g(E) = +\infty$. Such an h is not generally unique, however; but dividing $\mathcal{H}$ into those h for which $\mathscr{H}^h(E) < +\infty$ and those h for which $\mathscr{H}^h(E) = +\infty$ (or, almost equivalently, those h for which $\mathscr{H}^h(E) = 0$ and those h for which $\mathscr{H}^h(E) > 0$) is an illuminating partition.

XVII.6.3.19. Because of XVII.6.3.8., the same phenomenon works for spherical measures (either radius-based or diameter-based) and, for diameter-based spherical measure, gives the same s as the transition point. Thus Hausdorff dimension can be calculated using diameter-based spherical measures, i.e. covering with balls. For more general h the phenomenon is not quite the same or so simple. And in general metric spaces, if one uses radius-based spherical measure, even the Hausdorff dimension can be different (it is smaller in general). Not a problem, of course, in Euclidean space.

XVII.6.4. Hausdorff Measure on $\mathbb{R}^n$

XVII.6.4.1. Hausdorff measures and Hausdorff dimension were originally defined and studied for subsets of $\mathbb{R}^n$, and are still of most interest there.

Unfortunately, there is no consistency in references about the exact definition of Hausdorff measure in $\mathbb{R}^n$. It appears that most references define the measures in the same way up to a constant factor, which can vary from reference to reference. There are two predominant definitions used in the literature: the "unnormalized Hausdorff measure," primarily used in the theory of fractals, and the "normalized Hausdorff measure," primarily used in Geometric Measure Theory (only s-dimensional Hausdorff measure can be reasonably normalized, and usually only integer values of s are used). The normalized version has nicer properties for most purposes in analysis, since it exactly coincides with Lebesgue measure and arc length/surface area/surface measure in appropriate dimensions. Thus we will concentrate on the normalized case. The distinction is not important for many purposes: the crucial distinction is often simply whether a set has zero, infinite, or finite nonzero measure. For example, both versions give the same Hausdorff dimension for subsets of $\mathbb{R}^n$.

Unnormalized Hausdorff Measure

XVII.6.4.2. Definition. Let $h \in \mathcal{H}$. Unnormalized Hausdorff h-measure $\mathscr{H}^h$ on $\mathbb{R}^n$ is $\mathscr{H}^h$ on $\mathbb{R}^n$ with the usual Euclidean metric ρ_2. If $0 \leq s$, unnormalized s-dimensional Hausdorff measure $\mathscr{H}^{(s)}$ on $\mathbb{R}^n$ is $\mathscr{H}^{(s)}$ for $(\mathbb{R}^n, \rho_2)$.

As usual, we will use the same notation for the outer measure and the Borel measure.

Normalized Hausdorff Measure

XVII.6.4.3. Definition. If $0 \leq s$, normalized s-dimensional Hausdorff measure $\mathscr{L}^{(s)}$ on $\mathbb{R}^n$ is $\alpha_s \mathscr{H}^{(s)}$ for $(\mathbb{R}^n, \rho_2)$, where

$$\alpha_s = \frac{\pi^{s/2}}{2^s \Gamma(\frac{s}{2}+1)}$$

is a normalizing constant (Γ is the Gamma function ().)

XVII.6.4.4. If s is an integer, α_s is the s-dimensional "volume" of a ball in $\mathbb{R}^s$ of diameter 1 (radius $\frac{1}{2}$). Thus $\alpha_1 = 1$, $\alpha_2 = \frac{\pi}{4}$, $\alpha_3 = \frac{1}{8}\left(\frac{4}{3}\pi\right) = \frac{\pi}{6}$, etc.; see ().

XVII.6.4.5. The notation $\mathscr{L}^{(s)}$ for normalized Hausdorff measure is nonstandard; it is most commonly called $\mathscr{H}^{(s)}$. But this notation is also used by some authors to mean unnormalized Hausdorff measure, which is consistent with its use in general metric spaces, so we use $\mathscr{H}^{(s)}$ to mean unnormalized Hausdorff measure.

We have chosen the notation $\mathscr{L}^{(s)}$ for normalized Hausdorff measure since for appropriate s it coincides with usual Lebesgue measure or standard variants. Note that we do not define normalized Hausdorff measure for general $h \in \mathcal{H}$.

XVII.6.4.6. Hausdorff measure on $\mathbb{R}^n$ can be defined in other ways too. For example, in [**?**] it is defined using the metric ρ_∞ (). Hausdorff measure defined in this way is not a constant multiple of the usual Hausdorff measure in most cases, but it is an equivalent measure and gives the same theory of Hausdorff dimension, and is technically simpler in some respects.

XVII.6.5. Hausdorff Measure and Hausdorff Dimension for Self-Similar Sets

XVII.6.6. Packing Measure and other Geometric Measures

Hausdorff measure, or more precisely spherical measure (XVII.6.3.8.) which is equivalent and hence gives the same dimension, uses good coverings by balls. There is a "dual" notion which uses dense packings by disjoint balls, not surprisingly called packing measure. Packing measure is a much more recent development than Hausdorff measure, being defined by TRICOT in 1982 [].

The definition of packing measure involves yet another level of approximation:

XVII.6.6.1. DEFINITION. Let (X, ρ) be a metric space, and $s \geq 0$. Let $E \subseteq X$. For each $\delta > 0$ set

$$\mathscr{P}_\delta^{(s)}(E) = \sup\left\{\sum_{k=1}^{\infty}(2r_k)^s\right\}$$

where the supremum is over all sequences (B_k) of pairwise disjoint open balls of radii $r_k \leq \delta$ with centers in E (empty balls are allowed with radius 0, i.e. finite sequences of B_k are allowed). Set

$$\mathscr{P}_0^{(s)}(E) = \inf_{\delta>0} \mathscr{P}_\delta^{(s)}(E)$$

(note that $\mathscr{P}_\delta^{(s)}(E)$ decreases as δ decreases) and

$$\mathscr{P}^{(s)}(E) = \inf\left\{\sum_{k=1}^{\infty} \mathscr{P}_0^{(s)}(E_k) : E \subseteq \bigcup_{k=1}^{\infty} E_k\right\}$$

i.e. $\mathscr{P}^{(s)}$ is the outer measure defined by the set function $\mathscr{P}_0^{(s)}$ (). $\mathscr{P}^{(s)}$ is called *s-dimensional packing measure* on (X, ρ).

XVII.6.6.2. Note that "packing measure" is actually an outer measure. But it is easily seen to be a metric outer measure, so all Borel sets are measurable and it is truly a measure on Borel sets. $\mathscr{P}_\delta^{(s)}$ is an outer measure for any $\delta > 0$, but $\mathscr{P}_0^{(s)}$ is not an outer measure in general (), so the last step is necessary even though it makes packing measure harder to compute. Roughly speaking, the last step allows not just packings by disjoint balls with centers in E, but "packings" by "almost disjoint" balls with centers close to E, which can be taken to be in E:

XVII.6.6.3. PROPOSITION. Let (X,ρ) be a metric space, and $E \subseteq X$. Then, for any s,

$$\inf\left\{\sum_{k=1}^{\infty} \mathscr{P}_0^{(s)}(E_k) : E \subseteq \bigcup_{k=1}^{\infty} E_k\right\} = \inf\left\{\sum_{k=1}^{\infty} \mathscr{P}_0^{(s)}(E_k) : E \subseteq \bigcup_{k=1}^{\infty} E_k, E_k \subseteq E\right\} .$$

PROOF: Let $E \subseteq \bigcup_{k=1}^{\infty} E_k$, and let $\tilde{E}_k = E \cap E_k$ for each k. Then $\mathscr{P}_0^{(s)}(\tilde{E}_k) \leq \mathscr{P}_0^{(s)}(E_k)$ since $\mathscr{P}_0^{(s)}$ is monotone, and $E \subseteq \bigcup_{k=1}^{\infty} \tilde{E}_k$. 👍

XVII.6.6.4. A routine argument shows that disjoint closed balls can be used instead of disjoint open balls in the definition of packing measure, since the diameter of an open ball is the supremum of the diameters of the closed balls contained in it, and two open balls are disjoint if and only if all pairs of closed balls contained in them are disjoint. It also suffices to use finite packings rather than sequences. Packing measures on separable complete metric spaces are semifinite [JP95], but not generally σ-finite.

It is also easily verified that $\mathscr{P}^{(0)}$ is counting measure on $(X, \mathcal{P}(X))$.

XVII.6.6.5. PROPOSITION. Let (X,ρ) be a metric space, $E \subseteq X$, and $s \geq 0$. Then $\mathscr{P}_0^{(s)}(E) = \mathscr{P}_0^{(s)}(\bar{E})$ (where $\bar{E}$ is the closure of E).

PROOF: Any δ-packing of $\bar{E}$ can be approximated by a δ'-packing of E, for any $\delta' < \delta$. 👍

XVII.6.6.6. This shows why $\mathscr{P}_0^{(s)}$ is not an outer measure: if E is a countable dense set in X, then $\mathscr{P}_0^{(s)}(E) = \mathscr{P}_0^{(s)}(X)$, which is not necessarily zero. But $\mathscr{P}_0^{(s)}(pt) = 0$. So $\mathscr{P}_0^{(s)}$ is not countably subadditive in general.

Packing measure behaves in a similar way to Hausdorff measure as s varies:

XVII.6.6.7. PROPOSITION. Let $0 \leq s < t$. Let (X,ρ) be a metric space, and $E \subseteq X$. Then
(i) If $\mathscr{P}^{(s)}(E) < +\infty$, then $\mathscr{P}^{(t)}(E) = 0$.
(ii) If $\mathscr{P}^{(t)}(E) > 0$, then $\mathscr{P}^{(s)}(E) = +\infty$.

PROOF: If (B_k) is a sequence of pairwise disjoint balls of radius $\leq \delta$ with centers in E, then $\sum_{k=1}^{\infty}(diam(B_k))^t \leq (2\delta)^{t-s}\sum_{k=1}^{\infty}(diam(B_k))^s$. Taking the infimum over all such sequences, we obtain $\mathscr{P}_\delta^{(t)}(E) \leq (2\delta)^{t-s}\mathscr{P}_\delta^{(s)}(E)$. Taking the infimum over all δ, if $\mathscr{P}_0^{(s)}(E) < +\infty$, then $\mathscr{P}_0^{(t)}(E) = 0$. Now suppose $\mathscr{P}^{(s)}(E) < +\infty$. Then there is a sequence (E_k) with $E \subseteq \cup_{k=1}^{\infty} E_k$ and $\sum_{k=1}^{\infty} \mathscr{P}_0^{(s)}(E_k) < +\infty$. Then $\mathscr{P}_0^{(s)}(E_k) < +\infty$ for all k, so $\mathscr{P}_0^{(t)}(E_k) = 0$ for all k, and thus

$$\mathscr{P}^{(t)}(E) \leq \sum_{k=1}^{\infty} \mathscr{P}_0^{(t)}(E) = 0 .$$

Statement (ii) is the contrapositive of (i). 👍

XVII.6.6.8. Corollary. Let (X, ρ) be a metric space, and E an infinite subset of X. Then

$$\sup\{s \geq 0 : \mathscr{P}^{(s)}(E) = \infty\} = \inf\{t > 0 : \mathscr{P}^{(t)}(E) = 0\}$$

(set $\inf(\emptyset) = +\infty$).

XVII.6.6.9. Definition. The common value in XVII.6.6.8. is called the *packing dimension* of E in (X, ρ), denoted $dim_P(E)$. We also define the packing dimension of a finite set to be 0.

XVII.6.6.10. So the packing dimension of E is the critical value s where the packing measure $\mathscr{P}^{(s)}(E)$ jumps from $+\infty$ to 0. Examples show that if $s = dim_P(E)$, then $\mathscr{P}^{(s)}(E)$ can be 0, $+\infty$, or any finite positive number.

XVII.6.6.11. We have that $dim_H(E) \leq dim_P(E)$ for any $E \subseteq X$ (XVII.6.6.18.). The Hausdorff and packing dimensions of a subset can differ (XVII.6.9.2.(d)), although for well-behaved subsets they frequently coincide; when they do, the "fractal dimension" of the set is well defined as the common value.

XVII.6.6.12. We can also consider "diameter-based packing measure." This is the same as radius-based packing measure on $\mathbb{R}^n$, but different in general metric spaces and can even yield a different packing dimension.

Packing Measure of Subsets

XVII.6.6.13. There is a potential ambiguity in packing measure: is the packing measure or packing dimension of a subset E of X the same whether E is regarded as a subset of X or just as a metric space with the restricted metric? More generally, if Y is a subset of X with the restricted metric, and $E \subseteq Y$, is the packing measure or packing dimension of E as a subset of Y the same as it is as a subset of X?

XVII.6.6.14. We may define some other variations of packing measure by using other types of "packing" (cf. [Edg98,]). If (X, ρ) is a metric space and $E \subseteq X$, we can describe a potential finite packing as a finite set $\{(x_j, r_j) : 1 \leq j \leq n\}$, where the $x_j \in E$ and $r_j > 0$, corresponding to the set $\{\bar{B}_{r_j}(x_j) : 1 \leq j \leq n\}$.

XVII.6.6.15. Fix $s > 0$ and define radius-based "packing" measures using (finite) "packings" of various forms:

$\mathscr{P}^{(s)}_{max}(E)$ by using "packings" for which $\rho(x_j, x_k) > \max(r_j, r_k)$ for all j, k.

$\mathscr{P}^{(s)}_{sub}(E)$ by using "packings" for which $\bar{B}_{r_j}(x_j) \cap \bar{B}_{r_k}(x_k) \cap E = \emptyset$ for all j, k.

$\mathscr{P}^{(s)}_{rad}(E)$ by using "packings" for which $\bar{B}_{r_j}(x_j) \cap \bar{B}_{r_k}(x_k) = \emptyset$ for all j, k.

$\mathscr{P}^{(s)}_{sum}(E)$ by using "packings" for which $\rho(x_j, x_k) > r_j + r_k$ for all j, k.

In each case, one takes for $\delta > 0$ the supremum of the sums $\sum_j (2r_j)^s$ over all "packings" with all $r_j < \delta$, then the inf for $\delta > 0$, and then the inf over all countable coverings of E.

A "packing" of any of these kinds is a "packing" of the kinds above it ("packings" of the last two kinds are true packings). The measure $\mathscr{P}^{(s)}_{rad}$ is just the previously-defined s-dimensional packing measure. As E varies, $\mathscr{P}^{(s)}_{sub}$ is not clearly an outer measure since a variable E is used in its definition, but if E is fixed, by XVII.6.6.3. $\mathscr{P}^{(s)}_{sub}(E)$ is the s-dimensional packing measure of the metric space $(E, \rho|_E)$. We have the inequalities

$$\mathscr{P}^{(s)}_{sum}(E) \leq \mathscr{P}^{(s)}_{rad}(E) \leq \mathscr{P}^{(s)}_{sub}(E) \leq \mathscr{P}^{(s)}_{max}(E) \leq 2^s \mathscr{P}^{(s)}_{sum}(E)$$

for any E. It follows that the packing dimension of E from any of these is the same.

In particular, although the s-dimensional radius-based packing *measure* of E as a subset of X may (potentially) differ from its packing *measure* as a metric space on its own, the radius-based packing *dimension* of E is the same whether regarded as a subset of X or as a metric space on its own.

The corresponding situation about diameter-based packing dimension is unclear. For one thing, the diameter of a ball B in X may be much larger than the diameter of the relative ball $B \cap E$.

XVII.6.6.16. We could also normalize packing measure on $\mathbb{R}^n$ as with Hausdorff measure. This is rarely done, however; the principal use of packing measure (at present) is to define and work with packing dimension, which is independent of normalization.

Relation Between Hausdorff and Packing Dimension

XVII.6.6.17. PROPOSITION. Let (X, ρ) be a metric space, and $E \subseteq X$. If $s \geq 0$, then

$$\mathscr{H}^{(s)}(E) \leq \mathscr{S}^{(s)}(E) \leq \mathscr{C}^{(s)}(E) \leq \mathscr{P}^{(s)}(E) \ .$$

PROOF: The first two inequalities were noted in XVII.6.3.11.. For the last inequality, it suffices to show that $\tilde{\mathscr{C}}^{(s)}(E) \leq \mathscr{P}^{(s)}_0(E)$ for all $E \subseteq X$. Fix E; we may assume $\mathscr{P}^{(s)}_0(E) < \infty$. Then, for some $\delta > 0$, $\mathscr{P}^{(s)}_\delta(E) < \infty$. Set $\mathcal{V} = \{\bar{B}_r(x) : x \in E, 0 < r < \delta\}$. By (), there is a disjoint family $\{\bar{B}_{r_j}(x_j)\}$ in $\mathcal{V}$ with

$$E \setminus \bigcup_{j=1}^{k} \bar{B}_{r_j}(x_j) \subseteq \bigcup_{j=k+1}^{\infty} \bar{B}_{3r_j}(x_j)$$

for all k.

Since $\{\bar{B}_{r_j}(x_j)\}$ is a packing of E by balls of radius $< \delta$,

$$\sum_{j=1}^{\infty} (2r_j)^s \leq \mathscr{P}^{(s)}_\delta(E) < \infty \ .$$

For each k,

$$\{\bar{B}_{r_j}(x_j) : 1 \leq j \leq k\} \cup \{\bar{B}_{3rj}(x_j) : j > k\}$$

is a cover of E with balls of radius $< 3\delta$, so

$$\mathscr{C}^{(s)}_{3\delta}(E) \leq \sum_{j=1}^{k}(2r_j)^s + \sum_{j=k+1}^{\infty}(6r_j)^s$$

for every k. Letting $k \to \infty$, we obtain

$$\mathscr{C}^{(s)}_{3\delta}(E) \leq \sum_{j=1}^{\infty}(2r_j)^s \leq \mathscr{P}^{(s)}_{\delta}(E) \ .$$

Letting $\delta \to 0$, the result follows. 👍

XVII.6.6.18. COROLLARY. Let (X, ρ) be a metric space, and $E \subseteq X$. Then

$$dim_H(E) \leq dim_P(E) \ .$$

Integralgeometric Measure

We now restrict to an integer value of s, which we will denote k. Another related measure on $\mathbb{R}^n$ "averages" over projections onto k-dimensional subspaces. For "nice" subspaces such as k-dimensional smooth submanifolds, this measure agrees with k-dimensional Hausdorff measure. This measure was defined by J. FAVARD in 1932 [].

XVII.6.6.19. If $0 \leq k \leq n$, let $Gr_k(\mathbb{R}^n)$ be the Grassmannian () of k-dimensional subspaces of $\mathbb{R}^n$, a manifold of dimension $k(n-k)$. Points of $Gr_k(\mathbb{R}^n)$ can be identified with the rank k orthogonal projections on $\mathbb{R}^n$. We can regard $Gr_k(\mathbb{R}^n)$ as the homogeneous space

$$O(n)/[O(k) \times O(n-k)]$$

and it thus has a unique Radon probability measure μ invariant under the natural action of O_n. (In fact μ is a smooth measure given by integration of a differential form.)

If $E \subseteq \mathbb{R}^n$ and P is a rank k projection on $\mathbb{R}^n$, for y in the range of P let $N(P|_E, y)$ be the number of points in $E \cap P^{-1}(\{y\})$. If E is a Borel set, then $N(P|_E, y)$ is a measurable function of y ().

XVII.6.6.20. DEFINITION. If E is a Borel set in $\mathbb{R}^n$, define the k-dimensional *integralgeometric measure* of E by

$$\mathscr{I}^{(k)}(E) = \frac{1}{\beta(n,k)} \int_{P \in Gr_k(\mathbb{R}^n)} \int_{P(\mathbb{R}^n)} N(P|_E, y)\, d\lambda^{(k)}(y)\, d\mu(P)$$

where $\lambda^{(k)}$ is k-dimensional Lebesgue measure on the range of P which is isomorphic to $\mathbb{R}^k$, and $\beta(n,k)$ is a normalizing constant

$$\beta(n,k) = \frac{\Gamma([k+1]/2)\Gamma([n-k+1]/2)}{\sqrt{\pi}\,\Gamma([n+1]/2)} \ .$$

Extend $\mathscr{I}^{(k)}$ to an outer measure on $\mathbb{R}^n$ in the usual way.

Note that $Gr_0(\mathbb{R}^n)$ and $Gr_n(\mathbb{R}^n)$ are one-point spaces, and it is easy to see that $\mathscr{I}^{(0)}$ is counting measure and $\mathscr{I}^{(n)} = \mathscr{L}^{(n)}$.

Many more similar measures are discussed in [Fed69, 2.10].

$\mathscr{H}$ is Hausdorff measure
$\mathscr{P}$ is packing measure
$\mathscr{I}$ is integralgeometric measure
$\mathscr{L}$ is Lebesgue measure

XVII.6.7. Metric Dimensions

Hausdorff and packing dimensions are two important examples of fractal dimensions for metric spaces. There are several more that have been considered in the literature, and two particularly important ones, Assouad dimension and metric dimension, and various box-counting dimensions. These fractal dimensions do not arise from measures in general, although see XVII.6.7.18.. It should be noted that these various notions of fractal dimension are *metric* invariants and not *topological* invariants, and must be carefully distinguished from topological dimension as discussed in () (although there are connections, e.g. XVII.6.7.23., XVII.6.7.25.).

Doubling and Homogeneous Metrics

XVII.6.7.1. Definition. A metric space (X, ρ) is *doubling* if there is a constant $K \in \mathbb{N}$ such that, for every $r > 0$, every open ball of radius r can be covered by at most K open balls of radius $r/2$.

XVII.6.7.2. If (X, ρ) is doubling, then any open ball of radius r can be covered by at most K^n open balls of radius $r/2^n$, for any n. It follows that every bounded set in X is totally bounded in a "uniform" way.

There is a (very!) closely related notion, based on packing:

XVII.6.7.3. Definition. If $C > 0$ and $s \geq 0$, a metric space (X, ρ) is (C, s)-*homogeneous* if for every $x \in X$, $r > 0$, and $\lambda > 1$, at most $C\lambda^s$ points can be placed in $B_{\lambda r}(x)$ which are at least r apart. (X, ρ) is s-*homogeneous* if it is (C, s)-homogeneous for some $C > 0$. (X, ρ) is *homogeneous* if it is s-homogeneous for some s.

XVII.6.7.4. The C in the definition is necessary: a finite metric space (X, ρ) is 0-homogeneous, but the smallest C necessary is $card(X)$.

XVII.6.7.5. The number of points a distance at least r apart which can be placed in $B_{\lambda r}(x)$ is roughly the number of disjoint open balls of radius $r/2$ which can be packed into $B_{\lambda r}(x)$ (it is not exactly the same since balls centered at points of $B_{\lambda r}(x)$ can overlap the boundary). In packing problems, closed balls are usually used. Since any closed ball is contained in an open ball with the same center and slightly larger radius, these quantities are "asymptotically" the same.

XVII.6.7.6. The connection between packing and covering is that if P is a set of points in a set S which are at least r apart, and $\mathcal{U}$ is a covering of S by open balls of radius $r/2$, then each ball in $\mathcal{U}$ contains at most one point of P, so $card(P) \leq card(\mathcal{U})$. Conversely, if P is a set of points of S at least r apart of maximum size, then the open balls of radius r centered at points of P cover S.

XVII.6.7.7. PROPOSITION. A metric space is doubling if and only if it is homogeneous.

PROOF: If (X, ρ) is (C, s)-homogeneous, $x \in X$, and $r > 0$, let P be a set of maximum cardinality of points of $B_{2r}(x)$ of distance at least r apart. Then the open balls of radius r centered at points of P cover $B_{2r}(x)$; thus there is a cover of cardinality $card(P) \leq C \cdot 2^s$, so (X, ρ) is doubling with $K = C \cdot 2^s$.

Conversely, suppose (X, ρ) is doubling with constant K. We may assume $K \geq 2$. We claim (X, ρ) is (C, s)-homogeneous, where $C = K^2$ and $s = \log_2 K$. Let $x \in X$ and $r > 0$. If $\lambda > 1$, choose n so that $2^{n-1} < \lambda \leq 2^n$. Then $B_{\lambda r}(x)$ can be covered by $K \cdot K^n = K^2 \cdot K^{n-1}$ open balls of radius $r/2$, so there can be no more than $K^2 \cdot K^{n-1}$ points in $B_{\lambda r}$ of distance at least r apart. And

$$K^2 \cdot K^{n-1} = K^2(2^s)^{n-1} = K^2(2^{n-1})^s < K^2\lambda^s \ .$$



XVII.6.7.8. There are several other equivalent conditions, described in the Appendix to [Luu98], which has a detailed discussion of Assouad dimension.

Assouad Dimension

The following definition was made in [Ass79] (with different notation):

XVII.6.7.9. DEFINITION. A metric space (X, ρ) which is homogeneous or, equivalently, doubling, has *finite Assouad dimension*, and the *Assouad dimension* of (X, ρ) is

$$dim_A(X, \rho) = \inf\{s : (X, \rho) \text{ is } s\text{-homogeneous}\} \ .$$

If (X, ρ) does not have finite Assouad dimension, we write $dim_A(X, \rho) = \infty$.

We often slightly abuse notation by writing $dim_A X$, although the Assouad dimension depends (greatly!) on the metric ρ.

The next proposition summarizes the most important basic properties of Assouad dimension.

XVII.6.7.10. PROPOSITION. Let (X, ρ) be a metric space. Then

(i) If $Y \subseteq X$, then $dim_A(Y, \rho|_Y) \leq dim_A(X, \rho)$, and $dim_A(Y, \rho|_Y) = dim_A(X, \rho)$ if Y is dense in X.

(ii) $dim_A(X, \rho) = dim_A(\bar{X}, \bar{\rho})$, where $(\bar{X}, \bar{\rho})$ is the completion of (X, ρ).

(iii) If $Y_1, \ldots, Y_n \subseteq X$, then $dim_A(Y_1 \cup \cdots \cup Y_n) = \max_{k=1}^n dim_A(Y_k)$.

(iv) If σ is a metric on X biLipschitz equivalent to ρ, i.e. there are constants $0 < a < b$ such that $a\rho \leq \sigma \leq b\rho$, then $dim_A(X, \sigma) = dim_A(X, \rho)$.

(v) If $dim_A(X, \rho) < \infty$, then X is separable and every bounded subset of X is totally bounded. If (X, ρ) is complete, then every closed bounded subset of X is compact, and X is locally compact and σ-compact.

XVII.6.7.11. Properties (i) and (iii) show that Assouad dimension behaves quite differently from topological dimension, cf. ().

XVII.6.7.12. Example. Let $X = \left\{\frac{1}{n} : n \in \mathbb{N}\right\} \cup \{0\} \subseteq \mathbb{R}$ with the induced metric ρ, i.e. $\rho\left(\frac{1}{n}, \frac{1}{m}\right) = \left|\frac{1}{n} - \frac{1}{m}\right|$. For any $t > 1$, the set of points at least $1/t^2$ apart of maximum size is $\{0\} \cup \left\{\frac{1}{n} : n \leq t\right\}$. It follows that $dim_A(X, \rho) = \frac{1}{2}$.

Assouad dimension behaves as expected for Euclidean space:

XVII.6.7.13. Proposition. Let X be a subset of $\mathbb{R}^n$ with nonempty interior (e.g. $\mathbb{R}^n$ itself). Then $dim_A(X) = n$.

Proof: $\mathbb{R}^n$ is n-homogeneous (), so $dim_A(X) \leq n$ (XVII.6.7.10.(i)). It suffices to show that an open cube E in $\mathbb{R}^n$ has Assouad dimension $\geq n$. By XVII.6.7.10.(iv) we may use $\|\cdot\|_\infty$, which is biLipschitz equivalent to the Euclidean norm. But for any $m \in \mathbb{N}$ and $r > 0$ we can put more than m^n points (in fact $(2m-1)^n$) of $\|\cdot\|_\infty$-distance at least r apart in an open cube of radius mr, so E cannot be s-homogeneous for any $s < n$. ☝

Doubling and Homogeneous Measures

There is a measure version of the metric notions of doubling and homogeneity:

XVII.6.7.14. Definition. Let (X, ρ) be a metric space, and μ a Borel measure on X.
(i) The measure μ on X is *doubling* for ρ if there is a $K \in \mathbb{N}$ such that

$$\mu(B_{2r}(x)) \leq K\mu(B_r(x))$$

for all $x \in X$ and $r > 0$, and $0 < \mu(B) < \infty$ for some (hence every) ball B.
(i) If $C > 0$, $s \geq 0$, the measure μ on X is (C, s)-*homogeneous* for ρ if

$$\mu(B_{\lambda r}(x)) \leq C\lambda^s\mu(B_r(x))$$

for all $x \in X$ and $\lambda > 1$, and $0 < \mu(B) < \infty$ for some (hence every) ball B. The measure μ is *s-homogeneous* if it is (C, s)-homogeneous for some C, and is *homogeneous* if it is s-homogeneous for some s.

XVII.6.7.15. A Borel measure is doubling if and only if it is homogeneous; the proof is virtually identical to the proof of ().

XVII.6.7.16. If (X, ρ) has a (C, s)-homogeneous measure, then it is (C, s)-homogeneous. The converse is false in general: $\mathbb{Q}$ with the metric from $\mathbb{R}$ is 1-homogeneous, but there is no doubling measure on $\mathbb{Q}$ (or on any countable metric space with no isolated points). However:

XVII.6.7.17. THEOREM. [LS98] Let $C > 1$ and $0 < s < t$. Then there is a $C' > 1$ such that every complete (C, s)-homogeneous metric space has a (C', t)-homogeneous measure.

XVII.6.7.18. COROLLARY. Let (X, ρ) be a complete metric space. Then

$$dim_A(X, \rho) = \inf\{s : X \text{ has a } s\text{-homogeneous measure}\} .$$

Metric Dimension

XVII.6.7.19. One feature of Assouad dimension is that it uses both large and small balls. This is an advantage in some contexts, but a defect in other contexts where only small balls are relevant (and there are also contexts where only large balls are important, e.g. [LeC73] and coarse geometry/topology ()). We can restrict to small balls in various ways:

XVII.6.7.20. DEFINITION. Let (X, ρ) be a metric space.

(i) [ML92] The *metric dimension* $dim_m(X, \rho)$ is the infimum of all s for which there is an $R > 0$ for which there is a $C > 0$ such that every ball of radius R is (C, s)-homogeneous (equivalently, there is a K such that every ball of radius R is doubling with constant K). If $dim_m(X, \rho) < \infty$, then (X, ρ) is *metrically finite-dimensional*.

(ii) [Lar67a] The *dimension number* $dim\text{-}n(X, \rho)$ is the infimum of all s such that every point of X has a neighborhood U with $dim_A(U) \leq s$.

There is inconsistency in the literature about this definition: some references define metric dimension differently. Our definition of metrically finite-dimensional is called "finite structural dimension" in [How95].

XVII.6.7.21. It is clear that, for any (X, ρ),

$$dim\text{-}n(X, \rho) \leq dim_m(X, \rho) \leq dim_A(X, \rho) .$$

These dimensions behave somewhat more like topological or Hausdorff dimension since they involve only small subsets of X.

The difference between metric dimension and Assouad dimension can be dramatic:

XVII.6.7.22. EXAMPLES. (i) Let X be an infinite set with the discrete metric. Then $dim_A(X) = \infty$ but $dim_m(X) = 0$.

(ii) Let (X, ρ) be a metric space with ρ unbounded. Let $\tilde{\rho}$ be either $\min(\rho, 1)$ or $\frac{\rho}{\rho+1}$ (). Then $\dim_A(X, \tilde{\rho}) = \infty$, but $dim_m(X, \tilde{\rho}) = dim_m(X, \rho)$.

XVII.6.7.23. PROPOSITION. Let (X, ρ) be a metric space. Then

$$dim(X) \leq dim_H(X, \rho) \leq dim_P(X, \rho) \leq dim_m(X, \rho) \leq dim_A(X, \rho) \ .$$

Examples show that these can all be different.

XVII.6.7.24. EXAMPLE. The packing dimension of any countable metric space is 0. If (X, ρ) is the metric space of XVII.6.7.12., then

$$0 = dim_P(X, \rho) < dim_m(X, \rho) = \frac{1}{2} \ .$$

Relation with Topological Dimension

If (X, ρ) is a metric space, then $dim_A(X, \rho) \geq dim(X)$ (XVII.6.7.23.). Since $dim(X)$ is a topological invariant of X, we have $dim_A(X, \sigma) \geq dim(X)$ for every metric σ equivalent to ρ. Can we find a σ which gives equality (as we can for dim_H or dim_P ())?

XVII.6.7.25. THEOREM. [Luu98, 4.3] Let X be a separable metrizable (second countable regular) topological space. Then there is a totally bounded metric σ on X compatible with its topology, with

$$dim(X) = dim_H(X, \sigma) = dim_P(X, \sigma) = dim_m(X, \sigma) = dim_A(X, \sigma)$$

(such a metric is called *flat*).

The proof is a combination of XIII.19.5.15. with the construction of a flat metric on the Menger compactum.

BiLipschitz Embedding in Euclidean Space

XVII.6.7.26. If ρ is a metric on a set X, then so is ρ^q for $0 < q < 1$ (), and ρ^q is uniformly equivalent to ρ. If (X, ρ) is (C, s)-homogeneous, then (X, ρ^q) is $(C, s/q)$-homogeneous, and conversely; Thus

$$dim_A(X, \rho^q) = \frac{1}{q} dim_A(X, \rho) \ .$$

It makes sense that the fractal dimension of (X, ρ^q) should be larger than (X, ρ) since (X, ρ^q) is "more fractal" than (X, ρ). However, it follows that if (X, ρ) has finite Assouad dimension, so does (X, ρ^q).

XVII.6.7.27. THEOREM. ([Ass79], cf. [Luu98, A13]) Let (X, ρ) be a metric space. The following are equivalent:

(i) $dim_A(X, \rho) < \infty$.

(ii) For some q, $0 < q < 1$, there is an n and a biLipschitz embedding of (X, ρ^q) into $\mathbb{R}^n$.

(iii) For every q, $0 < q < 1$, there is an n and a biLipschitz embedding of (X, ρ^q) into $\mathbb{R}^n$.

XVII.6.7.28. There is an (X, ρ) with $dim_A(X, \rho) < \infty$ for which there is no biLipschitz embedding of (X, ρ) into a Euclidean space, cf. [Luu98, A15].

Other Fractal Dimensions

XVII.6.7.29. There are some other notions of fractal dimension generally intermediate between Hausdorff and packing dimension, e.g. upper and lower box-counting dimension and modified upper and lower box-counting dimensions (sometimes called *Minkowski dimensions*), which do not arise from measures and which are not as nicely behaved, and some others which are more subtle and variable. These are defined and discussed in [Fal14].

XVII.6.8. Radon Measures on Metric Spaces

XVII.6.9. Exercises

XVII.6.9.1. (a) Let ρ be a metric on $\mathbb{R}^n$ coming from a norm. Show that $\mathscr{H}_\rho^{(n)}$ is a scalar multiple $\alpha\mathscr{L}^{(n)}$, where α is the reciprocal of the Lebesgue measure of the unit ball for ρ.

(b) Show that unless ρ is a constant multiple of the usual Euclidean distance, $\mathscr{H}_\rho^{(s)}$ for $0 < s < n$ is not a constant multiple of $\mathscr{L}^{(s)}$. [For $s = 1$ consider horizontal and diagonal line segments.]

XVII.6.9.2. [Edg98] (a) If F is a closed set in $\mathbb{R}^n$ and $\mathscr{P}^{(n)}(F) = 0$, then F is nowhere dense.

(b) If $A \subseteq \mathbb{R}^n$ and $dim_P(A) < n$, fix s with $dim_P(A) < s < n$. Show that A can be covered by a sequence (F_k) of closed sets with $\mathscr{P}^{(s)}(F_k) < \infty$ for all k (in fact one can do much better!) [Use that if $B \subseteq \mathbb{R}^n$, then $\mathscr{P}_0^{(s)}(B) = \mathscr{P}_0^{(s)}(\bar{B})$ (XVII.6.6.5.).]

(c) Show that if $A \subseteq \mathbb{R}^n$ and $dim_P(A) < n$, then A is meager in $\mathbb{R}^n$.

(d) Let L be the set of Liouville numbers in $\mathbb{R}$. Then $dim_P(L) = 1$ although $dim_H(L) = 0$ (III.8.2.37.).

(e) If $A \subseteq \mathbb{R}^n$ and $\mathscr{P}^{(n)}(A) = 0$, is A necessarily meager in $\mathbb{R}^n$?

XVII.6.9.3. [Mor09, p. 19] **A Triangular Cantor Set.** In $\mathbb{R}^2$, let T_1 be a solid equilateral triangle of side 1. Let T_2 be the union of the three corner solid equilateral triangles of T_1 of side 1/3, and for each n, T_{n+1} the union of all solid corner triangles of the triangles in T_n of 1/3 the side length. Set $T = \cap_n T_n$. See Figure XVII.1 (the edge segments between the triangles are not part of T).
(a) Show that T is homeomorphic to the Cantor set. [Use ().] In particular, $dim(T) = 0$.

(b) Show that $\mathscr{L}^{(1)}(T) = 1$, and hence $dim_H(T) = 1$. [Use self-similarity.]

(c) Describe a homeomorphism of $\mathbb{R}^2$ sending T onto the usual Cantor set as a subset of the x-axis (cf. XIII.19.5.43.).

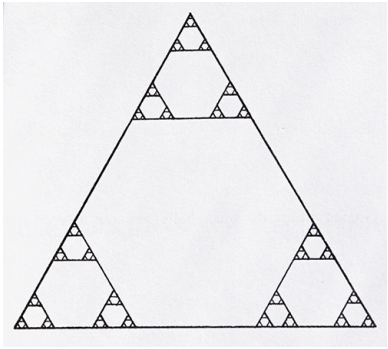


Figure XVII.1: A triangular Cantor set

XVII.7. Measures on Product Spaces

In this section, we will discuss a natural way of constructing a Cartesian product of two or more measure spaces.

XVII.7.1. Product σ-Algebras

XVII.7.1.1. Suppose $(X, \mathcal{A})$ and $(Y.\mathcal{B})$ are measurable spaces. If $A \subseteq X$ and $B \subseteq Y$, then $A \times B$ is called a *rectangle* in $X \times Y$. If $A \in \mathcal{A}$ and $B \in \mathcal{B}$, then $A \times B$ is called a *measurable rectangle* (technically, $(\mathcal{A}, \mathcal{B})$-*measurable rectangle*). It is natural that such subsets of $X \times Y$ should be regarded as measurable sets.

XVII.7.1.2. Definition. The *product σ-algebra* of $(\mathcal{A}, \mathcal{B})$, denoted $\mathcal{A} \otimes \mathcal{B}$, is the σ-algebra on $X \times Y$ generated by the measurable rectangles.

Although some authors use the notation $\mathcal{A} \times \mathcal{B}$ for the product σ-algebra, we will use the symbol $\otimes$ since $\mathcal{A} \otimes \mathcal{B}$ is much larger than just the set of measurable rectangles in general; we will use $\mathcal{A} \times \mathcal{B}$ to denote the set of measurable rectangles. (The symbol $\otimes$ is normally used for "tensor product" in abstract algebra; there is a close analogy between our use of the symbol here and algebraic tensor products.)

XVII.7.1.3. Examples.

(i) Let $X = Y = \{0, 1\}$, with $\mathcal{A} = \mathcal{B} = \mathcal{P}(\{0, 1\})$. Then $X \times Y = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Each singleton in $X \times Y$ is a (measurable) rectangle. The set $\{(0, 0), (1, 1)\}$ is not a (measurable) rectangle in $X \times Y$, but it is a union of two rectangles, hence is in $\mathcal{A} \otimes \mathcal{B}$; thus $\mathcal{A} \otimes \mathcal{B}$ is strictly larger than $\mathcal{A} \times \mathcal{B}$. In fact, $\mathcal{A} \otimes \mathcal{B} = \mathcal{P}(X \times Y)$.

(ii) If X and Y are sets, and at least one is countable, then it is easily checked that $\mathcal{P}(X) \otimes \mathcal{P}(Y) = \mathcal{P}(X \times Y)$ (Exercise ().) In particular, $\mathcal{P}(\mathbb{N}) \otimes \mathcal{P}(\mathbb{N}) = \mathcal{P}(\mathbb{N}^2)$. But this result does not hold if X and Y are of large cardinality (Exercise ()).

(iii) If $\mathcal{B}$ is the σ-algebra of Borel sets in $\mathbb{R}$, then $\mathcal{B} \otimes \mathcal{B} = \mathcal{B}^{(2)}$, the σ-algebra of Borel sets in $\mathbb{R}^2$. Neither inclusion is completely obvious. To show that $\mathcal{B} \otimes \mathcal{B} \subseteq \mathcal{B}^{(2)}$, it suffices to show that every measurable rectangle is in $\mathcal{B}^{(2)}$. But $\mathcal{B}_x = \{B \times \mathbb{R} : B \in \mathcal{B}\}$ is a σ-algebra which is generated by $\{U \times \mathbb{R} : U \text{ open}\}$, and $U \times \mathbb{R}$ is open in $\mathbb{R}^2$, hence in $\mathcal{B}^{(2)}$; thus $\mathcal{B}_x \subseteq \mathcal{B}^{(2)}$. Similarly, $B^y = \{\mathbb{R} \times B : B \in \mathcal{B}\} \subseteq \mathcal{B}^{(2)}$. So if $A, B \in \mathcal{B}$, then $A \times B = (A \times \mathbb{R}) \cap (\mathbb{R} \times B) \in \mathcal{B}^{(2)}$, so $\mathcal{B} \otimes \mathcal{B} \subseteq \mathcal{B}^{(2)}$. For the other inclusion, it suffices to show that every open set in $\mathbb{R}^2$ is in $\mathcal{B} \otimes \mathcal{B}$. But it is a simple exercise () that every nonempty open set in $\mathbb{R}^2$ is a (countable) union of open rectangles of the form $(a, b) \times (c, d)$ with $a, b, c, d \in \mathbb{Q}$. (This example generalizes to the Borel sets in any product of second countable topological spaces; see ().)

Although the set of measurable rectangles is not a σ-algebra or even an algebra because it is not closed under complements or finite unions (except in trivial cases), we have:

XVII.7.1.4. PROPOSITION. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces. Then the set

$$\mathcal{A} \times \mathcal{B} = \{A \times B : A \in \mathcal{A}, B \in \mathcal{B}\}$$

of measurable rectangles is a semialgebra.

PROOF: We have $X \times Y$ and $\emptyset = \emptyset \times \emptyset$ in $\mathcal{A} \times \mathcal{B}$. If $\{A_k \times B_k : 1 \leq k \leq n\}$ is a finite collection in $\mathcal{A} \times \mathcal{B}$, then

$$\bigcap_{k=1}^{n} (A_k \times B_k) = \left(\bigcap_{k=1}^{n} A_k \right) \times \left(\bigcap_{k=1}^{n} B_k \right) \in \mathcal{A} \times \mathcal{B}$$

so $\mathcal{A} \times \mathcal{B}$ is closed under finite intersections (in fact, the same argument shows that $\mathcal{A} \times \mathcal{B}$ is closed under countable intersections). And if $A \times B \in \mathcal{A} \times \mathcal{B}$, then

$$(A \times B)^c = (A \times B^c) \cup (A^c \times Y)$$

so $(A \times B)^c$ is a disjoint union of two sets in $\mathcal{A} \times \mathcal{B}$. 👍

XVII.7.1.5. PROPOSITION. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces. Let π_X and π_Y be the projections from $X \times Y$ to X and Y respectively, i.e. $\pi_X(x, y) = x$, $\pi_Y(x, y) = y$. Then

(i) π_X is an $(\mathcal{A} \otimes \mathcal{B}, \mathcal{A})$-measurable function and π_Y is an $(\mathcal{A} \otimes \mathcal{B}, \mathcal{B})$-measurable function.

(ii) If $f : Z \to (X \times Y)$ is a function, then f is $(\mathcal{C}, \mathcal{A} \otimes \mathcal{B})$-measurable if and only if $\pi_X \circ f : Z \to X$ is $(\mathcal{C}, \mathcal{A})$-measurable and $\pi_Y \circ f : Z \to Y$ is $(\mathcal{C}, \mathcal{B})$-measurable.

PROOF: (i): If $A \in \mathcal{A}$, then $\pi_X^{-1}(A) = A \times Y \in \mathcal{A} \otimes \mathcal{B}$, so π_X is measurable. The argument for π_Y is similar. (ii): If f is measurable, then $\pi_X \circ f$ and $\pi_Y \circ f$ are measurable by XVI.1.6.3.. Conversely, if $\pi_X \circ f$ is measurable, then $f^{-1}(A \times Y) = [\pi_X \circ f]^{-1}(A) \in \mathcal{C}$ for all $A \in \mathcal{A}$. Similarly, if $\pi_Y \circ f$ is measurable, then $f^{-1}(X \times B) \in \mathcal{C}$ for all $B \in \mathcal{B}$. Since $\{A \times Y : A \in \mathcal{A}\} \cup \{X \times B : B \in \mathcal{B}\}$ generates $\mathcal{A} \otimes \mathcal{B}$ as a σ-algebra, f is measurable by XVI.1.6.10.. 👍

XVII.7.2. Product Measures

Now suppose that $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ are measure spaces. We want to obtain a reasonable measure $\mu \times \nu$ on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$. It is natural to want the measure of a measurable rectangle $A \times B$ to be $\mu(A)\nu(B)$. It turns out that it is possible to obtain such a measure through the Caratheodory extension process.

XVII.7.2.6. THEOREM. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces. For $A \times B$ in the semialgebra $\mathcal{C}$ of measurable rectangles (XVII.7.1.4.), set $\iota(A \times B) = \mu(A)\nu(B)$ (where $0 \cdot \infty = \infty \cdot 0 = 0$ by convention). Then ι is a premeasure on $\mathcal{C}$, and hence ι extends to a measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ denoted $\mu \times \nu$ by the Caratheodory extension process. If μ and ν are σ-finite, then $\mu \times \nu$ is the unique extension of ι to a measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$, and is also σ-finite.

PROOF: We verify that ι satisfies (i), (ii′), and (iii) of XVII.4.5.2. and XVII.4.5.6. (note that (ii′) $\Rightarrow$ (ii) since $\emptyset \in \mathcal{A} \times \mathcal{B}$). Condition (i) is obvious. We prove (ii′) and (iii) simultaneously since the arguments are almost identical (we could simplify the argument slightly by using the somewhat nontrivial result XVII.4.5.6. to avoid proving (iii), but this is not necessary). Suppose $A_1 \times B_1, A_2 \times B_2, \cdots$ is a sequence of measurable rectangles and

$$\bigcup_{k=1}^{\infty} (A_k \times B_k) \supseteq A \times B \in \mathcal{A} \times \mathcal{B}$$

(for (ii′), assume the $A_k \otimes B_k$ are pairwise disjoint and $\cup_{k=1}^{\infty}(A_k \times B_k) = A \times B$). We must show $\mu(A)\nu(B) \leq \sum_{k=1}^{\infty} \mu(A_k)\nu(B_k)$, with equality for (ii′). We may assume A and B are nonempty. Set

$$\phi = \mu(A)\chi_B, \quad \psi = \sum_{k=1}^{\infty} \mu(A_k)\chi_{B_k}$$

(where the convention $\infty \cdot 0 = 0$ is followed). Then ϕ and ψ are extended real-valued measurable functions on $(Y, \mathcal{B})$. We claim $\phi \leq \psi$, with equality in (ii′). If $y \notin B$, then $\phi(y) = 0$, and in case (ii′) also $\psi(y) = 0$. If $y \in B$, set

$$I_y = \{k \in I : y \in B_k\} \subseteq \mathbb{N} .$$

If $x \in A$, then $(x, y) \in A \times B$, so $(x, y) \in A_k \times B_k$ for at least one k (exactly one k in case (ii′)), i.e. $x \in A_k$ for at least one $k \in I_y$ (exactly one in case (ii′)). Thus A is contained in the union of $\{A_k : k \in I_y\}$. In case (ii′), if $x \notin A$, then for every $k \in I_y$ $(x, y) \notin A_k \times B_k$; since $y \in B_k$, we have $x \notin A_k$ for every $k \in I_y$, i.e. A is the disjoint union of $\{A_k : k \in I_y\}$. Thus

$$\phi(y) = \mu(A) \leq \sum_{k \in I_y} \mu(A_k) = \sum_{k \in I_y} \mu(A_k)\chi_{B_k}(y)$$

$$= \sum_{k \in \mathbb{N}} \mu(A_k)\chi_{B_k}(y) = \psi(y)$$

(with equality in case (ii′)), where the first equality in the last line is because the additional terms are zero. Thus $\phi \leq \psi$, with equality in case (ii′). So

$$\mu(A)\nu(B) = \int_Y \phi \, d\nu \leq \int_Y \psi \, d\nu = \int_Y \left[\sum_{k=1}^{\infty} \mu(A_k)\chi_{B_k} \right] d\nu$$

$$= \sum_{k=1}^{\infty} \mu(A_k) \left[\int_Y \chi_{B_k} \, d\nu \right] = \sum_{k=1}^{\infty} \mu(A_k)\nu(B_k)$$

(with equality in case (ii′)), where interchange of the sum and integral is justified by the MCT (XVIII.1.5.5.). Thus ι satisfies (ii′) and (iii) and is a premeasure.

The other conclusions then follow from XVII.4.5.8.. 👍

XVII.7.2.7. DEFINITION. The measure $\mu \times \nu$ is called the *product measure* of μ and ν.

Completeness

XVII.7.2.8. There is a slight ambiguity in XVII.7.2.7.. By the Extension Theorem, the σ-algebra $\mathcal{M}$ of $(\mu \times \nu)$-measurable sets contains the σ-algebra $\mathcal{A} \otimes \mathcal{B}$. But $(X \times Y, \mathcal{M}, \mu \times \nu)$ is a complete measure space and $\mathcal{M}$ is strictly larger than $\mathcal{A} \otimes \mathcal{B}$ in general. By slight abuse of notation, we will write $\mu \times \nu$ for the measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ and $\overline{\mu \times \nu}$ for the measure on $\mathcal{M}$, and we will write $\mathcal{A} \bar{\otimes} \mathcal{B}$ for $\mathcal{M}$; we will sometimes use the notation $\mu \times \nu$ for both the measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ and its extension on $(X \times Y, \mathcal{M})$ when a distinction is unimportant.

If μ and ν are σ-finite, then $(X \times Y, \mathcal{M}, \overline{\mu \times \nu})$ is precisely the completion of $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ (XVII.4.5.8.). If $(X, \bar{\mathcal{A}}, \bar{\mu})$ and $(Y, \bar{\mathcal{B}}, \bar{\nu})$ are the completions of $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ respectively, then the outer measures induced by ι on $\mathcal{A} \times \mathcal{B}$ and $\bar{\iota}$ on $\bar{\mathcal{A}} \times \bar{\mathcal{B}}$ coincide (Exercise XVIII.3.3.2.); so if $\bar{\mathcal{M}}$ is the σ-algebra of $(\bar{\mu} \times \bar{\nu})$-measurable sets, then $\bar{\mathcal{M}} = \mathcal{M}$. Thus in the σ-finite case we can also identify the completion of $(X \times Y, \bar{\mathcal{A}} \otimes \bar{\mathcal{B}}, \bar{\mu} \times \bar{\nu})$ with $(X \times Y, \mathcal{M}, \overline{\mu \times \nu})$, i.e. the completions of $(X \times Y, \mathcal{A} \otimes \mathcal{B}, \mu \times \nu)$ and $(X \times Y, \bar{\mathcal{A}} \otimes \bar{\mathcal{B}}, \bar{\mu} \times \bar{\nu})$ coincide, and in particular $\mathcal{A} \bar{\otimes} \mathcal{B} = \bar{\mathcal{A}} \bar{\otimes} \bar{\mathcal{B}}$.

Products of Finitely Many Spaces

XVII.7.2.9. The results about products of two measure spaces can easily be generalized to products of finitely many. If $(X_1, \mathcal{A}_1), \cdots, (X_n, \mathcal{A}_n)$ are measurable spaces, then a set of the form $A_1 \times \cdots \times A_n$ with $A_i \subseteq X_i$, is called a *rectangle*, a *measurable rectangle* if $A_i \in \mathcal{A}_i$ for each i. Then $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$ is the σ-algebra generated by the measurable rectangles in $X_1 \times \cdots \times X_n$. If $I = \{1, \ldots, n\}$, $\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n$ is often denoted $\bigotimes_{i \in I} \mathcal{A}_i$ or $\bigotimes_{i=1}^{n} \mathcal{A}_i$.

The next proposition shows that the product σ-algebra can be defined also by successive products of two σ-algebras, using any convenient grouping:

XVII.7.2.10. Proposition. Let $(X, \mathcal{A})$, $(Y, \mathcal{B})$, and $(Z, \mathcal{C})$ be measurable spaces. If $X \times Y \times Z$ is identified with $(X \times Y) \times Z$ and $X \times (Y \times Z)$ in the standard way, we have $\mathcal{A} \otimes \mathcal{B} \otimes \mathcal{C} = (\mathcal{A} \otimes \mathcal{B}) \otimes \mathcal{C} = \mathcal{A} \otimes (\mathcal{B} \otimes \mathcal{C})$.

The proof is routine, and is left as an exercise.

XVII.7.2.11. Corollary. Let $(X_1, \mathcal{A}_1), \ldots, (X_n, \mathcal{A}_n)$ be measurable spaces, and let I_1, I_2 be disjoint subsets of $I = \{1, \ldots, n\}$ with $I = I_1 \cup I_2$. If $\prod_{i \in I} X_i$ is identified with $\left(\prod_{i \in I_1} X_i\right) \times \left(\prod_{i \in I_2} X_i\right)$ in the standard way, then

$$\bigotimes_{i \in I} \mathcal{A}_i = \left(\bigotimes_{i \in I_1} \mathcal{A}_i\right) \otimes \left(\bigotimes_{i \in I_2} \mathcal{A}_i\right).$$

There is then an analog of XVII.7.1.5.. This can be proved directly in the same way as XVII.7.1.5., or by repeated application of XVII.7.1.5. and XVII.7.2.10..

XVII.7.2.12. Theorem. Let $(X_1, \mathcal{A}_1), \ldots, (X_n, \mathcal{A}_n)$ and $(Z, \mathcal{C})$ be measurable spaces. Let π_k $(1 \leq k \leq n)$ be the projections from $X_1 \times \cdots \times X_n$ to X_k, i.e. $\pi_k(x_1, \ldots, x_n) = x_k$. Then

(i) π_k is an $(\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n, \mathcal{A}_k)$-measurable function.

(ii) If $f : Z \to (X_1 \times \cdots \times X_n)$ is a function, then f is $(\mathcal{C}, \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_n)$-measurable if and only if $\pi_k \circ f : Z \to X$ is $(\mathcal{C}, \mathcal{A}_k)$-measurable for each k, $1 \leq k \leq n$.

We also have the analog of XVII.7.1.4., which can also either be proved directly in the same way as XVII.7.1.4. or by repeated application of XVII.7.1.4. and XVII.7.2.10..

XVII.7.2.13. PROPOSITION. Let $(X_1, \mathcal{A}_1), \ldots, (X_n, \mathcal{A}_n)$ be measurable spaces. Then the set $\mathcal{C} = \{A_1 \times \cdots \times A_n : A_i \in \mathcal{A}_i\}$ of measurable rectangles is a semialgebra.

XVII.7.2.14. If $(X_1, \mathcal{A}_1, \mu_1), \cdots, (X_n, \mathcal{A}_n, \mu_n)$ are measure spaces, we can form a product measure $\mu_1 \times \cdots \times \mu_n$ on $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)$ such that $(\mu_1 \times \cdots \times \mu_n)(A_1 \times \cdots \times A_n) = \mu_1(A_1)\mu_2(A_2)\cdots\mu_n(A_n)$ for any $A_i \in \mathcal{A}_i$. The proof of the next theorem is an adaptation of the proof of () and is left as an exercise ().

XVII.7.2.15. THEOREM. Let $\{(X_i, \mathcal{A}_i, \mu_i) : 1 \leq i \leq n\}$ be measure spaces. For $\prod_{i=1}^n A_i$ in the semialgebra $\mathcal{C}$ of measurable rectangles (XVII.7.2.13.), set $\tau(\prod_{i=1}^n A_i) = \mu_1(A_1)\cdots\mu_n(A_n)$. Then τ is a premeasure on $\mathcal{C}$, and hence τ extends to a measure on $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)$ denoted $\mu_1 \times \cdots \times \mu_n$ or $\times_{i=1}^n \mu_i$, by the Caratheodory extension process. If all μ_i are σ-finite, then $\times_{i=1}^n \mu_i$ is the unique extension of τ to a measure on $(\prod_{i=1}^n X_i, \bigotimes_{i=1}^n \mathcal{A}_i)$.

The next fact is immediate from the uniqueness in the σ-finite case, but holds in general. The proof in the general case is routine but somewhat tedious, and the result in the non-σ-finite case is of limited interest, so we omit the details.

XVII.7.2.16. PROPOSITION. Let $(X_1, \mathcal{A}_1, \mu_1), \ldots, (X_n, \mathcal{A}_n, \mu_n)$ be measure spaces, and let I_1, I_2 be disjoint subsets of $I = \{1, \ldots, n\}$ with $I = I_1 \cup I_2$. If $\prod_{i \in I} X_i$ is identified with $\left(\prod_{i \in I_1} X_i\right) \times \left(\prod_{i \in I_2} X_i\right)$ in the standard way, then

$$\times_{i \in I} \mu_i = (\times_{i \in I_1} \mu_i) \times (\times_{i \in I_2} \mu_i) .$$

XVII.7.3. Lebesgue Measure on $\mathbb{R}^n$

XVII.7.4. Infinite Product Spaces

It is also possible to define an infinite Cartesian product of measure spaces provided that all (or all but finitely many) of the measures are probability measures.

XVII.7.4.1. Suppose that $\{(X_i, \mathcal{A}_i) : i \in I\}$ is an indexed collection of measurable spaces. A *rectangle* in $\prod_i X_i$ is a finite set $I_0 = \{i_1, \ldots, i_n\}$ of coordinates, and for each i_k a subset A_{i_k} of X_{i_k}. The corresponding set

$$\{(\cdots x_i \cdots) \in \prod_i X_i : x_{i_k} \in A_{i_k} \text{ for all } k, 1 \leq k \leq n\}$$

is often denoted $A_{i_1} \times \cdots \times A_{i_n}$, although it should more correctly be denoted $\prod_{i \in I_0} A_i \times \prod_{i \notin I_0} X_i$. Such a rectangle is a *measurable rectangle* if each A_{i_k} is in $\mathcal{A}_{i_k}$.

If A_i is a subset of X_i for each i, then the subset

$$\prod_i A_i = \{(\cdots x_i \cdots) \in \prod_i X_i : x_i \in A_i \text{ for all } i \in I\}$$

is called a *box*, and, if each $A_i \in \mathcal{A}_i$, a *measurable box*. Note that a [measurable] box is not a [measurable] rectangle in general if I is infinite, unless $A_i = X_i$ for all but finitely many i.

XVII.7.4.2. DEFINITION. The *product σ-algebra* $\bigotimes_{i \in I} \mathcal{A}_i$ is the σ-algebra on $\prod_i X_i$ generated by the measurable rectangles.

XVII.7.4.3. PROPOSITION. If I is countable, then every measurable box is in $\bigotimes_i \mathcal{A}_i$.

XVII.7.4.4. This statement is false in general if I is uncountable. We say that a subset A of $\prod_i X_i$ *depends only on the coordinates in $J \subseteq I$* if, whenever $x = (\cdots x_i \cdots)$ and $y = (\cdots y_i \cdots)$ are elements of $\prod_i X_i$, and $x_j = y_j$ for all $j \in J$, then $x \in A$ if and only if $y \in A$. A subset A of $\prod_i X_i$ *depends on finitely many* [*countably many*] *coordinates* if it depends only on some finite [countable] set of coordinates. Thus every rectangle depends only on finitely many coordinates.

XVII.7.4.5. PROPOSITION. Every set in $\bigotimes_i \mathcal{A}_i$ depends only on countably many coordinates.

PROOF: It is easily checked that the collection of all sets which depend on only countably many coordinates is a σ-algebra. 👍

XVII.7.4.6. COROLLARY. A nonempty measurable box $\prod_i A_i$ is in $\bigotimes_i \mathcal{A}_i$ if and only if $A_i = X_i$ for all but countably many i.

Infinite product measures are particularly important in probability. The next example is a good illustration of an infinite product measure which has a nice probabilistic interpretation.

XVII.7.4.7. EXAMPLE. Let $X = \{0,1\}$, and for each n let $(X_n, \mathcal{A}_n) = (X, \mathcal{P}(X))$. Put a measure μ on $(X, \mathcal{P}(X))$ by $\mu(\{0\}) = \mu(\{1\}) = 1/2$, and let μ_n be the corresponding measure on $(X_n, \mathcal{A}_n)$. Let $X^{\mathbb{N}} = \prod_{n\in\mathbb{N}} X_n$, $\mathcal{A}^{\mathbb{N}} = \bigotimes_{n\in\mathbb{N}} \mathcal{A}_n$, and $\mu^{\mathbb{N}} = \times_{n\in\mathbb{N}}\mu_n$. $(X, \mathcal{P}(X), \mu)$ is not a very interesting or complicated measure space, but $(X^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, \mu^{\mathbb{N}})$ is quite interesting and nontrivial. Elements of $X^{\mathbb{N}}$ are, of course, all sequences of 0's and 1's.

If instead of $\{0,1\}$ we take $X = \{H,T\}$, we may interpret $(X, \mathcal{P}(X), \mu)$ as the sample space of the experiment of tossing a fair coin once. Then $(X^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}}, \mu^{\mathbb{N}})$ can be interpreted as the sample space of the experiment of tossing a fair coin infinitely many times. Probabilists use this measure space to analyze such things as the likelihood of long runs and other details in a sequence of coin tosses. This is a good example of a *stochastic process*, the study of which is an entire branch of probability theory.

This example also has other interpretations not closely connected with probability. The product measure can be interpreted as a measure on the Cantor set () or as Haar measure on a certain compact topological group (). There is also a close connection with Lebesgue measure on the unit interval ().

The example could be modified by taking $\mu_\alpha(\{0\}) = \alpha$, $\mu_\alpha(\{1\}) = 1-\alpha$ for some fixed α, $0 < \alpha < 1$ (the original μ is $\mu_{1/2}$.) The measure μ_α could be interpreted probabilistically (using $\{H,T\}$ for $\{0,1\}$) as tossing an unfair coin with probability of heads equal to α. It turns out that the $\mu_\alpha^{\mathbb{N}}$ are mutually singular continuous probability measures on $(X^{\mathbb{N}}, \mathcal{A}^{\mathbb{N}})$ as α ranges over $(0,1)$ (). One could even similarly define μ_0 and μ_1; $\mu_0^{\mathbb{N}}$ is then the point mass at the constant sequence of 0's, and $\mu_1^{\mathbb{N}}$ the point mass at the constant sequence of 1's.

XVII.7.4.8. One might try to construct the infinite product of $\{(X_i, \mathcal{A}_i, \mu_i) : i \in I\}$ (where $\mu_i(X_i) = 1$ for all i) as follows. For each finite subset F of I, let $\mathcal{A}_F$ be the σ-algebra on $X = \prod_i X_i$ generated by the measurable rectangles depending only on F, and μ_F the measure on $(X, \mathcal{A}_F)$ which is effectively $\times_{i\in F}\mu_i$ using the obvious identification of $\mathcal{A}_F$ with $\otimes_{i\in F}\mathcal{A}_i$ on $\prod_{i\in F} X_i$. Then $\mathcal{B} = \cup_F \mathcal{A}_F$ is an algebra of subsets of X, and the μ_F give a well-defined function ι from $\mathcal{B}$ to $[0,1]$. The σ-algebra generated by $\mathcal{B}$ is $\otimes_{i\in I}\mathcal{A}_i$, and $\times_i\mu_i$ is an extension of ι to $\otimes_{i\in I}\mathcal{A}_i$, so one might try to obtain $\times_{i\in I}\mu_i$ directly from ι by the Caratheodory extension process.

The difficulty in this approach is showing that ι is a premeasure on $\mathcal{B}$. This is not automatic: it is not true in general that if $(X, \mathcal{A}_n, \nu_n)$ is a compatible family of probability measures on an increasing sequence $(\mathcal{A}_n)$ of σ-algebras, that the function ι defined by the ν_n on the algebra $\cup_n\mathcal{A}_n$ is a premeasure. For example, let $X = \mathbb{N}$, $\mathcal{A}_n$ the σ-algebra of all subsets of $\{1,\dots,n\}$ and their complements in $\mathbb{N}$, and ν_n the measure on $(\mathbb{N}, \mathcal{A}_n)$ with $\nu_n(A) = 0$ if $A \subseteq \{1,\dots,n\}$ and $\nu_n(A) = 1$ if $\mathbb{N} \setminus A \subseteq \{1,\dots,n\}$. Then the function ι is not a premeasure on $\cup_n\mathcal{A}_n$.

XVII.7.5. Exercises

XVII.7.5.1. Let X and Y be sets, and $E \subseteq X \times Y$. For each $x \in X$, set

$$E_x = \{y \in Y : (x,y) \in E\} \subseteq Y$$

$$\mathcal{S}_X(E) = \{E_x : x \in X\} \subseteq \mathcal{P}(Y)\ .$$

$\mathcal{S}_X(E)$ is the set of "vertical cross-sections" of the set E. The *X-variety* of E is $var_X(E) = card(\mathcal{S}_X(E))$, the "number" of different sets occurring as vertical cross sections of E. Similarly, for $y \in Y$ define

$$E^y = \{x \in X : (x,y) \in E\} \subseteq X$$

$$\mathcal{S}_Y(E) = \{E^y : y \in Y\} \subseteq \mathcal{P}(X)$$

and the *Y-variety* of E to be $var_Y(E) = card(\mathcal{S}_Y(E))$.

(a) Prove that $\mathcal{C}_X = \{E \subseteq X \times Y : var_X(E) \leq 2^{\aleph_0}\}$ is a σ-algebra of subsets of $X \times Y$ containing $\mathcal{P}(X) \otimes \mathcal{P}(Y)$. Similarly, $\mathcal{C}_Y = \{E \subseteq X \times Y : var_Y(E) \leq 2^{\aleph_0}\}$ is a σ-algebra containing $\mathcal{P}(X) \otimes \mathcal{P}(Y)$.

(b) If $card(X) > 2^{\aleph_0}$ and $card(Y) \geq 2^{\aleph_0}$, then $\mathcal{P}(X) \otimes \mathcal{P}(Y) \neq \mathcal{P}(X \times Y)$. [Let $\kappa = \min(card(X), 2^{2^{\aleph_0}})$. Construct an $E \subseteq X \times Y$ with $var_X(E) = \kappa$. (What is really needed is a function from X to $\mathcal{P}(Y)$ whose range has cardinality greater than $2^{\aleph_0}$; AC is needed in general to insure that $card(X)$ is comparable with $2^{2^{\aleph_0}}$.)]

(c) Is it true that $\mathcal{P}(X) \otimes \mathcal{P}(Y) = \mathcal{C}_X \cap \mathcal{C}_Y$ for all X, Y?

(d) Is $\mathcal{P}(\mathbb{R}) \otimes \mathcal{P}(\mathbb{R}) = \mathcal{P}(\mathbb{R}^2)$? [Unfair question: it cannot be answered in ZFC! The CH, or, more generally, Martin's Axiom (II.10.5.6.), implies a positive answer, but (if ZFC is consistent) there are models of ZFC where the answer is negative. See ().]

XVII.7.5.2. ([GP10], attributed to von Neumann) Let $R = \prod_{k=1}^n [a_k, b_k]$ be a closed rectangle in $\mathbb{R}^n$. Note that $\lambda(R) \leq \prod_{k=1}^n (b_k - a_k)$ is immediate from the definition. The aim of this problem is to show the opposite inequality.

(a) Suppose $R_1, \ldots, R_m$ are open rectangles covering R (note that any cover of R by open rectangles has a finite subcover). Let $R_j = \prod_{k=1}^n (a_k^{(j)}, b_k^{(j)})$. If $S \subseteq \mathbb{R}^n$, let $\#(S)$ be the number of points with integer coordinates in S. If $\alpha > 0$ is large enough that $\alpha(b_k - a_k) > 1$ and $\alpha(b_k^{(j)} - a_k^{(j)}) > 1$ for all k and j, then

$$\prod_{k=1}^n (\alpha b_k - \alpha a_k - 1) \leq \#(\alpha R) \leq \sum_{j=1}^m \#(\alpha R_j) \leq \sum_{j=1}^m \prod_{k=1}^n (\alpha b_k^{(j)} - \alpha a_k^{(j)} + 1) \ .$$

(b) Let $\alpha \to +\infty$ to conclude that

$$\prod_{k=1}^n (b_k - a_k) \leq \sum_{j=1}^m \prod_{k=1}^n (b_k^{(j)} - a_k^{(j)}) \ .$$

Taking the infimum over all such open covers, conclude that

$$\prod_{k=1}^n (b_k - a_k) \leq \lambda(R) \ .$$

XVII.7.5.3. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, with completions $(X, \bar{\mathcal{A}}, \bar{\mu})$ and $(Y, \bar{\mathcal{B}}, \nu)$ respectively, and let ι and $\bar{\iota}$ be the induced premeasures on $\mathcal{A} \times \mathcal{B}$ and $\bar{\mathcal{A}} \times \bar{\mathcal{B}}$, i.e. $\iota(A \times B) = \mu(A)\mu(B)$ for $A \in \mathcal{A}$, $B \in \mathcal{B}$ and $\bar{\iota}(C \times D) = \bar{\mu}(C)\bar{\nu}(D)$ for $C \in \bar{\mathcal{A}}$, $D \in \bar{\mathcal{B}}$. Let θ^* and $\bar{\theta}^*$ be the outer measures on $X \times Y$ induced by ι and $\bar{\iota}$ respectively.

(a) If $C \in \bar{\mathcal{A}}$ and $D \in \bar{\mathcal{B}}$, show that there are $A \in \mathcal{A}$ and $B \in \mathcal{B}$ such that $C \subseteq A$, $D \subseteq B$, and $\iota(A \times B) = \bar{\iota}(C \times D)$.

(b) Show that $\bar{\theta}^* = \theta^*$.

XVII.8. The Besicovitch Covering Theorem

The Besicovitch Covering Theorem is a more flexible version of the Vitali Covering Theorem () which applies to arbitrary locally finite Borel measures on subsets of $\mathbb{R}^n$ and, more generally, on metric spaces which have a certain "finite-dimensionality."

This theorem was first stated in [Bes45] for $\mathbb{R}^2$ and soon extended to $\mathbb{R}^n$ [Mor47], and a very general version was given (somewhat cryptically) in [Fed69]. Our treatment is adapted from [Fed69], [FL94], and [Mor09]. Other relevant references include [KP86], [EoGKP88], [Pre83], and [Sul94] (there is a large additional literature on this and related matters). Our main interest is the $\mathbb{R}^n$ case, but there is no additional work in proving a very general version, even more general than the one in [Fed69]. The results are related to sphere-packing problems and sphere-of-influence graphs (cf. [FL94], [Sul94], [MQ94].

XVII.8.1. Controlled Metrics

We fix a metric space (X, ρ). We really want to work with closed balls in (X, ρ), but we must refer to the center and radius of a ball, neither of which is unique in general, so instead we work with pairs $(x, r) \in X \times (0, \infty)$, where (x, r) corresponds to the closed ball $\bar{B}_r(x)$ (i.e. to each (x, r) we associate a closed ball with specified center and radius). In many cases (e.g. in $\mathbb{R}^n$) this correspondence is a bijection.

XVII.8.1.1. Definition. If $\tau \geq 1$, a subset $\mathcal{P} = \{(x_i, r_i) : i \in I\}$ of $X \times (0, \infty)$ is *τ-controlled* if, for every $i \neq j$, either $\rho(x_i, x_j) > r_i \geq r_j/\tau$ or $\rho(x_i, x_j) > r_j \geq r_i/\tau$.

If $0 < R \leq \infty$ and $\tau \geq 1$, and $(x, r) \in X \times (0, \infty)$ with $r \leq R$, an *(R, τ)-controlled expansion* of (x, r) is a τ-controlled set $\mathcal{P} = \{(x_i, r_i) : i \in I\}$ such that

(i) $r/\tau \leq r_i \leq R$ for all $i \in I$.

(ii) For each $i \in I$, $\bar{B}_{r_i}(x_i) \cap \bar{B}_r(x) \neq \emptyset$.

(iii) $x \notin \bar{B}_{r_i}(x_i)$ (i.e. $r_i < \rho(x, x_i)$) for all $i \in I$.

A *strictly controlled expansion* of $(x, r) \in X \times (0, \infty)$ is a 1-controlled expansion in which $r_i = r$ for all i.

The metric ρ is *(R, τ)-controlled* ($\tau \geq 1$) if there is an $M \in \mathbb{N}$ such that, whenever $(x, r) \in X \times (0, \infty)$, $r < R$, and $\mathcal{C} = \{(x_i, r_i) : i \in I\}$ is an (R, τ)-controlled expansion of (x, r), then $card(I) \leq M - 1$. The metric ρ is *controlled* if it is (R, τ)-controlled for some $R > 0$ and $\tau > 1$.

The metric ρ is *strictly controlled* if there is an R and $M \in \mathbb{N}$ such that, whenever $(x, r) \in X \times (0, \infty)$, $r < R$, and $\mathcal{C} = \{(x_i, r) : i \in I\}$ is a strictly controlled expansion of (x, r), then $card(I) \leq M$.

XVII.8.1.2. The definition of a controlled metric is a little difficult to digest. Strictly controlled is easier to understand: it means that for any closed ball B in X of (specified) radius $r \leq R$, there are at most M closed balls of (specified) radius r intersecting B, including B, with no ball containing the (specified) center of any other ball. The variations with $\tau > 1$ and allowing the r_i to be greater than r are needed for technical reasons. ("Strictly controlled" should perhaps be called "strictly 1-controlled"; one could also make a definition of a strictly τ-controlled metric for $\tau > 1$, but this does not seem to be useful.)

Being controlled is a metric property. Even replacing a metric with a uniformly equivalent metric need not preserve control (Exercise XVII.8.4.8.).

XVII.8.1.3. PROPOSITION. (i) Any subspace of a controlled metric space is controlled (with the restriction metric).

(ii) A metric space is controlled (resp. strictly controlled) if and only if its completion is controlled (resp. strictly controlled).

Although checking that a metric is controlled can be difficult, there are important special cases which cover most situations of interest.

XVII.8.1.4. EXAMPLE. Let (X, ρ) be an ultrametric space (). Then any for any two closed balls, either the balls are disjoint or one is contained in the other. An ultrametric space is $(\infty, 1)$-controlled, with $M = 1$. But if $\tau > 1$, the situation is dicier: a τ-controlled expansion P of (x, r) would consist of a set of "almost disjoint" balls of radius greater than r/τ contained in $\bar{B}_r(x)$. These need not be bounded in number (XVII.8.4.8.). However, the p-adic numbers $\mathbb{Q}_p$, or any finite-dimensional ultranormed vector space over $\mathbb{Q}_p$, is (∞, τ)-controlled with $M = 1$ for $\tau > 1$ sufficiently close to 1.

XVII.8.1.5. If $\tau \geq \tau' \geq 1$, an (R, τ)-controlled metric is (R, τ')-controlled for any R, and is strictly controlled. The main difference between "controlled" and "strictly controlled," besides the presence of $\tau > 1$ (which is relatively harmless (XVII.8.1.12.)), is that for controlled expansions the r_i are allowed to be much larger than r, and there is no control (pun intended) over the ratios r_i/r and thus no ball of controlled size containing all the (x_i, r_i). This turns out to be a considerable technical difficulty (cf. Exercise XVII.8.4.4.) and can only be overcome if the metric space has enough "approximately interpolating" points.

The most important example is:

XVII.8.1.6. THEOREM. Let $(X, \|\cdot\|)$ be a finite-dimensional normed vector space, and ρ the induced metric on X. Then ρ is controlled.

Even if one is only interested in the case $X = \mathbb{R}^n$, one must essentially prove this theorem as part of the (standard) proof of the Besicovitch Covering Theorem; in fact, proving this theorem is arguably the most delicate part of the proof of the Besicovitch Covering Theorem. The proof relates this problem to the sphere-packing problem, and thus is geometric in nature. Our exposition of the proof is adapted from [FL94]. See [Sul94] for an alternate argument for $\mathbb{R}^n$ based on the Law of Cosines, although the reduction to $\tau = 1$ there is incomplete and flawed (the remark on p. 200 contains an incorrect statement and an inadequately justified statement; the paper does not have a proof that $\beta_n(\epsilon) < \infty$ for any $\epsilon > 0$). See [Fed69, 2.8] for another argument (cf. Exercise XVII.8.4.2.).

XVII.8.1.7. For the proof, we introduce some notation. For $\tau \geq 1$ let $M(X, \tau)$ be the smallest M such that every (∞, τ)-controlled expansion of a (closed) ball in X has cardinality $\leq M - 1$. By homogeneity and translation-invariance, $M(X, \tau)$ is the smallest M such that every (∞, τ)-controlled expansion of the closed unit ball in X has cardinality $\leq M - 1$. To prove the theorem we must show that $M(X, \tau)$ is finite for some $\tau > 1$. $M(X, \tau)$ is an increasing function of τ. Set $M(X) = M(X, 1)$.

XVII.8.1.8. For technical reasons, we slightly modify the definition of a controlled expansion, replacing all strict inequalities with nonstrict ones. If $\tau \geq 1$, a *τ-satellite configuration* is a set of closed balls $\{(x_i, r_i) : 0 \leq i\}$ with each $r_i \geq 1/\tau$, $x_0 = 0$, $r_0 = 1$, such that $\bar{B}_{r_i}(x_i) \cap \bar{B}_1(0) \neq \emptyset$ for all i, $\rho(x_i, 0) = \|x_i\| \geq \max(r_i, 1)/\tau$ and, for $i \neq j$, either $\rho(x_i, x_j) = \|x_i - x_j\| \geq r_i \geq r_j/\tau$ or $\rho(x_i, x_j) \geq r_j \geq r_i/\tau$. Let $\tilde{M}(X, \tau)$ be the maximum cardinality of a τ-satellite configuration. A τ-controlled expansion of the unit ball, along with the unit ball itself, is a τ-satellite configuration; thus $M(X, \tau) \leq \tilde{M}(X, \tau)$. $\tilde{M}(X, \tau)$ is an increasing function of τ. Set $\tilde{M}(X) = \tilde{M}(X, 1)$.

XVII.8.1.9. We also consider the sphere-packing constant $\theta(X)$, the largest number of points of distance at least 1 apart which can be put into the closed ball of radius 2 around the origin in X with one of the points at the origin. If we consider the closed balls of radius 1 around such a collection, we obtain a 1-satellite configuration; thus $\theta(X) \leq \tilde{M}(X)$.

XVII.8.1.10. PROPOSITION. If X is a normed vector space of dimension d, then $\theta(X) \leq 5^d$ (and in particular is finite).

PROOF: If $\{x_i : 1 \leq i \leq m\}$ is a set of points of distance ≥ 1 apart in $\bar{B}(0, 2)$, then the open balls of radius $1/2$ around the x_i form a pairwise disjoint set of balls of radius $1/2$ in the ball of radius $5/2$ around 0, of cardinality m. Identifying X with $\mathbb{R}^d$ and taking λ to be Lebesgue measure on $\mathbb{R}^d$, we have

$$m\lambda(B_{1/2}(0)) \leq \lambda(B_{5/2}(0)) = 5^d \lambda(B_{1/2}(0)) \ .$$



XVII.8.1.11. PROPOSITION. Let $(X, \|\cdot\|)$ be a finite-dimensional normed vector space. Then there is an ϵ, $0 < \epsilon < 1$, such that $\theta(X)$ is the largest number of points of distance $\geq 1 - \epsilon$ apart which can be put into the closed ball $B = \bar{B}_2(0)$, with one point at the origin.

PROOF: Suppose not. Set $m = \theta(X) + 1$. Then for each n there are points $x_1^{(n)}, \ldots, x_m^{(n)}$ in B of distance $\geq 1 - \frac{1}{n}$ apart, with $x_1^{(n)} = 0$. By compactness of B, passing to subsequences we may assume that for each i, $x_i^{(n)} \to x_i$ for some $x_i \in B$. Then the x_i are a set of points of B of distance ≥ 1 apart, of cardinality m, with $x_1 = 0$, a contradiction. 

Theorem XVII.8.1.6. is a corollary of the following result:

XVII.8.1.12. THEOREM. Let $(X, \|\cdot\|)$ be a finite-dimensional normed vector space. Let $\epsilon > 0$ be as in XVII.8.1.11.. Then, if $1 \leq \tau < 1 + \frac{\epsilon}{4}$, we have

$$\theta(X) = \tilde{M}(X) = \tilde{M}(X, \tau) \ .$$

In particular, $M(X, \tau) \leq \theta(X) \leq 5^{dim(X)} < \infty$.

PROOF: By XVII.8.1.9., we only need to show that $\tilde{M}(X,\tau) \leq \theta(X)$. To understand the proof, first consider the case where $\tau = 1$ (1-satellite configuration), which gives the inequality $\tilde{M}(X) \leq \theta(X)$. The general proof needs some estimates to handle the case where $\tau > 1$. The two estimates needed are

$$1 - \frac{1}{\tau} = \frac{\tau - 1}{\tau} \leq \tau - 1 < \frac{\epsilon}{4} < \epsilon$$

$$2\tau^2\left(1 - \frac{1}{\tau}\right) \leq 2\tau^2(\tau - 1) < 2\left(1 + \frac{\epsilon}{4}\right)^2 \frac{\epsilon}{4} < 2\left(\frac{5}{4}\right)^2 \frac{\epsilon}{4} = \frac{25}{32}\epsilon < \epsilon$$

since $\epsilon < 1$.

Suppose $\{(x_i, r_i) : 0 \leq i \leq m\}$ is a τ-satellite configuration. For each i set

$$y_i = \begin{cases} x_i & \text{if} \quad \|x_i\| \leq 2 \\ \frac{2}{\|x_i\|} x_i & \text{if} \quad \|x_i\| > 2 \end{cases}$$

i.e. the x_i outside the ball of radius 2 around 0 are radially projected onto the sphere of radius 2. We claim that $\|y_i - y_j\| \geq 1 - \epsilon$ for all $i \neq j$. This is obvious if $\|x_i\|$ and $\|x_j\|$ are ≤ 2.

If $\|x_i\| > 2$, $(y_i, 1) \subseteq (x_i, r_i)$ since $(x_i, r_i) \cap (0, 1) \neq \emptyset$. If $\|x_i\| > 2$, $\|x_j\| \leq 2$, and $\|x_i - x_j\| \geq r_i$, we have $\|y_i - y_j\| = \|y_i - x_j\| \geq 1$. $\|x_i\| > 2$, $\|x_j\| \leq 2$, and $\|x_i - x_j\| < r_i$, we have $\|x_i - x_j\| \geq r_j \geq r_i/\tau$, so

$$2\tau \geq \|x_j\|\tau \geq r_j$$

so $2\tau^2 \geq r_i$. Thus

$$r_i - r_i/\tau = r_i\left(1 - \frac{1}{\tau}\right) \leq 2\tau^2\left(1 - \frac{1}{\tau}\right) < \epsilon$$

and

$$\|y_i - y_j\| = \|y_i - x_j\| \geq \|x_i - x_j\| - \|x_i - y_i\| \geq r_i/\tau - (r_i - 1) = 1 - (r_i - r_i/\tau) > 1 - \epsilon \ .$$

Now consider the case $\|x_i\| \geq \|x_j\| > 2$. Set $s = \|x_j\|$ and $z = \frac{s}{\|x_i\|} x_i$ (z is the point on the ray from 0 to x_i of norm s). We have

$$\|x_i - x_j\| \leq \|x_i - z\| + \|z - x_j\| = (\|x_i\| - s) + \|z - x_j\| \ .$$

If $a \in (x_i, r_i) \cap (0, 1)$, we have

$$\|x_i\| \leq \|x_i - a\| + \|a\| \leq r_i + 1$$

so

$$\|z - x_j\| \geq (s + \|x_i - x_j\|) - \|x_i\| \geq (s + \|x_i - x_j\|) - (r_i + 1) = (s - 1) + (\|x_i - x_j\| - r_i) \ .$$

We have $y_i = \frac{2}{s} z$ and $y_j = \frac{2}{s} x_j$, so if $\|x_i - x_j\| \geq r_i$ we have

$$\|y_i - y_j\| = \frac{2}{s}\|z - x_j\| \geq \frac{2}{s}(s - 1) = 2 - \frac{2}{s} > 1$$

since $s > 2$. If $\|x_i - x_j\| < r_i$, then $\|x_i - x_j\| \geq r_j \geq r_i/\tau$ so

$$r_i - \|x_i - x_j\| \leq r_i - r_j \leq \tau r_j - r_j = r_j(\tau - 1) \leq \tau s(\tau - 1) \ .$$

We have $2\tau(\tau-1) \leq 2\tau^2(\tau-1) < \epsilon$, so

$$\|y_i - y_j\| \geq \frac{2}{s}[(s-1) - (\|x_i - x_j\| - r_i)] \geq \left(2 - \frac{2}{s}\right) - 2\tau(\tau-1) > 1 - \epsilon \ .$$

Thus, since $y_0 = 0$, $m \leq \theta(X)$. 👍

XVII.8.1.13. It is generally difficult to find $M(X)$ for a given normed vector space; the bound $M(X) \leq 5^{dim(X)}$ is usually crude. $M(\mathbb{R}^2)$ with the usual metric is 19 (Exercise XVII.8.4.3.). Exact values for $M(\mathbb{R}^n)$ for $n > 2$ do not appear to be known. However, for our purposes the value does not matter; we only need to know it is finite.

XVII.8.1.14. There are many other metric spaces of interest which are controlled besides subsets of Euclidean space (or finite-dimensional normed vector spaces); cf. Exercise XVII.8.4.2..

Finite-Dimensionality

XVII.8.1.15. PROPOSITION. (i) If (X, ρ) is a controlled metric space, then $dim(X) < \infty$ ().
(ii) If X is a separable metrizable (second countable regular) topological space and $dim(X) < \infty$, then there is a controlled metric on X giving its topology.

PROOF: (i) Let M be a constant for (X, ρ) as in the definition (for some R and $\tau > 1$). If $\mathcal{U}$ is an open cover of X, then $\mathcal{U}$ has a refinement $\mathcal{C}$ which is a Besicovitch cover. By (), $\mathcal{C}$ has a subcover $\mathcal{D}$ which can be split into at most M pieces $\mathcal{D}_i$, each of which consists of pairwise disjoint sets. Thus the nuclear dimension () of X is $\leq M - 1$.
(ii) X can be embedded in $\mathbb{R}^n$ for some n (XIII.22.6.3.). Apply XVII.8.1.3.(i). 👍

XVII.8.1.16. The upper bound of $M-1$ in (i) for $dim(X)$ is generally crude: $dim(X)$ is typically more on the order of $\log M$ or even smaller (cf. XVII.8.4.7.).

XVII.8.1.17. Not every metric on a finite-dimensional metric space, even a compact one, is controlled (XVII.8.4.8.).

XVII.8.2. The Besicovitch Partition Theorem

We first obtain a version of the Besicovitch Covering Theorem which has no reference to Borel measures. Actually the definition of a controlled metric was made specifically so this result holds.

There is inconsistency in the literature about names: some authors call this the Besicovitch Covering Theorem. ✋

Besicovich Coverings

XVII.8.2.1. DEFINITION. Let (X, ρ) be a metric space, and $A \subseteq X$. A *Besicovitch covering* (sometimes called a *fine covering*) of A in X is a collection $\mathcal{C} = \{(x_i, r_i) : i \in I\} \subseteq X \times (0, \infty)$ of closed balls in X such that for every $x \in A$ and $\epsilon > 0$, there is a δ with $0 < \delta \leq \epsilon$ with $(x, \delta) \in \mathcal{C}$.

In other words, for every $x \in A$, there are balls *centered at* x of arbitrarily small radius in the cover $\mathcal{C}$. This is stronger than the notion of a Vitali cover ().

Only small balls are needed, and we can restrict to small balls:

XVII.8.2.2. PROPOSITION. Let $\mathcal{C}$ be a Besicovitch covering of a subset A in a metric space (X, ρ).

(i) For any $\epsilon > 0$, the collection $\mathcal{C}_\epsilon$ of balls $(x, r) \in \mathcal{C}$ with $r < \epsilon$ is also a Besicovitch covering of A in X.

(ii) If W is an open subset of X containing A, then the set $\mathcal{C}_W$ of balls $(x, r) \in \mathcal{C}$ with $\bar{B}_r(x) \subseteq W$ is also a Besicovitch covering of A in X.

(iii) The subset $\mathcal{C}'$ of $\mathcal{C}$ consisting of balls in $\mathcal{C}$ with centers in A is also a Besicovitch covering of A.

XVII.8.2.3. Thus we may assume a Besicovitch covering of A in X has any or all of these properties. But it is convenient not to require them in the definition of Besicovitch covering.

The Partition Theorem

XVII.8.2.4. THEOREM. [BESICOVITCH PARTITION THEOREM] Let (X, ρ) be a metric space. If ρ is controlled with constants $M \in \mathbb{N}$, $R > 0$, and $\tau > 1$, $A \subseteq X$, and $\mathcal{C}$ is a Besicovitch covering of A, then there are M pairwise disjoint subcollections $\mathcal{D}_1, \ldots, \mathcal{D}_M$ of $\mathcal{C}$ such that $\cup_i \mathcal{D}_i$ covers A and each $\mathcal{D}_i$ consists of pairwise disjoint balls.

While a proof of this theorem using Zorn's Lemma can be given (cf. [Fed69, 2.8]), it is more illuminating to give a (semi)constructive proof by transfinite induction, along the lines of the argument outlined in [Mor09].

PROOF: We choose a set of balls by transfinite induction. We want at each step to choose the next ball to be the "largest" remaining ball; but since there may not be a largest ball, we will choose one which is "almost the largest." First well order the balls (x_α, r_α) in $\mathcal{C}$. By transfinite induction, we will select a ball, place it in a row, discard some other balls, and reindex the remaining balls.

Let $s = \sup\{r_\alpha : (x_\alpha, r_\alpha) \in \mathcal{C}\}$. At the first step, select the first (x_α, r_α) with $r_\alpha > s/\tau$, put it in the first row, and discard all balls (x_β, r_β) with $x_\beta \in \bar{B}_{r_\alpha}(x_\alpha)$. Reindex the remaining balls, calling the chosen ball B_0. Note that the remaining balls (B_0 along with the ones not discarded) still cover A.

Suppose $B_0, \ldots, B_\alpha$ have been chosen and put in rows, and some other ones discarded, so that

The balls already chosen form an (R, τ)-controlled set.

The chosen balls in each row are pairwise disjoint.

No chosen ball is disjoint from all the previously chosen balls in any higher row.

All remaining balls (x, r) not chosen have $r/\tau \leq r_\beta$ for all $\beta \leq \alpha$, and x not in any chosen ball.

The remaining balls (chosen or not) cover A.

Let s_α be the supremum of the radii of the remaining balls not chosen. Choose $B_{\alpha+1}$ to be the first remaining ball not yet chosen with radius greater than s_α/τ. By construction, the center $x_{\alpha+1}$ of $B_{\alpha+1}$ is not in any previously chosen ball; and $B_{\alpha+1}$ along with the previously chosen balls still form a τ-controlled set. Put $B_{\alpha+1}$ in the first row for which it is disjoint from the balls already in that row, or put it in a new row if there is no existing row for which this holds. Discard any succeeding balls whose center is in $\bar{B}_{r_\alpha}(x_\alpha)$. The remaining balls (chosen or not) still cover A. Reindex the remaining balls.

If α is a limit ordinal and B_β have been chosen for all $\beta < \alpha$. Set $t = \inf_{\beta<\alpha} r_\beta$. There are two cases:

(1) $t > 0$. Then the previously chosen B_β, along with the remaining balls, cover A: if $x \in A \setminus \cup_{\beta<\alpha} B_\beta$ and $(x, r) \in \mathcal{C}$ with $r < t$, then (x, r) has not been discarded. In this case we can choose B_α and discard and reindex the same way as above.

(2) $t = 0$. If $(x, r) \in \mathcal{C}$, choose $\beta < \alpha$ with $r_\beta/\tau < r$; then (x, r) must have been either chosen or discarded by the β'th step. Thus x is in one of the previously chosen balls, and the already chosen balls cover A. In fact, in this case there are no remaining balls not chosen, and the process terminates.

Continue transfinitely until there are no remaining balls not either chosen or discarded. (This will happen when (2) occurs.)

At the end of the process, we have balls from $\mathcal{C}$ in rows with the following properties:

All the balls cover A.

If (x, r) is a chosen ball and (y, s) a previously chosen ball, then $r/\tau \leq s$.

All the balls in each row are pairwise disjoint.

No ball in any row is disjoint from all the previously chosen balls in any previous row.

There could *a priori* be infinitely many rows, even uncountably many. But the key fact is:

XVII.8.2.5. LEMMA. There are at most M rows.

PROOF: Suppose there is a chosen ball B, of radius r, which is put in the $M+1$ row. Then in each preceding row k, there is a previously chosen ball B_k not disjoint from B; the radius of B_k is at least r/τ. Then, since all balls have radius no more than R, the balls $B_1, \ldots, B_M$ form an (R, τ)-controlled expansion of B of cardinality M, contradicting the choice of M, R, and τ. 👍

We let $\mathcal{D}_i$ be the set of balls in the i'th row (some $\mathcal{D}_i$ may be empty if there are less than M rows). This completes the proof of Theorem XVII.8.3.1.. 👍

XVII.8.3. The Besicovitch Covering Theorem

Informally, the Besicovitch Covering Theorem says that if μ is a locally finite Borel measure on a controlled metric space (X, ρ), $A \subseteq X$ of finite measure, and $\mathcal{C}$ a Besicovitch covering of A in X, then there is a sequence of pairwise disjoint closed balls in $\mathcal{C}$ which cover μ-almost all of A. We use the outer measure μ^* defined by μ (XVII.6.1.6.): if $Y \subseteq X$, set

$$\mu^*(Y) = \inf\{\mu(B) : B \text{ Borel}, Y \subseteq B\} = \min\{\mu(B) : B \text{ Borel}, Y \subseteq B\} .$$

We need this outer measure since A is not necessarily a Borel set in X.

XVII.8.3.1. Theorem. [Besicovitch Covering Theorem] Let (X, ρ) be a controlled metric space and μ a locally finite Borel measure () on X, $A \subseteq X$ with $\mu^*(A) < \infty$, and $\mathcal{C}$ a Besicovitch covering of A in X. Assume that

(i) There is a largest open set W in X with $\mu(W) = 0$.

((i) is automatic if (X, ρ) is separable). Then there is a countable subset $\mathcal{B} = \{(x_n, r_n)\}$ of $\mathcal{C}$ with the $\bar{B}_{r_n}(x_n)$ pairwise disjoint, such that

$$\mu^*\left(A \setminus \bigcup_{\mathcal{B}} \bar{B}_{r_n}(x_n)\right) = 0 .$$

($\mathcal{B}$ will usually be countably infinite, i.e. a sequence.) For any $\epsilon > 0$, we may choose $\mathcal{B}$ so that

$$\sum_{B \in \mathcal{B}} \mu(B) < \mu^*(A) + \epsilon .$$

Proof: We may assume $\mu^*(A) > 0$ since otherwise there is nothing to prove. Fix an $M \in \mathbb{N}$, $R > 0$, and $\tau > 1$ for ρ as in the definition of a controlled metric (XVII.8.1.1.). We may assume all balls $(x, r) \in \mathcal{C}$ have $x \in A$ and $r < R$ (XVII.8.2.2.).

To see that (i) is automatic if (X, ρ) is separable (second countable), let $\mathcal{U}$ be a countable base for the topology, and let W be the union of all $U \in \mathcal{U}$ with $\mu(U) = 0$.

Suppose (i) holds. Then any closed ball whose center is not in W has positive measure. We have

$$\mu^*(A) = \mu^*(A \cap W) + \mu^*(A \cap W^c) = \mu^*(A \cap W^c)$$

since W is measurable, and $\mathcal{C}$ is a Besicovitch covering of $A \cap W^c$. So it suffices to obtain the conclusion for $A \cap W^c$, i.e. replacing A by $A \cap W^c$ and $\mathcal{C}$ by the subset consisting of all $(x, r) \in \mathcal{C}$ with $x \in A \cap W^c$, we may assume that all balls in $\mathcal{C}$ have strictly positive measure.

Let $\epsilon > 0$. Since μ is a locally finite Borel measure on X and $\mu(A) < \infty$, there is an open set V in X with $A \subseteq V$ and $\mu(V) < \mu^*(A) + \epsilon$ (). We may assume all $(x, r) \in \mathcal{C}$ have $\bar{B}_r(x) \subseteq V$ (XVII.8.2.2.).

By Theorem XVII.8.2.4. there are subsets $\mathcal{D}_1, \ldots, \mathcal{D}_M$ of $\mathcal{C}$ so that $\cup_i \mathcal{D}_i$ covers A and each $\mathcal{D}_i$ consists of pairwise disjoint balls. Each $\mathcal{D}_i$ is countable since all the balls are contained in V and have positive measure.

For each i, $1 \le i \le M$, let D_i be the union of the balls in $\mathcal{D}_i$. We have $A \subseteq \cup_i D_i$, so

$$\mu^*(A) \le \sum_{i=1}^{M} \mu^*(A \cap D_i)$$

and thus there is an i for which

$$\mu^*(A \cap D_i) \ge \frac{1}{M} \mu^*(A) \ .$$

Fix such an i, and, changing notation, let $B_1, B_2, \ldots$ be an enumeration of the balls in row i. Let Z be a Borel set containing $A \cap D_i$ with $\mu(Z) = \mu^*(A \cap D_i)$. By regularity, there is a closed set E with $E \subseteq Z$ and with

$$\mu(E) > \frac{1}{M+1} \mu^*(A) \ .$$

Since the B_k are disjoint and cover almost all of E, we have

$$\mu(E) = \sum_{k=1}^{\infty} \mu(E \cap B_k)$$

so there is an n such that

$$\sum_{k=1}^{n} \mu(E \cap B_k) > \frac{1}{M+1} \mu^*(A) \ .$$

Put $B_1, \ldots, B_n$ into $\mathcal{B}$.

Set $F = B_1 \cup \cdots \cup B_n$. Then F is a closed set, and

$$\mu^*(A \cap F) \ge \mu(E \cap F) > \frac{1}{M+1} \mu^*(A) \ .$$

Set $A_1 = A \setminus F$. Since F is measurable,

$$\mu^*(A) = \mu^*(A \cap F) + \mu(A_1)$$

so

$$\mu^*(A_1) < \frac{M}{M+1} \mu^*(A) \ .$$

Set $V_1 = V \cap F^c$. Then V_1 is an open set of finite measure disjoint from $B_1, \ldots, B_n$, and $A_1 \subseteq V_1$. Let $\mathcal{C}_1$ be the set of balls in $\mathcal{C}$ contained in V_1; then $\mathcal{C}_1$ is a Besicovitch covering of A_1. Repeat the construction with $\mathcal{C}_1$, A_1, V_1 in place of $\mathcal{C}$, A, V, to obtain a closed set F_1 which is a finite union of disjoint balls in $\mathcal{C}_1$, with

$$\mu^*(A_1 \cap F_1) > \frac{1}{M+1} \mu^*(A_1) \ .$$

Put these balls into $\mathcal{B}$, and set $A_2 = A_1 \setminus F_1 = A \setminus (F \cup F_1)$; then

$$\mu^*(A_2) < \frac{1}{M+1} \left(\frac{M}{M+1} \right) \mu^*(A) \ .$$

Repeat the process inductively; after n steps we have that the part A_n of A not covered by $\mathcal{B}$ has

$$\mu^*(A_n) < \frac{1}{M+1}\left(\frac{M}{M+1}\right)^{n-1}\mu^*(A)\ .$$

Since

$$\frac{M}{M+1} < 1$$

at the end of the process (it might end after finitely many steps if $\mu^*(A_n) = 0$ for some n) we have a countable set $\mathcal{B}$ of pairwise disjoint balls in $\mathcal{C}$ which covers almost all of A. Since all the balls in $\mathcal{B}$ are contained in V, the last assertion follows. 👍

Euclidean Space

The hypotheses of the theorem are satisfied if X is a subset of $\mathbb{R}^n$ (or a finite-dimensional normed vector space) with the metric induced by the norm (XVII.8.1.6.). Thus we obtain the common version of the theorem:

XVII.8.3.2. Corollary. [Besicovitch Covering Theorem, Euclidean Version] Let $X \subseteq \mathbb{R}^n$ with the induced metric and μ a locally finite Borel measure on X, $A \subseteq X$ with $\mu^*(A) < \infty$, and $\mathcal{C}$ a Besicovitch covering of A in X. Then there is a countable subset $\mathcal{B} = \{(x_n, r_n)\}$ of $\mathcal{C}$ with the $\bar{B}_{r_n}(x_n)$ pairwise disjoint, such that

$$\mu^*\left(A \setminus \bigcup_{\mathcal{B}} \bar{B}_{r_n}(x_n)\right) = 0\ .$$

($\mathcal{B}$ will usually be countably infinite, i.e. a sequence.) For any $\epsilon > 0$, we may choose $\mathcal{B}$ so that

$$\sum_{B\in\mathcal{B}} \mu(B) < \mu^*(A) + \epsilon\ .$$

XVII.8.3.3. Actually the X can be eliminated since any (locally finite) Borel measure on a subset of a topological space can be converted to a (locally finite) Borel measure on the whole space ().

XVII.8.3.4. Example. Here is an example illustrating why assumption (i) (or something similar) in the statement of XVII.8.3.1. is necessary. Let X be a set of measurable cardinality (), ρ the discrete metric, μ a nonatomic $\{0,1\}$-valued measure on $(X, \mathcal{P}(X))$, and $\mathcal{C}$ the collection of all balls of radius $< \frac{1}{2}$ (each such ball is actually a singleton). The metric ρ is controlled (with, say, constants $R = 1/2$, $M = 1$, $\tau = 3/2$). But (i) is not satisfied. If $A = X$, the process in the proof can be carried out, and there is only one $\mathcal{D}_i$, but it contains uncountably many balls. Actually, the conclusion of the theorem holds except that $\mathcal{B}$ is necessarily uncountable (in fact, it must be all the singleton subsets).

XVII.8.3.5. The proof uses the AC. But if $card(\mathcal{C}) \leq 2^{\aleph_0}$ (e.g. if (X, ρ) is separable), then one only needs that $\mathbb{R}$ can be well ordered (i.e. that $2^{\aleph_0}$ is an aleph), which is weaker than the full AC (e.g. the CH suffices).

XVII.8.4. Exercises

XVII.8.4.1. (a) Verify the assertions in Example XVII.8.1.4..

(b) Show that the following version of the Besicovitch Partition Theorem holds:

THEOREM. Let (X, ρ) be an ultrametric space, $A \subseteq X$, and $\mathcal{C}$ a Besicovitch covering of A. Then there is a subset $\mathcal{D}$ of $\mathcal{C}$ consisting of pairwise disjoint balls which covers A.

[In the proof of XVII.8.2.4., fix $\tau > 1$ and at each stage throw away not only any remaining balls (not yet chosen) with center in $B_{\alpha+1}$, but also any remaining balls of larger radius. All balls go in the first row.]

(c) Show that the statement of the Besicovitch Covering Theorem holds if (X, ρ) is an ultrametric space, with a simplified proof, even with "almost all" replaced by "all."

XVII.8.4.2. [Fed69, 2.8] Let (X, ρ) be a metric space. If $\zeta \in \mathbb{N}$, $0 < \eta \leq \frac{1}{3}$, and $0 < R \leq \infty$, and $A \subseteq X$, say (X, ρ) is R, η, ζ-*directionally limited at* A if, whenever $a \in A$, $P \subseteq B_R(a) \setminus \{a\}$, and $b, c \in P$ with $b \neq c$ and $\rho(a, b) \geq \rho(a, c)$, and

$$\rho(a, c) + \rho(c, b) - \rho(a, b) \geq \eta\rho(a, b)$$

then $card(P) \leq \zeta$.

(a) Show that if X is a finite-dimensional normed vector space, then X is $\infty, 1/3, \zeta$-directionally limited at X for some $\zeta \in \mathbb{N}$.

(b) Show that if X is a Riemannian manifold and K is a compact subset, then X is R, η, ζ-directionally limited at K for some R, η, ζ.

(c) Show that if (X, ρ) is R, η, ζ-directionally limited at A, and $Y \subseteq X$, then $(Y, \rho|_Y)$ is R, η, ζ-directionally limited at $A \cap Y$.

(d) Show that if (X, ρ) is R, η, ζ-directionally limited at A, then its completion $(\bar{X}, \bar{\rho})$ is R, η, ζ-directionally limited at $\bar{A}$.

(e) If (X, ρ) is R, η, ζ-directionally limited on every ball of radius R in X, and if $\tau > 1$ satisfies

$$\eta + \tau/(2 - \eta) + \tau(\tau - 1) < 1$$

then (X, ρ) is (R, τ)-controlled, with constant $M = 2\zeta + 1$.

(f) In (e), show that any $\tau > 1$ which is sufficiently close to 1 satisfies the condition. If $\eta = 1/3$, the restriction is

$$1 < \tau < \frac{1}{5} + \sqrt{\frac{53}{75}} \approx 1.04 \ .$$

Parts (a) and (e) give an alternate proof of Theorem XVII.8.1.6.. The *directionally limited* definition here is a less restrictive version of the one in [Fed69]: the definition there presupposes the existence of interpolating points; in the absence of such points the restriction is vacuous (always satisfied) and the condition essentially forces the metric to be discrete (in the presence of the interpolating points the two definitions are the same by the argument in the proof of [Fed69, 2.8.11]).

XVII.8.4.3. [Sul94] For each n let $M(n) = M(n,1)$ be the best constant for strict control of $\mathbb{R}^n$ ().
(a) [Rei48] Show that $M(2) = 19$. See Figure XVII.2 (from [Sul94]).
(b) Show that

$$67 \leq M(3) \leq 87$$

$$226 \leq M(4) \leq 331$$

$$681 \leq M(5) \leq 1159$$

and that $M(n)$ grows exponentially in n.

It is conjectured in [Sul94] that $M(3) = 67$ and that $M(4)$ and $M(5)$ are close to the lower bounds 226 and 681 respectively. The value of $M(n)$ is closely related to sphere-packing problems.

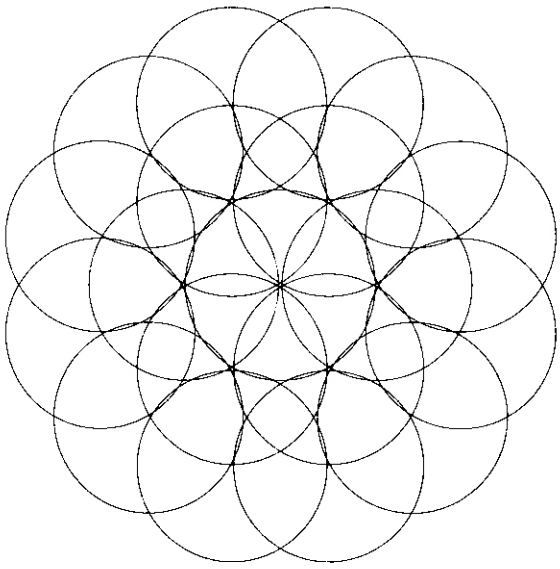


Figure XVII.2: A (strictly) controlled expansion consisting of 19 balls (disks) in $\mathbb{R}^2$

XVII.8.4.4. (a) Show that a metrically finite-dimensional metric space () is strictly controlled.
(b) Try to prove that a metrically finite-dimensional metric space is controlled. What goes wrong? (See XVII.8.1.5..)

In fact, metrically finite-dimensional metric spaces are not controlled in general [LDR19].

XVII.8.4.5. Show that a metric space satisfying the conclusion of the Besicovitch Partition Theorem is topologically finite-dimensional. [See the proof of XVII.8.1.15.(i).]

XVII.8.4.6. (a) Show that the proof of Theorem XVII.8.2.4. works verbatim if the cover $\mathcal{C}$ of A is not necessarily a Besicovitch cover, but merely a cover containing a ball centered at each point of A, of uniformly bounded radius.

(b) In verifying the hypotheses of this version, it suffices to take $A = X$. [If $\mathcal{C}$ is such a cover, for each $x \in \bar{A} \setminus A$ add a ball centered at x contained in a ball in $\mathcal{C}$ of almost the largest radius possible, and for points of $X \setminus \bar{A}$ an almost largest ball centered at x disjoint from $\bar{A}$.]

(c) Show that a metric space satisfying the conclusion of this more general version of the Besicovitch Partition Theorem is controlled. [Let A be the set of centers of a controlled expansion.]

XVII.8.4.7. Let $1 \leq N \leq \infty$, and let X be a countable dense subset of $[0,1]^N$, with the restriction of a metric giving the topology of $[0,1]^N$. Then X is homeomorphic to $\mathbb{Q}$ (), and in particular zero-dimensional, but its completion is N-dimensional. The metric is not controlled if $N = \infty$ (XVII.8.1.3., XVII.8.1.15.).

XVII.8.4.8. Let $X = \prod_{n=1}^{\infty} \{1, \ldots, n\}$. Then X consists of all sequences in $\mathbb{N}$ for which the n'th term is $\leq n$ for all n. If x and y are such sequences, set $\rho(x,y) = \frac{1}{n}$ if x and y first differ in the n'th term. Then ρ is an ultrametric on X. The ρ-topology on X is the product topology, where each coordinate has the discrete topology; thus (X, ρ) is compact and zero-dimensional (and homeomorphic to the Cantor set, cf. ()). Since $\frac{n}{n+1}$ becomes arbitrarily close to 1, (X, ρ) is not (R, τ)-controlled for any $R > 0$ and $\tau > 1$.

XVII.9. Signed and Complex Measures

In this section we examine the theory of measures which are not necessarily nonnegative or even real-valued. It turns out that the theory of such measures can be nicely reduced to the theory of ordinary (nonnegative) measures.

The only insurmountable difficulty comes with trying to make sense of the expression $\infty - \infty$. It will be implicit in our definitions that this possibility will never arise, which will entail that our measures do not take both values $\pm\infty$. In most applications, we will only consider real-valued measures, i.e. ones not taking either value $\pm\infty$. For complex measures we will not be able to satisfactorily allow any type of infinite value. There is an alternate approach to avoiding the $\infty - \infty$ problem, using local measures only defined on δ-rings ().

XVII.9.1. Signed Measures

XVII.9.1.1. DEFINITION. Let $(X, \mathcal{A})$ be a measurable space. A *signed measure* on $(X, \mathcal{A})$ is a function $\nu : \mathcal{A} \to [-\infty, +\infty]$ such that

(i) $\nu(\emptyset) = 0$.

(ii) If $\{A_n\}$ is a sequence of pairwise disjoint sets in $\mathcal{A}$, then either

$$\sum_{\nu(A_n)>0} \nu(A_n) < +\infty \quad \text{or} \quad \sum_{\nu(A_n)<0} \nu(A_n) > -\infty$$

(so $\sum_n \nu(A_n)$ is well defined), and

$$\nu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \nu(A_n) \ .$$

The signed measure ν is *finite* if $\nu(A) \in \mathbb{R}$ for all $A \in \mathcal{A}$.

Note that, as with measures, the countable additivity of (ii) includes finite additivity (taking the remaining sets to be $\emptyset$). In particular, there cannot be disjoint sets $A_1, A_2 \in \mathcal{A}$ with $\nu(A_1) = +\infty$ and $\nu(A_2) = -\infty$.

XVII.9.1.2. EXAMPLES. (i) Any measure is a signed measure.

(ii) If μ_1 and μ_2 are (ordinary) measures on a measurable space $(X, \mathcal{A})$, at least one of which is finite, then $\mu_1 - \mu_2$ is a well-defined signed measure on $(X, \mathcal{A})$. More generally, any (finite) linear combination of measures, all but at most one of which are finite, is a signed measure. Any (finite) linear combination of finite signed measures is a (finite) signed measure.

(iii) Let $(X, \mathcal{A}, \mu)$ be an (ordinary) measure space, and $f \in \mathcal{L}^1(X, \mathcal{A}, \mu)$. For $A \in \mathcal{A}$, let

$$\nu(A) = \int_A f \, d\mu \ .$$

It is easy to check (cf. ()) that ν is a finite signed measure.

A signed measure of the form in (ii) can be regarded as "trivial", since its structure is completely determined and easy to describe in terms of the properties of μ_1 and μ_2. The main result of this section is that every signed measure is of this form, with the positive and negative pieces supported on disjoint sets:

XVII.9.1.3. THEOREM. [JORDAN DECOMPOSITION THEOREM] Let ν be a signed measure on a measure space $(X, \mathcal{A})$. Then there are unique measures ν^+ and ν^- on $(X, \mathcal{A})$, with $\nu = \nu^+ - \nu^-$ and $\nu^+ \perp \nu^-$. At least one of ν^+ and ν^- is a finite measure.

Thus the theory of signed measures turns out to be "trivial." (However, the *proof* that the theory is actually trivial is definitely nontrivial!)

It is useful to rephrase this theorem in a somewhat different way. We first make some definitions:

XVII.9.1.4. DEFINITION. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$. A set $P \in \mathcal{A}$ is a *positive set* for ν if $\nu(A) \geq 0$ for every $A \in \mathcal{A}$, $A \subseteq P$. Similarly, $N \in \mathcal{A}$ is a *negative set* for ν if $\nu(A) \leq 0$ for all $A \in \mathcal{A}$, $A \subseteq N$. A set $E \in \mathcal{A}$ is a *null set* for ν if it is both positive and negative, i.e. $\nu(A) = 0$ for all $A \in \mathcal{A}$, $A \subseteq E$.

Note that a set $E \in \mathcal{A}$ with $\nu(E) = 0$ is not necessary a null set: such an E could be a disjoint union $A \cup B$, where $\nu(A) > 0$ and $\nu(B) = -\nu(A)$. Similarly, an $E \in \mathcal{A}$ with $\nu(E) > 0$ is not necessarily a positive set.

It is obvious that any (measurable) subset of a positive set is positive, and a finite or countable union of positive sets is positive (Exercise ()). The same holds for negative sets and null sets.

XVII.9.1.5. DEFINITION. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$. A *Hahn decomposition* for ν is a pair (P, N), where $P, N \in \mathcal{A}$, $N = X \setminus P$, P is positive for ν, and N is negative for ν.

We can then rephrase the Jordan Decomposition Theorem in the following way:

XVII.9.1.6. THEOREM. [HAHN DECOMPOSITION THEOREM] Let ν be a signed measure on a measurable space $(X, \mathcal{A})$. Then there is a Hahn decomposition for ν. If (P, N) and (P', N') are Hahn decompositions for ν, then $P \triangle P'$ and $N \triangle N'$ are null sets for ν, i.e. the Hahn decomposition of ν is unique up to null sets.

XVII.9.1.7. To see that the Hahn Decomposition Theorem is equivalent to the Jordan Decomposition Theorem (we will only need one direction of this equivalence), suppose (P, N) is a Hahn decomposition for ν. For $A \in \mathcal{A}$, set $\nu^+(A) = \nu(A \cap P)$ and $\nu^-(A) = -\nu(A \cap N)$. Then it is easily checked that ν^+ and ν^- are measures, $\nu^+ \perp \nu^-$, and $\nu = \nu^+ - \nu^-$. These measures do not change if P and N are changed by null sets. At least one of P and N must have finite measure. Conversely, let $\nu = \mu_1 - \mu_2$, where μ_1 and μ_2 are measures with $\mu_1 \perp \mu_2$. Let $P \in \mathcal{A}$ with $\mu_2(P) = 0$ and $\mu_1(X \setminus P) = 0$; such a P is unique up to null sets. Then $(P, X \setminus P)$ is a Hahn decomposition for ν. Another Hahn decomposition for ν not differing from this one by null sets would give a pair (ν^+, ν^-) different from (μ_1, μ_2).

XVII.9.1.8. EXAMPLE. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f \in \mathcal{L}^1(X, \mathcal{A}, \mu)$, and for $A \in \mathcal{A}$ set $\nu(A) = \int_A f \, d\mu$ as in XVII.9.1.2.. A Hahn decomposition for ν is precisely a pair (P, N), where $N = X \setminus P$, $f \geq 0$ a.e. on P, and $f \leq 0$ a.e. on N. One Hahn decomposition is given by $P = \{x : f(x) \geq 0\}$ and $N = \{x : f(x) < 0\}$. The Jordan decomposition of ν is $\nu = \nu^+ - \nu^-$, where $\nu^+(A) = \int_A f_+ \, d\mu$ and $\nu^-(A) = \int_A f_- \, d\mu$ for $A \in \mathcal{A}$ and f_+, f_- the positive and negative parts of f ().

The proof of Theorem XVII.9.1.6. (and therefore Theorem XVII.9.1.3.) will consist of a series of lemmas, all of which would be easy corollaries of the theorem. Our proof is based on [Dos80] [note that the argument of [Dos80] is incomplete as written since it assumes the conclusion of XVII.9.1.14. without comment.]

The first result about signed measures is that a signed measure cannot take both the values $+\infty$ and $-\infty$:

XVII.9.1.9. LEMMA. Let ν be a signed measure on $(X, \mathcal{A})$. Then there do not exist $A, B \in \mathcal{A}$ with $\nu(A) = +\infty$ and $\nu(B) = -\infty$.

PROOF: It follows immediately from XVII.9.1.1.(ii) that there do not exist *disjoint* sets $A_1, A_2 \in \mathcal{A}$ with $\nu(A_1) = +\infty$ and $\nu(A_2) = -\infty$. Suppose $\nu(A) = +\infty$ and $\nu(B) = -\infty$. Then $A \cap B$, $A \setminus B$, and $B \setminus A$ are pairwise disjoint, and:

At least one of $A \cap B$ and $A \setminus B$ has measure $+\infty$.

At least one of $A \cap B$ and $B \setminus A$ has measure $-\infty$.

Thus there are disjoint sets with measure $+\infty$ and $-\infty$, a contradiction. 👍

XVII.9.1.10. LEMMA. Let ν be a signed measure on $(X, \mathcal{A})$, and $A, B \in \mathcal{A}$ with $B \subseteq A$. If $\nu(A) \in \mathbb{R}$ (i.e. $|\nu(A)| < \infty$), then $\nu(B) \in \mathbb{R}$.

PROOF: We have $\nu(A) = \nu(B) + \nu(A \setminus B)$, and the sum on the right is well defined. If $\nu(B) = +\infty$, then we cannot have $\nu(A \setminus B) = -\infty$, and hence we would have to have $\nu(A) = +\infty$. The argument if $\nu(B) = -\infty$ is similar. 👍

The next result is an exact analog of ().

XVII.9.1.11. LEMMA. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$.

(i) If $E_1, E_2, \cdots \in \mathcal{A}$ with $E_1 \subseteq E_2 \subseteq \cdots$, then $\lim_{n\to\infty} \nu(E_n)$ exists as an extended real number and equals $\nu(\cup_{n=1}^{\infty} E_n)$.

(ii) If $E_1, E_2, \cdots \in \mathcal{A}$ with $E_1 \supseteq E_2 \supseteq \cdots$, and $\nu(E_n) \in \mathbb{R}$ for some n, then $\lim_{n\to\infty} \nu(E_n)$ exists as a real number and equals $\nu(\cap_{n=1}^{\infty} E_n)$.

PROOF: (i): Let $A_1 = E_1$ and $A_n = E_n \setminus E_{n-1}$ for $n > 1$. Then the A_n are pairwise disjoint, $E_n = \cup_{k=1}^{n} A_k$ for all n, $\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} E_n$, and the series $\sum_{k=1}^{\infty} \nu(A_k)$ converges to $\nu(\cup_{n=1}^{\infty} E_n)$. So

$$\nu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{k=1}^{\infty} \nu(A_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \nu(A_k) = \lim_{n\to\infty} \nu\left(\bigcup_{k=1}^{n} A_k\right) = \lim_{n\to\infty} \nu(E_n) \ .$$

(ii): We may assume $\nu(E_1) \in \mathbb{R}$. Set $E = \cap_{n=1}^{\infty} E_n$, $B_n = E_1 \setminus E_n$, and $B = \cup_{n=1}^{\infty} B_n = E_1 \setminus E$. Then by (i)

$$\nu(B) = \lim_{n\to\infty} \nu(B_n) = \lim_{n\to\infty} \nu(E_1 \setminus E_n) \ .$$

Thus, since $\nu(E_1) \in \mathbb{R}$,

$$\lim_{n\to\infty} \nu(E_1 \setminus E_n) = \lim_{n\to\infty} [\nu(E_1) - \nu(E_n)]$$

exists and hence $\lim_{n\to\infty} \nu(E_n)$ exists and equals $\nu(E_1) - \lim_{n\to\infty} \nu(B_n)$, and we have

$$\nu(E) = \nu(E_1) - \nu(B) = \nu(E_1) - \lim_{n\to\infty} \nu(B_n) = \lim_{n\to\infty} \nu(E_n) \ .$$

Furthermore, since $E \subseteq E_1$ and $\nu(E_1) \in \mathbb{R}$, we have $\nu(E) \in \mathbb{R}$ by XVII.9.1.10.. 👍

XVII.9.1.12. LEMMA. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$ not taking the value $+\infty$, and let $\epsilon > 0$. Then for every $A \in \mathcal{A}$ there is a set $A_\epsilon \in \mathcal{A}$ with $A_\epsilon \subseteq A$, $\nu(A_\epsilon) \geq \nu(A)$, and $\nu(B) > -\epsilon$ for all $B \in \mathcal{A}$, $B \subseteq A_\epsilon$.

PROOF: Fix $\epsilon > 0$ and $A \in \mathcal{A}$. If $\nu(A) \leq 0$, we may take $A_\epsilon = \emptyset$. Assume $\nu(A) > 0$ (in particular, $\nu(A) \in \mathbb{R}$). Suppose the conclusion is false. Then there is a $B_1 \in \mathcal{A}$, $B_1 \subseteq A$, with $\nu(B_1) \leq -\epsilon$ (otherwise we could take $A_\epsilon = A$). Then $\nu(A \setminus B_1) = \nu(A) - \nu(B_1) \geq \nu(A) + \epsilon$. There is a $B_2 \in \mathcal{A}$, $B_2 \subseteq (A \setminus B_1)$, with $\nu(B_2) \leq -\epsilon$ (otherwise we could take $A_\epsilon = A \setminus B_1$). Proceeding in this way, we generate a sequence (B_n) of pairwise disjoint sets in $\mathcal{A}$ with $B_n \subset A$ and $\nu(B_n) \leq -\epsilon$ for all n. If $B = \cup_{n=1}^{\infty} B_n$, then $B \in \mathcal{A}$, $B \subseteq A$, and $\nu(B) = -\infty$, contradicting XVII.9.1.10.. 👍

XVII.9.1.13. LEMMA. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$ not taking the value $+\infty$. If $A \in \mathcal{A}$, then A contains a subset P with $P \in \mathcal{A}$, $\nu(P) \geq \nu(A)$, and P a positive set for ν.

PROOF: Let $A \in \mathcal{A}$. If $\nu(A) \leq 0$ (in particular, if $\nu(A) = -\infty$), we may take $P = \emptyset$. Assume $\nu(A) > 0$ (in particular, $\nu(A) \in \mathbb{R}$). Generate a decreasing sequence (A_n) of subsets of A by setting $A_0 = A$ and $A_{n+1} = (A_n)_\epsilon$ (XVII.9.1.12.) with $\epsilon = \frac{1}{n+1}$. Set $P = \cap_{n=1}^{\infty} A_n$. Then $P \in \mathcal{A}$, $P \subseteq A$, and

$$\nu(P) = \lim_{n\to\infty} \nu(A_n) \geq \nu(A)$$

by XVII.9.1.11.(ii) and the fact that $\nu(A_n) \geq \nu(A)$ for all n. Also, if $B \in \mathcal{A}$, $B \subseteq P$, then $B \subseteq A_n$ for each n, and so $\nu(B) > -\frac{1}{n}$ for all $n \geq 1$. Hence $\nu(B) \geq 0$. Thus P is positive for ν. 👍

We can now prove that a signed measure which does not take an infinite value is actually bounded on that side.

XVII.9.1.14. LEMMA. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$. If ν does not take the value $+\infty$, then $\{\nu(A) : A \in \mathcal{A}\}$ is bounded above. If ν does not take the value $-\infty$, then $\{\nu(A) : A \in \mathcal{A}\}$ is bounded below.

PROOF: We prove only the first statement; the second follows by replacing ν by $-\nu$.

Suppose for each n there is a set $A_n \in \mathcal{A}$ with $\nu(A_n) > n$. By XVII.9.1.13. there is a set $P_n \in \mathcal{A}$ with $\nu(P_n) \geq \nu(A_n) > n$ which is positive for ν. Set $P = \cup_{n=1}^{\infty} P_n$. Then $P \in \mathcal{A}$ and is positive for ν; thus $\nu(P) \geq \nu(P_n) > n$ for all n, i.e. $\nu(P) = +\infty$. 👍

We can now prove Theorem XVII.9.1.6..

PROOF: Replacing ν by $-\nu$ if necessary, we may assume that ν does not take the value $+\infty$. Let $M = \sup\{\nu(A) : A \in \mathcal{A}\}$. Then $0 \leq M < +\infty$ by XVII.9.1.14. ($M \geq 0$ since $\nu(\emptyset) = 0$). Let (A_n) be a sequence of sets in $\mathcal{A}$ with $\sup_n \nu(A_n) = M$. By XVII.9.1.13., for each n there is a set $P_n \in \mathcal{A}$ with $\nu(P_n) \geq \nu(A_n)$ which is positive for ν. Then $\sup_n \nu(P_n) = M$. Set $P = \cup_{n=1}^{\infty} P_n$. Then P is a positive set for ν, and $\nu(P) \geq \nu(P_n)$ for all n; thus $\nu(P) = M$. Let $N = X \setminus P$. If $B \in \mathcal{A}$, $B \subseteq N$, then

$$M + \nu(B) = \nu(P \cup B) \leq M$$

and hence $\nu(B) \leq 0$; thus N is a negative set for ν, and (P, N) is a Hahn decomposition for ν.

To show uniqueness, if (P', N') is another Hahn decomposition, let A be a subset of $P \setminus P'$. Then $\nu(A) \geq 0$ since $A \subseteq P$, and $\nu(A) \leq 0$ since $A \subseteq N'$. Hence $P \setminus P'$ is a null set. Similarly, $P' \setminus P$, $N \setminus N'$, and $N' \setminus N$ are null sets.

This completes the proof of XVII.9.1.6.. 👍

The proof actually shows the following:

XVII.9.1.15. COROLLARY. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$. If $L = \inf\{\nu(A) : A \in \mathcal{A}\}$ and $M = \sup\{\nu(A) : A \in \mathcal{A}\}$, then:

(i) At least one of L and M is finite (note that $L \leq 0$ and $M \geq 0$ since $\nu(\emptyset) = 0$).

(ii) A pair (P, N) of sets in $\mathcal{A}$ is a Hahn decomposition for ν if and only if $N = X \setminus P$, $\nu(P) = M$, and $\nu(N) = L$.

(iii) There exists at least one pair (P, N) as in (ii).

See [Kön97] for a detailed discussion of signed premeasures and versions of the Carathéodory Extension Theorem.

The Total Variation of a Signed Measure

XVII.9.1.16. DEFINITION. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$, with Jordan decomposition $\nu = \nu^+ - \nu^-$. The measure $|\nu| = \nu^+ + \nu^-$ is called the *total variation measure*, or *total variation*, of ν.

The next proposition is nearly trivial (given our previous results). Part (ii) is preparation for the definition of the total variation of a complex measure (XVII.9.2.6.), which is less trivial. This part is strongly reminiscent of the definition of the total variation of a function (). (This similarity is no accident! See ().)

XVII.9.1.17. PROPOSITION. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$. If $E \in \mathcal{A}$, then

(i) $|\nu(E)| \leq |\nu|(E)$.

(ii) $|\nu|(E) = \sup(\sum_{n=1}^{\infty} |\nu(E_n)|)$, where the supremum runs over all measurable partitions $\{E_n\}$ of E.

PROOF: (i) is obvious. For (ii), let S be the supremum. If $\{E_n\}$ is a measurable partition of E, then since $|\nu|$ is a measure,

$$\sum_{n=1}^{\infty} |\nu(E_n)| \leq \sum_{n=1}^{\infty} |\nu|(E_n) = |\nu| \left(\bigcup_{n=1}^{\infty} E_n \right) = |\nu|(E)$$

using (i), so $S \leq |\nu|(E)$. On the other hand, if (P, N) is a Hahn decomposition for ν, we have

$$|\nu|(E) = \nu^+(E) + \nu^-(E) = |\nu(E \cap P)| + |\nu(E \cap N)| \leq S \ .$$

An easy consequence of (ii) is:

XVII.9.1.18. COROLLARY. Let ν_1 and ν_2 be signed measures on a measurable space $(X, \mathcal{A})$. Suppose $\nu_1 + \nu_2$ is well defined as a signed measure (i.e. both ν_1 and ν_2 are bounded above or both are bounded below, e.g. if at least one is finite). Then $|\nu_1 + \nu_2| \leq |\nu_1| + |\nu_2|$ (i.e. $|\nu_1 + \nu_2|(E) \leq |\nu_1|(E) + |\nu_2|(E)$ for all $E \in \mathcal{A}$).

Note that there is no simple way to obtain a Hahn decomposition for $\nu_1 + \nu_2$ from Hahn decompositions for ν_1 and ν_2 in general.



XVII.9.2. Complex Measures

XVII.9.2.1. DEFINITION. If $(X, \mathcal{A})$ is a measurable space, a *complex measure* on $(X, \mathcal{A})$ is a function $\nu : \mathcal{A} \to \mathbb{C}$ such that

(i) $\nu(\emptyset) = 0$.

(ii) If $A_1, A_2, \ldots$ is a sequence of pairwise disjoint sets in $\mathcal{A}$, then $\sum_{n=1}^{\infty} \nu(A_n)$ converges to $\nu(\cup_{n=1}^{\infty} A_n)$.

XVII.9.2.2. EXAMPLES. (i) Any finite signed measure (in particular, any finite measure) is a complex measure.
(ii) If ν_1 and ν_2 are finite signed measures on $(X, \mathcal{A})$, then $\nu = \nu_1 + i\nu_2$ is a complex measure on $(X, \mathcal{A})$. More generally, any (finite) complex linear combination of complex measures on $(X, \mathcal{A})$ is a complex measure on $(X, \mathcal{A})$.

The following converse to (ii) essentially reduces the theory of complex measures to the theory of finite signed measures, and hence to the theory of ordinary finite measures:

XVII.9.2.3. PROPOSITION. Let ν be a complex measure on a measurable space $(X, \mathcal{A})$. For $E \in \mathcal{A}$, define $\nu_{re}(E) = Re(\nu(E))$ and $\nu_{im}(E) = Im(\nu(E))$. Then ν_{re} and ν_{im} are finite signed measures on $(X, \mathcal{A})$, called the *real* and *imaginary parts* of ν respectively.

Conversely, if ν_1 and ν_2 are finite signed measures on $(X, \mathcal{A})$, and $\nu = \nu_1 + i\nu_2$, then $\nu_{re} = \nu_1$, $\nu_{im} = \nu_2$.

XVII.9.2.4. To a complex measure ν there are thus four canonically associated finite measures ν_{re}^+, ν_{re}^-, ν_{im}^+, and ν_{im}^-. Note, however, that there may be overlap between the supports of $\nu_{re}^\pm$ and $\nu_{im}^\pm$, i.e. the measures $|\nu_{re}|$ and $|\nu_{im}|$ may not be mutually singular. Thus there is no analog of a Hahn decomposition of X into "real" and "imaginary" subsets for ν in general.

XVII.9.2.5. There is a slight difficulty in making a good definition of the total variation of a complex measure. A reasonable approach is to take the measure

$$|\nu_{re}| + |\nu_{im}| = \nu_{re}^+ + \nu_{re}^- + \nu_{im}^+ + \nu_{im}^-$$

but this measure does not have all the desirable properties of a total variation. Instead it is customary to take the following definition:

XVII.9.2.6. DEFINITION. Let ν be a complex measure on a measurable space $(X, \mathcal{A})$. If $E \in \mathcal{A}$, define $|\nu|(E) = \sup(\sum_{n=1}^\infty |\nu(E_n)|)$, where the supremum runs over all measurable partitions $\{E_n\}$ of E.

XVII.9.2.7. THEOREM. Let ν be a complex measure on a measurable space $(X, \mathcal{A})$. Then

(i) $|\nu|$ is a measure on $(X, \mathcal{A})$.

(ii) If $E \in \mathcal{A}$, then $|\nu(E)| \le |\nu|(E)$.

(iii) If $E \in \mathcal{A}$, then

$$\max(|\nu_{re}|(E), |\nu_{im}|(E)) \le |\nu|(E) \le |\nu_{re}|(E) + |\nu_{im}|(E) \ .$$

In particular, $|\nu|$ is a finite measure, and since

$$\frac{1}{2}\big[|\nu_{re}|(E) + |\nu_{im}|(E)\big] \le \max(|\nu_{re}|(E), |\nu_{im}|(E))$$

we have

$$\frac{1}{2}\big[|\nu_{re}|(E) + |\nu_{im}|(E)\big] \le |\nu|(E) \le |\nu_{re}|(E) + |\nu_{im}|(E) \ .$$

PROOF: (ii): This is trivial since $\{E\}$ is a measurable partition of E.
(iii): Note that if $x, y \in \mathbb{R}$, then

$$\frac{1}{2}\big[|x| + |y|\big] \le \max(|x|, |y|) \le |x + iy| \le |x| + |y|$$

(). If $\{E_n\}$ is a measurable partition of E, then

$$\max\left(\sum_{n=1}^\infty |\nu_{re}(E_n)|, \sum_{n=1}^\infty |\nu_{im}(E_n)|\right) \le \sum_{n=1}^\infty \big[\max(|\nu_{re}(E_n)|, |\nu_{im}(E_n)|)\big]$$

$$\le \sum_{n=1}^\infty |\nu(E_n)| \le \sum_{n=1}^\infty \big[|\nu_{re}(E_n)| + |\nu_{im}(E_n)|\big] = \sum_{n=1}^\infty |\nu_{re}(E_n)| + \sum_{n=1}^\infty |\nu_{im}(E_n)|$$

and the result follows by taking the supremum over all measurable partitions of E and applying XVII.9.1.17.(ii). In particular, since $|\nu_{re}|(E)$ and $|\nu_{im}|(E)$ are finite, so is $|\nu|(E)$.

(i): It is obvious that $|\nu|(\emptyset) = 0$. Let $\{A_m\}$ be a pairwise disjoint sets in $\mathcal{A}$, and let $A = \cup_{m=1}^{\infty} A_m$. Fix $\epsilon > 0$. Let $\{E_n\}$ be a measurable partition of A with $\sum_{n=1}^{\infty} |\nu(E_n)| > |\nu|(A) - \epsilon$ ($|\nu|(A)$ is finite by (iii)). Set $A_{n,m} = E_n \cap A_m$ for each n, m. Then

$$|\nu|(A) - \epsilon \le \sum_{n=1}^{\infty} |\nu(E_n)| = \sum_{n=1}^{\infty} \left| \sum_{m=1}^{\infty} \nu(A_{n,m}) \right|$$

$$\le \sum_{n=1}^{\infty} \left[\sum_{m=1}^{\infty} |\nu(A_{n,m})| \right] = \sum_{m=1}^{\infty} \left[\sum_{n=1}^{\infty} |\nu(A_{n,m})| \right] \le \sum_{m=1}^{\infty} |\nu|(A_m)$$

(the equality in the last line uses ().) Since $\epsilon > 0$ is arbitrary, $|\nu|(A) \le \sum_{m=1}^{\infty} |\nu|(A_m)$. Conversely, if $\epsilon > 0$, for each m let $\{E_{n,m}\}$ be a measurable partition of A_m with

$$\sum_{n=1}^{\infty} |\nu(E_{n,m})| > |\nu|(A_m) - 2^{-m}\epsilon \ .$$

Then $\{E_{n,m}\}$ is a measurable partition of A, so

$$\left(\sum_{m=1}^{\infty} |\nu|(A_m) \right) - \epsilon = \sum_{m=1}^{\infty} [|\nu|(A_m) - 2^{-m}\epsilon] < \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |\nu(E_{n,m})| \le |\nu|(A) \ .$$

Since $\epsilon > 0$ is arbitrary, $\sum_{m=1}^{\infty} |\nu|(A_m) \le |\nu|(A)$. 👍

XVII.9.2.8. PROPOSITION. Let ν_1 and ν_2 be complex measures on $(X, \mathcal{A})$, and let $\alpha \in \mathbb{C}$. Then

(i) $|\nu_1 + \nu_2| \le |\nu_1| + |\nu_2|$ (i.e. $|\nu_1 + \nu_2|(E) \le |\nu_1|(E) + |\nu_2|(E)$ for all $E \in \mathcal{A}$).

(ii) $|\alpha\nu_1| = |\alpha||\nu_1|$.

XVII.9.3. Integration With Respect to Signed and Complex Measures

There is nothing difficult or complicated about extending integration to the case of signed or complex measures; we need only make the right definitions, which are fairly obvious, guided by the desire to preserve linearity.

The only slight difficulty is deciding what should be meant by an integrable function. But the definition here is fairly obvious too, because of the following fact:

XVII.9.3.1. PROPOSITION. (i) Let ν be a signed measure on a measurable space $(X, \mathcal{A})$, and $f : X \to \mathbb{R}$ an $\mathcal{A}$-measurable function. Then $f \in \mathcal{L}^1(X, \mathcal{A}, |\nu|)$ if and only if $f \in \mathcal{L}^1(X, \mathcal{A}, \nu^+)$ and $f \in \mathcal{L}^1(X, \mathcal{A}, \nu^-)$.
(ii) Let ν be a complex measure on a measurable space $(X, \mathcal{A})$, and $f : X \to \mathbb{C}$ an $\mathcal{A}$-measurable function. Then $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, |\nu|)$ if and only if $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, |\nu_{re}|)$ and $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, |\nu_{im}|)$.

PROOF: Part (i) is obvious since $\int_X |f|\, d|\nu| = \int_X |f|\, d\nu^+ + \int_X |f|\, d\nu^-$, and (ii) follows easily from the same argument and XVII.9.2.7.(iii). 👍

XVII.9.3.2. DEFINITION. Let ν be a signed measure or complex measure on a measurable space $(X, \mathcal{A})$, and f a real- or complex-valued $\mathcal{A}$-measurable function. Then $f \in \mathcal{L}^1(X, \mathcal{A}, \nu)$ [$\mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$ if f is complex-valued] if $|f| \in \mathcal{L}^1(X, \mathcal{A}, |\nu|)$.

We then make the following obvious definition:

XVII.9.3.3. DEFINITION. (i) Let ν be a signed measure on a measurable space $(X, \mathcal{A})$, and $f \in \mathcal{L}^1(X, \mathcal{A}, \nu)$ or $\mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$. Then

$$\int_X f \, d\nu = \int_X f \, d\nu^+ - \int_X f \, d\nu^- \ .$$

(ii) Let ν be a complex measure on a measurable space $(X, \mathcal{A})$, and $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$. Then

$$\int_X f \, d\nu = \int_X f \, d\nu_{re} + i \int_X f \, d\nu_{im} \ .$$

(iii) If $A \in \mathcal{A}$, then in either of the above cases $\int_A f \, d\nu = \int_X f \chi_A \, d\nu$.

XVII.9.3.4. The integrals defined in this way have almost all the properties of integrals with respect to ordinary measures; the only obvious exceptions are positivity and monotonicity. It thus only makes sense to integrate $\mathcal{L}^1$-functions, not arbitrary nonnegative measurable functions.

We list the important properties of this extended integral. The proofs are simple applications of the results for nonnegative measures, and are mostly omitted.

XVII.9.3.5. PROPOSITION. Let ν be a signed or complex measure on a measurable space $(X, \mathcal{A})$. If $f, g \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$, then

$$\int_A (f + g) \, d\nu = \int_A f \, d\nu + \int_A g \, d\nu$$

for all $A \in \mathcal{A}$. If $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$ and $\alpha \in \mathbb{C}$, then

$$\int_A \alpha f \, d\nu = \alpha \int_A f \, d\nu$$

for all $A \in \mathcal{A}$.

For the proof of the last statement, see Exercise XVII.9.5.2..

XVII.9.3.6. If ν is a signed or complex measure and $f \in \mathcal{L}_1(\nu)$, then in general

$$\left| \int_X f \, d\nu \right| \not\leq \int |f| \, d\nu$$

(the right side is not necessarily nonnegative or even real). However, we have:

XVII.9.3.7. PROPOSITION. Let ν be a signed or complex measure on a measurable space $(X, \mathcal{A})$, and let $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$. Then

$$\left| \int_A f \, d\nu \right| \leq \int_A |f| \, d|\nu|$$

for any $A \in \mathcal{A}$.

Because of the lack of monotonicity, there are no versions of the Monotone Convergence Theorem or Fatou's Lemma for signed or complex measures. But the Dominated Convergence Theorem holds with an identical statement:

XVII.9.3.8. THEOREM. [DOMINATED CONVERGENCE THEOREM] Let ν be a signed or complex measure on a measurable space $(X, \mathcal{A})$. Let (f_n) be a sequence of functions in $\mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$ converging pointwise a.e. to a function f. If there is a $g \in \mathcal{L}^1_+(X, \mathcal{A}, \nu)$ with $|f_n| \leq g$ a.e. for all n, then

$$\int_X f \, d\nu = \lim_{n \to \infty} \int_X f_n \, d\nu \ .$$

There is a version for infinite series. We omit the general statement, but note the following important special case:

XVII.9.3.9. PROPOSITION. Let ν be a signed or complex measure on a measurable space $(X, \mathcal{A})$, and let $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$. Let $\{A_n\}$ be a sequence of pairwise disjoint sets in $\mathcal{A}$, with $A = \cup_{n=1}^{\infty} A_n$. Then the series

$$\sum_{n=1}^{\infty} \int_{A_n} f \, d\nu$$

converges absolutely to $\int_A f \, d\nu$.

XVII.9.3.10. As a consequence, if ν is a complex measure on $(X, \mathcal{A})$, and $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$, and σ is defined on $\mathcal{A}$ by $\sigma(A) = \int_A f \, d\nu$, then σ is also a complex measure on A. In this case, we have that $|\sigma|(A) = \int_A |f| \, d|\nu|$ (**??**).

XVII.9.4. Another Approach via Local Measures

There is another way to avoid the $\infty - \infty$ problem: considering signed and complex local measures defined on δ-rings () rather than σ-algebras. This approach allows for a more general theory which is important, for example, in the general form of the Riesz Representation Theorem ().

XVII.9.4.1. DEFINITION. Let X be a set, and $\mathcal{D}$ a δ-ring of subsets of X. A *(finite) signed local measure* on $(X, \mathcal{D})$ is a function $\nu : \mathcal{D} \to \mathbb{R}$ such that

(i) $\nu(\emptyset) = 0$.

(ii) If $\{A_n\}$ is a sequence of pairwise disjoint sets in $\mathcal{D}$ with $A := \cup_{n=1}^{\infty} A_n \in \mathcal{D}$, then

$$\nu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \nu(A_n)$$

(in particular, the infinite series in (ii) converges).

XVII.9.4.2. We could allow ν to take either the value $+\infty$ or $-\infty$, but not both; but the utility of this generalization is modest at best since working on δ-rings is primarily designed to eliminate infinite values of the measure, so we will stick to finite signed local measures, and the "finite" will be understood. Note that in the definition, the series $\sum_{n=1}^{\infty} \nu(A_n)$ necessarily converges absolutely under the hypotheses, since

$$A_+ = \cup\{A_n : \nu(A_n) \geq 0\} \quad \text{and} \quad A_- = \cup\{A_n : \nu(A_n) < 0\}$$

are in $\mathcal{D}$.

XVII.9.4.3. EXAMPLES. (i) A finite signed measure on $(X, \mathcal{A})$ is a signed local measure, with $\mathcal{D} = \mathcal{A}$.

(ii) More generally, let μ be a signed measure on $(X, \mathcal{A})$, not necessarily finite. Set $\mathcal{D} = \{A \in \mathcal{A} : \mu(A) \in \mathbb{R}\}$. Then $\mathcal{D}$ is a δ-ring, and $\nu = \mu|_{\mathcal{D}}$ is a signed local measure (XVII.9.1.10.).

(iii) Here is an example not of the form in (ii), which illustrates the main point in passing to signed local measures. Let $\mathcal{D}$ be the collection of Lebesgue measurable sets in $\mathbb{R}$ of finite Lebesgue measure, and for $A \in \mathcal{D}$ set

$$\nu(A) = \lambda(A \cap [0, \infty)) - \lambda(A \cap (-\infty, 0)) \ .$$

Then $\mathcal{D}$ is a δ-ring of subsets of $\mathbb{R}$, and ν is a signed local measure on $\mathcal{D}$. This $\mathcal{D}$ can be replaced by the δ-ring of bounded Lebesgue measurable sets or the δ-ring of bounded Borel sets (or Borel sets of finite Lebesgue measure). Note that there is no way to work with a signed measure in this case which avoids the $\infty - \infty$ problem.

(iv) A local measure on a δ-ring $\mathcal{D}$ is a signed local measure. Such a local measure extends canonically to a measure on the σ-algebra generated by $\mathcal{D}$, so there is little advantage to considering local measures on δ-rings. But Example (iii) shows that this is not true for signed local measures in general.

(v) Since we are only considering finite signed local measures, any linear combination of signed local measures on the *same* δ-ring $\mathcal{D}$ is a signed local measure on $\mathcal{D}$. In particular, the difference of two finite local measures on $\mathcal{D}$ is a local signed measure on $\mathcal{D}$.

XVII.9.4.4. If ν is a signed local measure on a δ-ring $\mathcal{D}$ of subsets of X, and $Y \in \mathcal{D}$, then the restriction of ν to the σ-algebra $\mathcal{D}_Y$ is a finite signed measure on $(Y, \mathcal{D}_Y)$. Thus, although the strict analog of XVII.9.1.14. is false in general for signed local measures, i.e. a signed local measure need not be bounded, as shown by Example XVII.9.4.3.(iii), a signed local measure is "locally bounded," and locally has all the other properties of a signed measure.

Jordan and Hahn Decompositions of Signed Local Measures

XVII.9.4.5. If ν is a signed local measure on a δ-ring $\mathcal{D}$, and $Y \in \mathcal{D}$, then ν defines a finite signed measure on $(Y, \mathcal{D}_Y)$, hence there is a Jordan decomposition for $\nu_Y := \nu|_{\mathcal{D}_Y}$, i.e. there are unique finite measures ν_Y^+ and ν_Y^- on $(Y, \mathcal{D}_Y)$ with $\nu_Y = \nu_Y^+ - \nu_Y^-$. If $Y, Z \in \mathcal{D}$, $Y \subseteq Z$, then ν_Y is the restriction of ν_Z to $\mathcal{D}_Y$, and it follows that $\nu_Y^\pm$ is the restriction of $\nu_Z^\pm$ to $\mathcal{D}_Y$. Since $\mathcal{D}$ is a directed set under inclusion, the $\nu_Y^\pm$ define finite local measures $\nu^\pm$ on $\mathcal{D}$ with $\nu = \nu^+ - \nu^-$. Thus a signed local measure has a well-defined and unique Jordan decomposition as a difference of finite local measures.

Both ν^+ and ν^- extend canonically to measures on the σ-algebra generated by $\mathcal{D}$. But the extensions are not finite in general, and ν may not extend as the difference due to the $\infty - \infty$ problem (cf. Example XVII.9.4.3.(iii)).

XVII.9.4.6. Although Jordan decomposition works nicely for signed local measures, Hahn decomposition is much more problematic since the Hahn decomposition of a signed measure is only uniquely determined up to null sets. And even if the Hahn decompositions (A_Y, B_Y) of the ν_Y can be chosen consistently, so a Hahn decomposition (A, B) can be defined as $A = \cup_Y A_Y$, $B = \cup_Y B_Y$, the A and B are only locally $\mathcal{D}$-measurable in general, and are only uniquely defined up to "locally null sets."

The Total Variation of a Signed Local Measure

XVII.9.4.7. If ν is a signed local measure on a δ-ring $\mathcal{D}$, with Jordan decomposition $\nu^+ - \nu^-$, we can define the total variation $|\nu|$ to be $\nu^+ + \nu^-$. Then $|\nu|$ is a finite local measure on $\mathcal{D}$. The exact analog of XVII.9.1.17. holds, with the same proof.

Complex Local Measures

XVII.9.4.8. The extension of the definition of signed local measures to complex measures presents no difficulty: just change $\mathbb{R}$ to $\mathbb{C}$ in Definition XVII.9.4.1.. The real and imaginary parts of a complex local measure are signed local measures.

XVII.9.4.9. The total variation of a complex local measure can be defined exactly as in XVII.9.2.6., and XVII.9.2.7. holds with trivial modifications of the proof (left to the reader).

Integration With Respect to Signed and Complex Local Measures

XVII.9.5. Exercises

XVII.9.5.1. (a) Let $(X, \mathcal{A}, \mu)$ be a measure space, and $f \in \mathcal{L}^1_{\mathbb{R}}(X, \mathcal{A}, \mu)$. Define $\nu(A) = \int_A f\, d\mu$ as in XVII.9.1.2. and XVII.9.1.8.. Show that, for all $A \in \mathcal{A}$, $|\nu|(A) = \int_A |f|\, d\mu$.

(b) Show the same for $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \mu)$. [Hint: for $n \in \mathbb{N}$ and $1 \leq k \leq n$, consider the set

$$A_{n,k} = \left\{ x \in A : \frac{2\pi(k-1)}{n} \leq arg(f(x)) < \frac{2\pi k}{n} \right\} .]$$

XVII.9.5.2. (a) Let ν be a complex measure on a measurable space $(X, \mathcal{A})$, and $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$. Show that

$$\int_A f\, d\nu = \left[\int_A Re(f)\, d\nu_{re} - \int_A Im(f)\, d\nu_{im}\right] + i\left[\int_A Re(f)\, d\nu_{im} + \int_A Im(f)\, d\nu_{re}\right]$$

for any $A \in \mathcal{A}$.

(b) Show that if $\alpha \in \mathbb{C}$, then

$$\int_X \alpha f\, d\nu = \alpha \int_X f\, d\nu\ .$$

XVII.9.5.3. Let ν be a signed measure on a measurable space $(X, \mathcal{A})$, and $f \in \mathcal{L}^1_{\mathbb{R}}(X, \mathcal{A}, \nu)$. Define a signed measure σ by $\sigma(A) = \int_A f\, d\nu$ for $A \in \mathcal{A}$ ().

(a) Find a Hahn decomposition for σ, and the Jordan decomposition.

(b) Show that $|\sigma|(A) = \int_A |f|\, d|\nu|$ for $a \in \mathcal{A}$.

XVII.9.5.4. Let ν be a complex measure on $(X, \mathcal{A})$, and $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$. Define a complex measure σ by $\sigma(A) = \int_A f\, d\nu$ for $A \in \mathcal{A}$ (). If g is an $\mathcal{A}$-measurable function, show that g is σ-integrable if and only if fg is ν-integrable, and that

$$\int_A g\, d\sigma = \int_A fg\, d\nu$$

for all $A \in \mathcal{A}$.

XVII.9.5.5. Try to extend the theory of complex measures to include measures of the form $\nu_1 + i\nu_2$, where ν_1 and ν_2 are signed measures, not necessarily finite. Try to give a definition as in XVII.9.1.1. and XVII.9.2.1. which covers this case (clumsy and unnatural, isn't it?) Discuss the properties which can and cannot be established in this case, and the difficulties and limitations of such a theory.

XVII.9.5.6. A finitely additive signed measure on an algebra can be finite but unbounded, and can fail to have a Hahn decomposition.

(a) Let $\mathcal{A}$ be the algebra of finite/cofinite subsets of $\mathbb{N}$. If $A \in \mathcal{A}$ is finite, let $\nu(A)$ be the number of elements in A; if A is cofinite, let $\nu(A) = -n$, where n is the number of elements in A^c. Show that ν is a finitely additive signed measure on $(\mathbb{N}, \mathcal{A})$.

(b) Show that there is no Hahn decomposition for ν.

XVII.9.5.7. **Products of Signed Measures.** Let μ and ν be finite signed measures on measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ respectively, with Hahn decompositions (P_μ, N_μ) and (P_ν, N_ν) and Jordan decompositions μ^+, μ^-, ν^+, ν^-. Let

$$(\mu \times \nu)^+ = (\mu^+ \times \nu^+) + (\mu^- \times \nu^-)$$
$$(\mu \times \nu)^- = (\mu^+ \times \nu^-) + (\mu^- \times \nu^+)$$

be corresponding measures on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$.

(a) Show that $(\mu \times \nu)^+$ and $(\mu \times \nu)^-$ are mutually singular finite measures on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$.

(b) Define $\mu \times \nu = (\mu \times \nu)^+ - (\mu \times \nu)^-$. Show that $\mu \times \nu$ is a finite signed measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ with Hahn decomposition

$$([P_\mu \times P_\nu] \cup [N_\mu \times N_\nu], [P_\mu \times N_\nu] \cup [N_\mu \times P_\nu])$$

satisfying $(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$ for $A \in \mathcal{A}$, $B \in \mathcal{B}$. Show that this measure is the unique finite signed measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ with this property.

(c) Discuss the difficulties in defining $\mu \times \nu$ if μ and ν are signed measures which are not finite.

XVII.9.5.8. Products of Complex Measures. Let $\mu = \mu_{re} + i\mu_{im}$ and $\nu = \nu_{re} + i\nu_{im}$ be complex measures on measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ respectively. Let

$$(\mu \times \nu)_{re} = (\mu_{re} \times \nu_{re}) - (\mu_{im} \times \nu_{im})$$

$$(\mu \times \nu)_{im} = (\mu_{re} \times \nu_{im}) + (\mu_{im} \times \nu_{re})$$

be corresponding signed measures on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$, defined as in XVII.9.5.7..

(a) Show that $(\mu \times \nu)_{re}$ and $(\mu \times \nu)_{im}$ are finite signed measures on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$.

(b) Show that $\mu \times \nu = (\mu \times \nu)_{re} + i(\mu \times \nu)_{im}$ is a complex measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ satisfying $(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$ for $A \in \mathcal{A}$, $B \in \mathcal{B}$. Show that this measure is the unique complex measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$ with this property.

XVII.9.5.9. Make an appropriate definition of signed and complex premeasures (XVII.4.5.2.) on semirings, and prove analogs of Carathéodory's Theorem (XVII.4.3.4.) and the Extension Theorem (XVII.4.5.8.) in this case. Use this to give an alternate proof of XVII.9.5.7. (and XVII.9.5.8.) which does not use Jordan decomposition.

XVII.9.6. Signed and Complex Measures on σ-Rings

There is no particular advantage in considering ordinary measures on σ-rings instead of σ-algebras, since an ordinary measure on a σ-ring can be extended in a canonical way to the σ-algebra it generates (). But there is utility in considering signed and complex measures on σ-rings, since they arise naturally and cannot generally be extended to the generated σ-algebra, essentially because of the $\infty - \infty$ problem. In this section, we describe the theory of such measures, which of course includes the theory of ordinary measures on σ-rings. We concentrate on the crucial case of Borel measures on locally compact Hausdorff spaces.

XVII.10. Limits of Measures

In this section, we discuss the following question: if (μ_i) is a sequence (or net) of measures, and $\mu_i \to \mu$, is μ a measure?

XVII.10.1. General Considerations

We must first make sense of what it means to say that $\mu_i \to \mu$. There is one obvious minimal sense to interpret this:

XVII.10.1.1. Definition. Let X be a set, $\mathcal{A}$ a collection of subsets of X, and Y a topological space. Then a sequence (or net) (f_i) of functions from $\mathcal{A}$ to Y *converges setwise* to the function $f : \mathcal{A} \to Y$ if $f_i(A) \to f(A)$ for all $A \in \mathcal{A}$. In particular, if (μ_i) is a sequence (or net) of measures on a measurable space $(X, \mathcal{A})$, and $\mu : \mathcal{A} \to [0, \infty]$, then $\mu_i \to \mu$ setwise if $\mu_i(A) \to \mu(A)$ for all $A \in \mathcal{A}$. Setwise convergence of a sequence (or net) (ν_i) of complex measures to a function $\nu : \mathcal{A} \to \mathbb{C}$ is defined in the same way.

In other words: a measure on $(X, \mathcal{A})$ may be thought of as an element of the product space $[0, \infty]^{\mathcal{A}}$, the set of all functions from $\mathcal{A}$ to $[0, \infty]$. Then $\mu_i \to \mu$ setwise if $\mu_i \to \mu$ in the product topology. Similarly, a complex measure on $(X, \mathcal{A})$ is an element of $\mathbb{C}^{\mathcal{A}}$, and setwise convergence is convergence in the product topology.

There are stronger notions of convergence; see, for example, Exercise XVII.10.4.1..

So the basic question is:

XVII.10.1.2. Question. If (μ_i) is a sequence (or net) of measures (ordinary or complex), and $\mu_i \to \mu$ setwise, is μ is also a measure (or complex measure)?

This question can be rephrased: is the set of measures a closed subset of $[0, \infty]^{\mathcal{A}}$? Is the set of complex measures a closed subset of $\mathbb{C}^{\mathcal{A}}$?

The following simple examples show that the answer is no in general.

XVII.10.1.3. Examples. (i) On $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, let μ_n be counting measure on $\{m : m > n\}$, regarded as a measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$ by setting $\mu_n(E) = 0$ if $E \subseteq \{1, \dots, n\}$. Then each μ_n is a σ-finite measure, $(\mu_n(E))$ is decreasing for each E, and $\mu(E) = \lim_{n\to\infty} \mu_n(E)$ exists for each $E \subseteq \mathbb{N}$. But $\mu(E) = 0$ if E is finite, and $\mu(\mathbb{N}) = \infty$ (in fact, $\mu(E) = \infty$ if E is any infinite set), so μ is not a measure.

(ii) Again on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$, let δ_n be the unit point mass at n, i.e. for $E \subseteq \mathbb{N}$, $\delta_n(E) = 1$ if $n \in E$ and $\delta_n(E) = 0$ if $n \notin E$. Each δ_n gives a point of $\{0, 1\}^{\mathcal{P}(\mathbb{N})}$, which is compact in the product topology; thus there is a subnet of the sequence (δ_n) which converges setwise to a function $\mu : \mathcal{P}(\mathbb{N}) \to \{0, 1\}$. But $\mu(E) = 0$ for all finite E, while $\mu(\mathbb{N}) = 1$, and thus μ cannot be a measure.

Example (ii) is rather generic: there are grave difficulties in taking limits of nets of measures which are not sequences. But, rather remarkably, often (i.e. under mild hypotheses) it will turn out that limits of *sequences* of measures are measures.

A setwise limit of measures is always at least a finitely additive measure:

XVII.10.1.4. PROPOSITION. Let $(X, \mathcal{A})$ be a measurable space. Then the set of finitely additive measures on $(X, \mathcal{A})$ is closed under setwise limits (i.e. is a closed subset of $[0, \infty]^{\mathcal{A}}$). The set of finitely additive complex measures is a closed subset of $\mathbb{C}^{\mathcal{A}}$.

This is an immediate consequence of ().

So the whole difficulty is with countable additivity.

XVII.10.2. Monotone and Dominated Convergence

The first result is an analog of the Monotone Convergence Theorem:

XVII.10.2.1. THEOREM. Let (μ_i) be a net of measures on a measurable space $(X, \mathcal{A})$. If $(\mu_i(A))$ is increasing for each $A \in \mathcal{A}$, then $\mu = \lim \mu_i = \sup \mu_i$ (defined by $\mu(A) = \sup_i \mu_i(A)$ for $A \in \mathcal{A}$) is a measure on $(X, \mathcal{A})$.

PROOF: We only need to show countable additivity. Suppose (A_n) is a pairwise disjoint sequence in $\mathcal{A}$, and set $A = \cup_{n=1}^{\infty} A_n$. We have

$$\sum_{n=1}^{\infty} \mu(A_n) = \sup_N \sum_{n=1}^{N} \mu(A_n) = \sup_N \left[\mu \left(\bigcup_{n=1}^{N} A_n \right) \right] \leq \mu(A)$$

by finite additivity and monotonicity. To show the reverse inequality, first suppose $\mu(A) < \infty$. Let $\epsilon > 0$ and choose i_0 such that $\mu_{i_0}(A) > \mu(A) - \epsilon$. Then we have

$$\mu(A) - \epsilon < \mu_{i_0}(A) = \sum_{n=1}^{\infty} \mu_{i_0}(A_n) \leq \sum_{n=1}^{\infty} \mu(A_n) \ .$$

Since $\epsilon > 0$ is arbitrary, $\mu(A) \leq \sum_{n=1}^{\infty} \mu(A_n)$.

Now suppose $\mu(A) = \infty$. Fix $M > 0$, and choose i_0 such that $\mu_{i_0}(A) > M$. We then have

$$M < \mu_{i_0}(A) = \sum_{n=1}^{\infty} \mu_{i_0}(A_n) \leq \sum_{n=1}^{\infty} \mu(A_n)$$

so $\sum_{n=1}^{\infty} \mu(A_n) = +\infty$. 👍

This implies that arbitrary sums of measures are measures (finite sums of measures are obviously measures ()).

XVII.10.2.2. COROLLARY. Let $\{\mu_i : i \in I\}$ be a set of measures on a measurable space $(X, \mathcal{A})$. For each $A \in \mathcal{A}$, define

$$\mu(A) = \sum_{i \in I} \mu_i(A) = \sup \left\{ \sum_{i \in F} \mu_i(A) : F \subseteq I \text{ finite} \right\} \ .$$

Then μ is a measure on $(X, \mathcal{A})$.

There is also a version of the Dominated Convergence Theorem:

XVII.10.2.3. THEOREM. Let (ν_i) be a net of complex measures on a measurable space $(X, \mathcal{A})$, converging setwise to a function ν. If there is a finite measure μ such that $|\nu_i| \leq \mu$ for all i, then ν is a complex measure.

(Of course, this theorem applies in the case the ν_i are ordinary measures uniformly dominated by a finite measure μ.)

In fact, this result holds in greater generality. We make some definitions:

XVII.10.2.4. DEFINITION. Let $(\nu_i)_{i\in I}$ be a net of complex measures on a measurable space $(X, \mathcal{A})$, and μ a measure on $(X, \mathcal{A})$. Then (ν_i) is *uniformly absolutely continuous* with respect to μ if for every $\epsilon > 0$ there is a $\delta > 0$ and an $i_0 \in I$ such that, for all $A \in \mathcal{A}$ with $\mu(A) < \delta$,

$$|\nu_i(A)| < \epsilon \text{ for all } i \geq i_0 \ .$$

The term "uniformly absolutely continuous" is appropriate for this notion, in analogy with the characterization of absolute continuity of measures in ().

Under the hypotheses of XVII.10.2.5., (ν_i) is uniformly absolutely continuous with respect to μ (take $\delta = \epsilon$). Thus Theorem XVII.10.2.5. is a special case of the following result:

XVII.10.2.5. THEOREM. Let (ν_i) be a net of complex measures on a measurable space $(X, \mathcal{A})$, converging setwise to a function $\nu : \mathcal{A} \to \mathbb{C}$. If there is a finite measure μ such that (ν_i) is uniformly absolutely continuous with respect to μ, then ν is a complex measure.

We can even go to one step greater generality:

XVII.10.2.6. DEFINITION. Let $(\nu_i)_{i\in I}$ be a net of complex measures on a measurable space $(X, \mathcal{A})$. Then (ν_i) is *uniformly countably additive* if, whenever (A_n) is a sequence of pairwise disjoint sets in $\mathcal{A}$ with $A = \cup_{n=1}^{\infty} A_n$, and $\epsilon > 0$, there is an $N_0 \in \mathbb{N}$ and an $i_0 \in I$ such that

$$\left|\nu_i(A) - \nu_i\left(\cup_{n=1}^{N} A_n\right)\right| < \epsilon$$

for all $N \geq N_0$ and $i \geq i_0$.

XVII.10.2.7. PROPOSITION. Let $(\nu_i)_{i\in I}$ be a net of complex measures on a measurable space $(X, \mathcal{A})$. Then (ν_i) is uniformly countably additive if and only if, whenever (B_n) is a decreasing sequence in $\mathcal{A}$ with $\cap_{n=1}^{\infty} B_n = \emptyset$, we have $\lim_{n\to\infty} \nu_i(B_n) = 0$ uniformly in i, i.e. for every $\epsilon > 0$ there is an $n_0 \in \mathbb{N}$ and $i_0 \in I$ such that

$$|\nu_i(B_n)| < \epsilon$$

for all $n \geq n_0$ and $i \geq i_0$.

PROOF: If (ν_i) is uniformly countably additive and (B_n) is decreasing with empty intersection, set $A_n = B_n \setminus B_{n+1}$. Conversely, if (A_n) is a pairwise disjoint sequence, let $B_n = \cup_{k\geq n} A_k$.

XVII.10.2.8. Corollary. Let $(\nu_i)_{i\in I}$ be a net of complex measures on a measurable space $(X, \mathcal{A})$. If (ν_i) is uniformly countably additive, then whenever (A_n) is a pairwise disjoint sequence in $\mathcal{A}$ we have $\lim_{n\to\infty} \nu_i(A_n) = 0$ uniformly in i.

The converse is also true for a *sequence* of complex measures, and is a key step in the direct proof of the Nikodým Convergence Theorem (Exercise XVII.10.4.2.).
Proof: Set $B_n = \cup_{k=n}^{\infty} A_k$. Then $\nu_i(B_n) \to 0$ uniformly in i, and $|\nu_i(A_n)| \leq |\nu_i(B_n)| + |\nu_i(B_{n+1})|$ for all n and i. 👍

XVII.10.2.9. Proposition. Let (ν_i) be a net of complex measures on a measurable space $(X, \mathcal{A})$. If there is a finite measure μ such that (ν_i) is uniformly absolutely continuous with respect to μ, then (ν_i) is uniformly countably additive.

The converse is true for a setwise convergent sequence of complex measures (XVII.10.3.8.).
Proof: Use XVII.10.2.7.. Let $\epsilon > 0$, and fix δ_0 and i_0 as in Definition XVII.10.2.4.. Let (B_n) be a decreasing sequence in $\mathcal{A}$ with $\cap_{n=1}^{\infty} B_n = \emptyset$. Fix n_0 such that $\mu(B_n) < \delta$ for all $n \geq n_0$. Then, for all $n \geq n_0$ and $i \geq i_0$, $|\nu_i(B_n)| < \epsilon$. 👍

We then obtain the most general version of the Dominated Convergence Theorem for measures (note that there is no actual domination in this version!)

XVII.10.2.10. Theorem. Let (ν_i) be a net of complex measures on a measurable space $(X, \mathcal{A})$, converging setwise to a function $\nu : \mathcal{A} \to \mathbb{C}$. If (ν_i) is uniformly countably additive, then ν is a complex measure.

Proof: We again need only show countable additivity. Since ν is finite-valued, it suffices to show () that if (B_n) is a decreasing sequence in $\mathcal{A}$ with $\cap_{n=1}^{\infty} B_n = \emptyset$, we have $\lim_{n\to\infty} \nu(B_n) = 0$. Let $\epsilon > 0$, and fix a corresponding n_0 and i_0 as in the definition of uniform countable additivity (XVII.10.2.6.). Then, if $n \geq n_0$, then $|\nu_i(B_n)| < \epsilon$ for all $i \geq i_0$, and thus $|\nu(A_n)| \leq \epsilon$. 👍

XVII.10.3. The Nikodým and Vitali-Hahn-Saks Theorems

We now restrict attention to sequences of measures. There is a circle of results due to Nikodým, Vitali, Hahn, and Saks (there were other contributors too) which say that under certain mild restrictions, the setwise limit of a sequence of complex measures is a complex measure, and that the convergence is "almost uniform."

These results are closely interconnected, and can be largely deduced from each other; various schemes for proving them are found in the literature. Even the names of the results are not uniform in references; what is called the "Vitali-Hahn-Saks Theorem" in some references is called the "Nikodým Convergence Theorem" in others, reflecting the close connections. We have tried to use names accurately reflecting the actual work of the originators. ✋

The first observation is that a sequence of complex measures is dominated by a single measure:

XVII.10.3.1. PROPOSITION. Let (ν_n) be a sequence of complex measures on a measurable space $(X, \mathcal{A})$. Then there is a finite measure μ on $(X, \mathcal{A})$ such that $\nu_n \ll \mu$ for all n.

PROOF: We may assume all the ν_n are nonzero. For $A \in \mathcal{A}$, set

$$\mu(A) = \sum_{n=1}^{\infty} \frac{|\nu_n|(A)}{2^n |\nu_n|(X)} \ .$$

By XVII.10.2.2., μ is a measure. It is obvious that μ is finite and $\nu_n \ll \mu$ for all n. 👍

The next theorem is the main result of this section. VITALI proved the first version as a special case of his convergence theorem (), and HAHN obtained the result for Lebesgue-Stieltjes measures. The general result is due to SAKS. Our proof, also due to SAKS (cf. [DS88]), uses the Baire Category Theorem and is strongly reminiscent of the proof of the Uniform Boundedness Theorem ().

XVII.10.3.2. THEOREM. [VITALI-HAHN-SAKS] Let μ be a measure, and (ν_n) a sequence of complex measures, on a measurable space $(X, \mathcal{A})$. If (ν_n) converges setwise to a function $\nu : \mathcal{A} \to \mathbb{C}$, and $\nu_n \ll \mu$ for all n, then (ν_n) is uniformly absolutely continuous with respect to μ.

PROOF: Let $\Sigma(\mu)$ be the metric space of μ. Then each ν_n gives a continuous function from $\Sigma(\mu)$ to $\mathbb{C}$, converging pointwise to $\nu : \Sigma(\mu) \to \mathbb{C}$. Fix $\epsilon > 0$. For each m and n, set

$$E_{m,n} = \{A \in \mathcal{A} : |\nu_n(A) - \nu_m(A)| \leq \epsilon\}$$

(regarded as a subset of $\Sigma(\mu)$) and set $F_k = \bigcap_{m,n \geq k} E_{m,n}$. Then F_k is a closed subset of $\Sigma(\mu)$ for each k, and since (ν_n) converges pointwise, $\cup_{k=1}^{\infty} F_k = \Sigma(\mu)$. Thus, since $\Sigma(\mu)$ is complete (), F_k has a nonempty interior for some k by the Baire Category Theorem, i.e. there is a k, an $A \in \mathcal{A}$, and an $r > 0$ such that

$$\{B \in \mathcal{A} : \mu(A \triangle B) < r\} \subseteq F_k \ .$$

There is a δ, $0 < \delta < r$, such that $|\nu_n(B)| < \epsilon$ for $1 \leq n \leq k$ whenever $B \in \mathcal{A}$ and $\mu(B) < \delta$ (). Then, if $B \in \mathcal{A}$, $\mu(B) < \delta$, we have that $A \cup B$ and $A \setminus B$ are in F_k, so for $n \geq k$ we have

$$\nu_n(B) = \nu_k(B) + (\nu_n(B) - \nu_k(B))$$

$$= \nu_k(B) + (\nu_n(A \cup B) - \nu_k(A \cup B)) - (\nu_n(A \setminus B) - \nu_k(A \setminus B))$$

and thus $|\nu_n(B)| < 3\epsilon$ for all n. 👍

XVII.10.3.3. COROLLARY. [NIKODÝM CONVERGENCE THEOREM] Let (ν_n) be a sequence of complex measures on a measurable space $(X, \mathcal{A})$, converging setwise to a function $\nu : \mathcal{A} \to \mathbb{C}$. Then (ν_n) is uniformly countably additive, and ν is a complex measure.

PROOF: There is a finite measure μ on $\mathcal{A}$ with $\nu_n \ll \mu$ for all n by XVII.10.3.1.. Then (ν_n) is uniformly absolutely continuous with respect to μ by XVII.10.3.2., and hence is uniformly countably additive by XVII.10.2.9.. Then ν is a measure by XVII.10.2.10.. 👍

XVII.10.3.4. The Nikodým Convergence Theorem was proved before the general theorem XVII.10.3.2. was obtained. See [Swa94] for a direct proof of the Nikodým Convergence Theorem using a matrix version () of the "gliding hump" argument (cf. Exercise ()). The Vitali-Hahn-Saks Theorem can then be derived from the Nikodým Convergence Theorem (). There is a direct proof of the Nikodým Convergence Theorem for ordinary (finite) measures not using the Baire Category Theorem; see Exercise XVII.10.4.3..

There are versions of these results for finitely additive measures and for vector measures; see [DU77].

A nice consequence of the Nikodým Convergence Theorem is the following result, which is a close analog of the Uniform Boundedness Principle:

XVII.10.3.5. THEOREM. [NIKODÝM BOUNDEDNESS THEOREM] Let (ν_n) be a sequence of complex measures on a measurable space $(X, \mathcal{A})$. If $\{\nu_n(A) : n \in \mathbb{N}\}$ is bounded for each $A \in \mathcal{A}$, then the ν_n are uniformly bounded, i.e. there is an M such that $|\nu_n(A)| \leq M$ for all n and for all $A \in \mathcal{A}$.

For the proof, we need a lemma:

XVII.10.3.6. LEMMA. Let (ν_n) be a sequence of complex measures on a measurable space $(X, \mathcal{A})$. If $\{\nu_n(A) : n \in \mathbb{N}\}$ is bounded for each $A \in \mathcal{A}$, then (ν_n) is uniformly bounded if and only if, for every sequence (A_n) of pairwise disjoint sets in $\mathcal{A}$, the set $\{\nu_n(A_n)\}$ is bounded.

PROOF: The condition is obviously necessary. For sufficiency, suppose (ν_n) is not uniformly bounded. Let $N_0 = \sup_n\{|\nu_n(X)|\}$. Let $E_1 \in \mathcal{A}$ with $|\nu_{n_1}(E_1)| > 1 + N_0$ for some n_1; then, if $F_1 = X \setminus E_1$, $|\nu_{n_1}(F_1)| \geq |\nu_{n_1}(E_1)| - |\nu_{n_1}(X)| > 1$. Either $\{|\nu_n(A \cap E_1)| : A \in \mathcal{A}, n \in \mathbb{N}\}$ or $\{|\nu_n(A \cap F_1)| : A \in \mathcal{A}, n \in \mathbb{N}\}$ is unbounded; set $B_1 = E_1$ in the first case, and $B_1 = F_1$ in the second case, and let A_{n_1} be the other (i.e. $A_{n_1} = X \setminus B_1$). Let $N_1 = \sup_n\{|\nu_n(B_1)|\}$. Let $E_2 \in \mathcal{A}$, $E_2 \subseteq B_1$, with $|\nu_{n_2}(E_2)| > 2 + N_1$ for some n_2; we may assume $n_2 \neq n_1$ since $\{\nu_{n_1}(A) : A \in \mathcal{A}\}$ is bounded [by $|\nu_{n_1}|(X)$]. Then, if $F_2 = B_1 \setminus E_2$, $|\nu_{n_2}(F_2)| > 2$. Let $B_2 = E_2$ if $\{|\nu_n(A \cap E_2)| : A \in \mathcal{A}, n \in \mathbb{N}\}$ is unbounded and $B_2 = F_2$ if $\{|\nu_n(A \cap F_2)| : A \in \mathcal{A}, n \in \mathbb{N}\}$ is unbounded, and set $A_{n_2} = B_1 \setminus B_2$. Continue inductively to obtain a sequence (n_k) in $\mathbb{N}$ and a pairwise disjoint sequence (A_{n_k}) in $\mathcal{A}$ such that $|\nu_{n_k}(A_{n_k})| > k$. Set $A_n = \emptyset$ for n not equal to any n_k. 👍

XVII.10.3.7. **Remark:** The hypothesis that (ν_n) be setwise bounded is omitted from the statement of this result in [Swa94, p. 68]. This hypothesis is easily seen to be necessary: let X be a singleton and $\nu_n(X) = n$. Also, the proof only requires that $\mathcal{A}$ be an algebra and that the ν_n be finitely additive if the additional hypothesis is added that each ν_n is bounded, which is not automatic in this case (Exercise XVII.9.5.6.). (Without this hypothesis, the result fails: let ν_1 be a finitely additive signed measure which is finite but not bounded, and $\nu_n = 0$ for $n > 1$.)

We now prove XVII.10.3.5..

PROOF: By XVII.10.3.6. and (), it suffices to show that if (A_n) is a pairwise disjoint sequence in $\mathcal{A}$ and (α_n) is a sequence in $\mathbb{C}$ with $\alpha_n \to 0$, that $\alpha_n\nu_n(A_n) \to 0$. But $(\alpha_n\nu_n)$ is a sequence of complex measures on $(X, \mathcal{A})$ converging setwise to the zero measure; hence by XVII.10.3.3. the sequence $(\alpha_n\nu_n)$ is uniformly countably additive, giving the conclusion from XVII.10.2.8.. 👍

As a related matter, we have the following converse to XVII.10.2.9.:

XVII.10.3.8. THEOREM. Let (ν_n) be a sequence of complex measures on a measurable space $(X, \mathcal{A})$, and μ a measure on $(X, \mathcal{A})$ such that $\nu_n \ll \mu$ for all n. If (ν_n) is uniformly countably additive, then (ν_n) is uniformly absolutely continuous with respect to μ.

PROOF: If (ν_n) is not uniformly absolutely continuous with respect to μ, we will construct a pairwise disjoint sequence (A_k) in $\mathcal{A}$ and indices n_k such that $\nu_{n_k}(A_k)$ does not approach 0, contradicting XVII.10.2.8..

If (ν_n) is not uniformly absolutely continuous with respect to μ, there is an $\epsilon > 0$ such that for all $\delta > 0$, there is a $B \in \mathcal{A}$ with $\mu(B) < \delta$ and $|\nu_n(B)| \geq \epsilon$ for infinitely many n. Fix such an ϵ. Choose $B_1 \in \mathcal{A}$ and n_1 with $|\nu_{n_1}(B_1)| \geq \epsilon$ and $\mu(B_1) < 1$. Since $\nu_{n_1} \ll \mu$, there is a $\delta_1 > 0$ such that $|\nu_{n_1}(B)| < \epsilon/2$ whenever $\mu(B) < \delta_1$. There is a $B_2 \in \mathcal{A}$ and an $n_2 > n_1$ such that $|\nu_{n_2}(B_2)| \geq \epsilon$ and $\mu(B_2) < \delta_1/2$. Continuing inductively, construct sets $B_k \in \mathcal{A}$, numbers δ_k, and a strictly increasing sequence (n_k) of indices, such that, for all k, $\delta_{k+1} < \delta_k/2$, $|\nu_{n_k}(B_k)| \geq \epsilon$, $\mu(B_{k+1}) < \delta_k/2$, and $|\nu_{n_k}(B)| < \epsilon/2$ whenever $\mu(B) < \delta_k$.

Set $C_k = \cup_{n=k+1}^{\infty} B_n$ for each k. Then

$$\mu(C_k) \leq \sum_{n=k+1}^{\infty} \mu(B_n) < \sum_{n=1}^{\infty} \frac{\delta_k}{2^n} = \delta_k$$

so that $|\nu_{n_k}(B_k \cap C_k)| < \epsilon/2$ for each k. Thus, if $A_k = B_k \setminus C_k$, the A_k are pairwise disjoint and, for all k,

$$|\nu_{n_k}(A_k)| \geq |\nu_{n_k}(B_k)| - |\nu_{n_k}(B_k \cap C_k)| > \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2} \ .$$



XVII.10.4. Exercises

XVII.10.4.1. Let (ν_n) be a sequence of complex measures on a measurable space $(X, \mathcal{A})$ which is a Cauchy sequence with respect to the total variation norm (). Give a direct proof (not using the Nikodým Convergence Theorem) that (ν_n) converges setwise to a complex measure.

XVII.10.4.2. (a) Prove the converse of XVII.10.2.8.. [If (B_n) is a decreasing sequence with $\cap_{n=1}^{\infty} B_n = \emptyset$ but $|\nu_n(B_n)| > \delta$ for all n, for some δ, choose n_1 with $\nu_1(B_{n_1}) < \delta/2$, and for each k choose n_{k+1} with $\nu_{n_k}(B_{n_{k+1}}) < \delta/2$. If $A_{n_k} = B_{n_k} \setminus B_{n_{k+1}}$, the A_{n_k} are pairwise disjoint and $|\nu_{n_k}(A_{n_k})| > \delta/2$ for all k.]
(b) Use (a) and () to give a direct proof of XVII.10.3.3. (cf. [Swa94, p. 68]).

XVII.10.4.3. [Doo94, III.10] Give a direct proof that if (μ_n) is a sequence of finite measures on a measurable space $(X, \mathcal{A})$, converging setwise to a function $\mu : \mathcal{A} \to [0, \infty)$, then μ is a measure (i.e. countably additive), as follows:
(a) Suppose (B_n) is a decreasing sequence in $\mathcal{A}$ with $\cap_{n=1}^{\infty} B_n = \emptyset$ but

$$\lim_{n \to \infty} \mu(B_n) = \epsilon > 0$$

(why does the limit exist?) Let $n_1 = m_1 = 1$ and inductively choose $n_{k+1} > n_k$ and $m_{k+1} > m_k$ with $\mu_{n_{k+1}}(B_{m_k}) > \frac{7\epsilon}{8}$ and $\mu_{n_{k+1}}(B_{m_{k+1}}) < \frac{\epsilon}{8}$.

(b) Let $A_k = B_{m_k} \setminus B_{m_{k+1}}$ for each k. Show that if $j > k$ is odd, then

$$\mu_{n_j}(\cup\{A_n : n \text{ even}, n \geq k\}) > \frac{3\epsilon}{4}$$

and if $j > k$ is even, then

$$\mu_{n_j}(\cup\{A_n : n \text{ odd}, n \geq k\}) > \frac{3\epsilon}{4} \ .$$

(c) Conclude that $\mu(B_{m_k}) \geq \frac{3\epsilon}{2}$ for all k, a contradiction.

Chapter XVIII

Modern Integration

There are four general approaches to integration: the Riemann integral and the "modern" approaches of Lebesgue, Daniell, and Denjoy-Perron-Henstock-Kurzweil. In this chapter, we discuss the last three; Riemann integration is treated in Chapter IX.

We begin with Lebesgue integration. The name "Lebesgue integral" has two standard meanings in mathematics: the Lebesgue integral on $\mathbb{R}$, or $\mathbb{R}^n$, with respect to Lebesgue measure, and the "abstract Lebesgue integral" on a general measure space.

Many general measure theory texts follow what I call the "reinventing the wheel" plan: first develop Lebesgue measure and integration, then redo everything in general, in many cases just referring back to the Lebesgue case for proofs of results. This approach is certainly defensible pedagogically, and is preferred by many people, but it is not to my personal taste; I much prefer to start out with general measure and integration theory, spending considerable time discussing and emphasizing the critical special case of Lebesgue measure and integration on $\mathbb{R}$ (and then $\mathbb{R}^n$) as I go, as well as discussing other special cases such as probability spaces.

XVIII.1. The Lebesgue Integral

This section treats the heart of modern integration theory, the general integral first defined by LEBESGUE and subsequently refined by other authors.

The term "Lebesgue integral" is commonly used to mean two different things:

(i) Lebesgue's generalization of the Riemann integral for functions from $\mathbb{R}$ to $\mathbb{R}$.

(ii) The general integral for real-valued functions on a measure space, often called the *abstract Lebesgue integral*.

As the integral is normally done today, the Lebesgue integral on $\mathbb{R}$ is just the special case of the general abstract Lebesgue integral for the measure space $(\mathbb{R}, \mathcal{M}, \lambda)$, and it is no easier and even no different to develop this special case than the general integral. Thus we will work generally and point out the few important special properties of the Lebesgue integral on $\mathbb{R}$ as we go. We will usually suppress the name "Lebesgue" from the general abstract integral, except when this integral must be distinguished from integrals defined in other ways, and mostly reserve it for the special case of Lebesgue measure on $\mathbb{R}$ (or $\mathbb{R}^n$).

If $(X, \mathcal{A}, \mu)$ is a measure space, we want to assign an extended real number to any extended real-valued measurable function f, and $E \in \mathcal{A}$, called the *integral of f over E* and denoted $\int_E f\, d\mu$, in such a way that the integral has standard nice properties we would like and expect. It turns out not to be possible to do this for every extended real-valued measurable function in general (because of the obstacle of defining $\infty - \infty$), but only for nonnegative functions and general functions whose positive and negative parts are not "too large" ("integrable" functions). We will build up the integral in three steps (the process is called *bootstrapping*):

(i) Simple functions.

(ii) Nonnegative extended real-valued measurable functions.

(iii) Integrable functions.

XVIII.1.1. Properties of the Integral

We can give a list of properties we want the integral to have. This list might be regarded as a "wish list" at the moment, but we will be able to get almost everything on the list.

XVIII.1.1.1. Let $(X, \mathcal{A}, \mu)$ be a measure space. We would like the assignment $(f, E) \mapsto \int_E f\, d\mu$ to have the following properties:

(i) [MONOTONICITY] If $f \le g$, then $\int_E f\, d\mu \le \int_E g\, d\mu$ for any E.

(ii) [LINEARITY] $\int_E (f+g)\, d\mu = \int_E f\, d\mu + \int_E g\, d\mu$ for any f, g, E.

(iii) [HOMOGENEITY] $\int_E \alpha f\, d\mu = \alpha \int_E f\, d\mu$ for any $f, E, \alpha \in \mathbb{R}$ (where we always interpret $0 \cdot \infty = 0$).

(iv) [ADDITIVITY] If (E_k) are pairwise disjoint in $\mathcal{A}$ and $E = \cup_k E_k$, then $\int_E f\, d\mu = \sum_{k=1}^{\infty} \int_{E_k} f\, d\mu$.

(v) If $f = g$ a.e. (μ-a.e.), then $\int_E f\, d\mu = \int_E g\, d\mu$ for any E.

(vi) If $\mu(E) = 0$, then $\int_E f\,d\mu = 0$ for any f.

(vii) [CONTINUITY] If $f_n \to f$ pointwise, then $\int_E f_n\,d\mu \to \int_E f\,d\mu$ for any E.

(viii) [NORMALIZATION] $\int_E 1\,d\mu = \mu(E)$ for any E.

XVIII.1.1.2. Actually, (ii) and (iii) together are usually called "linearity," and can be replaced by the single condition

(ii′) $\int_E (\alpha f + g)\,d\mu = \alpha \int_E f\,d\mu + \int_E g\,d\mu$ for any f, g, E, and $\alpha \in \mathbb{R}$.

We have used "linearity" for (ii) instead of "additivity" to distinguish it from (iii), although the way we will proceed it will turn out that (iii), at least for finitely many sets, is a consequence of (ii).

XVIII.1.1.3. We will be able to get everything on this list except Continuity, although in several others we will have to make modest restrictions to avoid the $\infty - \infty$ problem. Continuity in the general form of (vii) is simply not possible in general (XVIII.1.7.1.); however, the major convergence theorems of integration theory (MCT, DCT) will show that continuity holds under some reasonable additional assumptions.

XVIII.1.2. Integral of Simple Functions

Recall the definition of a simple function (). To avoid verbosity, we will restrict the definition to only include functions which are nonnegative and measurable:

XVIII.1.2.1. DEFINITION. Let $(X, \mathcal{A})$ be a measurable space. A *simple function* on $(X, \mathcal{A})$ is a nonnegative real-valued measurable function which takes only finitely many values.

If ϕ is a simple function on $(X, \mathcal{A})$ taking exactly the values $\alpha_1, \ldots, \alpha_n$, for each k set

$$A_k = \{x \in X : \phi(x) = \alpha_k\}\ .$$

Then $\{A_1, \ldots, A_n\}$ is a partition of X into nonempty measurable subsets (), and

$$\phi = \sum_{k=1}^{n} \alpha_k \chi_{A_k}$$

is the *canonical representation* of ϕ.

In some references, a term with $\alpha_k = 0$ is deleted from the canonical representation, but we prefer to keep such a term to force $\{A_1, \ldots, A_n\}$ to be a partition of X. This will have no practical effect.

XVIII.1.2.2. Any nonnegative (finite) linear combination of indicator functions of measurable sets is a simple function. A sum

$$\phi = \sum_{j=1}^{m} \beta_j \chi_{B_j}$$

may not be a canonical representation of ϕ for four reasons:

(1) The B_j may not be disjoint.

(2) The β_j may not all be distinct.

(3) We may not have $\cup_j B_j = X$.

(4) Some of the B_j may be empty.

(3) and (4) are easy to remedy: for (3), set $B_{m+1} = X \setminus \cup_{j=1}^m B_j$ and $\beta_{m+1} = 0$, and for (4), just delete any terms with $B_j = \emptyset$. The others can also be systematically fixed: for (1), let $A_1, \ldots, A_n$ consist of all nonempty finite intersections of the B_j and their complements (i.e. the nonempty sets in the partition of $\cup_{j=1}^m B_j$ generated by the B_j), and if $A_k = B_{j_1} \cap \cdots \cap B_{j_r} \cap B^c_{j_{r+1}} \cap \cdots B^c_{j_s}$, set $\alpha_k = \beta_{j_1} + \cdots + \beta_{j_r}$. Then for (2) just combine together terms with equal β_j, using that $\chi_{A \cup B} = \chi_A + \chi_B$ if A and B are disjoint. If these corrections are done in the order (1), (3), (4), (2), each only needs to be done once and the representation is converted into canonical form.

A representation in which all the B_j are disjoint, but (2), (3), (4) are allowed, is said to be in *semicanonical form*.

XVIII.1.2.3. Note that if ϕ is simple and $E \in \mathcal{A}$, then $\phi\chi_E$ is also simple: if $\phi = \sum_{j=1}^m \beta_j \chi_{B_j}$, then $\phi\chi_E = \sum_{j=1}^m \beta_j \chi_{B_j \cap E}$. If the representation of ϕ is in semicanonical form, so is this representation of $\phi\chi_E$; but the representation of $\phi\chi_E$ is not necessarily in canonical form even if the representation of ϕ is.

The Integral of a Simple Function

We can define the integral of a simple function in the only way possible if we hope to have linearity and normalization:

XVIII.1.2.4. Definition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and ϕ a simple function on $(X, \mathcal{A})$. Write $\phi = \sum_{k=1}^n \alpha_k \chi_{A_k}$ in canonical form. Then

$$\int_X \phi \, d\mu = \sum_{k=1}^n \alpha_k \mu(A_k)$$

where $0 \cdot \infty$ is always interpreted to be 0, and $r \cdot \infty = \infty$ if $r > 0$. If $E \in \mathcal{A}$, set

$$\int_E \phi \, d\mu = \int_X \phi\chi_E \, d\mu \ .$$

A few simple properties are immediate from the definition:

XVIII.1.2.5. Proposition. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $E \in \mathcal{A}$.
(i) If ϕ is a simple function on $(X, \mathcal{A})$, and $r \geq 0$, then

$$\int_E r\phi \, d\mu = r \int_E \phi \, d\mu \ .$$

(ii) If ϕ and ψ are simple functions and $\phi = \psi$ a.e. on E (i.e. $\phi\chi_E = \psi\chi_E$ a.e.), then

$$\int_E \phi\, d\mu = \int_E \psi\, d\mu \ .$$

(iii) If $\mu(E) = 0$, then $\int_E \phi\, d\mu = 0$ for any simple function ϕ.

PROOF: (i): If $r > 0$ and $\phi\chi_E = \sum_{k=1}^n \alpha_k \chi_{A_k}$ is the canonical representation, then the canonical representation of $r\phi\chi_E$ is $\sum_{k=1}^n \alpha_k \chi_{A_k}$. And if $r = 0$, both sides are 0 (recall the convention that $0 \cdot \infty = 0$).
(ii): If $\phi\chi_E = \sum_{k=1}^n \alpha_k \chi_{A_k}$ and $\psi\chi_E = \sum_{j=1}^m \beta_j \chi_{B_j}$ are the canonical representations, then each α_k corresponds to a unique β_{j_k} unless $\mu(A_k) = 0$ in which case the term $\alpha_k \mu(A_k)$ can be deleted from the sum, and conversely; thus once zero terms are deleted the terms in the sums are in one-one correspondence. For each k the sets A_k and B_{j_k} differ by a null set, so $\mu(A_k) = \mu(B_{j_k})$ (). Thus the remaining terms are numerically identical.
(iii): If $\mu(E) = 0$, then in the canonical representation $\phi\chi_E = \sum_{k=1}^n \alpha_k \chi_{A_k}$, in any term with $\alpha_k > 0$ we have $A_k \subseteq E$, so $\mu(A_k) = 0$. Thus all the terms in $\sum_{k=1}^n \alpha_k \mu(A_k)$ are zero. 👍

The key technical fact giving reasonable properties of the integral defined this way is that the formula in the definition holds for any representation of ϕ, not just the canonical one:

XVIII.1.2.6. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and ϕ a simple function on $(X, \mathcal{A})$. If $\phi = \sum_{j=1}^m \beta_j \chi_{B_j}$ is any representation of ϕ, then

$$\int_X \phi\, d\mu = \sum_{j=1}^m \beta_j \mu(B_j) \ .$$

PROOF: We must observe that the sum does not change as the four corrections of XVIII.1.2.2. are made to convert the given representation of ϕ to the canonical one. Corrections (3) and (4) obviously do not affect the sum (using the convention $0 \cdot \infty = 0$ for (3)), and correction (2) is easily seen to also leave the sum unchanged if the B_j are disjoint (using finite additivity of μ).

Correction (1) requires a little more work. We disjointize B_1 and B_2 into $B_1 \cap B_2$, $B_1 \setminus B_2$, and $B_2 \setminus B_1$. Observe that

$$\beta_1\chi_{B_1} + \beta_2\chi_{B_2} = \beta_1\chi_{B_1\setminus B_2} + \beta_2\chi_{B_2\setminus B_1} + (\beta_1 + \beta_2)\chi_{B_1\cap B_2}$$

and that

$$\beta_1\mu(B_1 \setminus B_2) + \beta_2\mu(B_2 \setminus B_1) + (\beta_1 + \beta_2)\mu(B_1 \cap B_2)$$
$$= \beta_1[\mu(B_1 \setminus B_2) + \mu(B_1 \cap B_2)] + \beta_2[\mu(B_1 \setminus B_2) + \mu(B_1 \cap B_2)] = \beta_1\mu(B_1) + \beta_2\mu(B_2)$$

by finite additivity of μ; thus the sum for the beginning representation equals the sum obtained by disjointizing B_1 and B_2. We can then successively disjointize any two nondisjoint remaining sets in the new representation in the same way without affecting the sum or the previous disjointizations. After a finite number of steps the representation is converted to semicanonical form and the sum is unchanged. 👍

Most of the important properties of the integral for simple functions follow easily. First is a formula for the integral of a simple function over a subset:

XVIII.1.2.7. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\phi = \sum_{k=1}^{n} \alpha_k \chi_{A_k}$ a simple function on $(X, \mathcal{A})$ (not necessarily in canonical form). If $E \in \mathcal{A}$, then

$$\int_E \phi \, d\mu = \int_X \phi\chi_E \, d\mu = \sum_{k=1}^{n} \alpha_k \mu(A_k \cap E) \ .$$

Linearity and (finite) additivity also follow easily:

XVIII.1.2.8. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and ϕ and ψ simple functions on $(X, \mathcal{A})$. If $E \in \mathcal{A}$, then

$$\int_E (\phi + \psi) \, d\mu = \int_E \phi \, d\mu + \int_E \psi \, d\mu \ .$$

Proof: If $\phi\chi_E = \sum_{k=1}^{n} \alpha_k \chi_{A_k}$ and $\psi\chi_E = \sum_{j=1}^{m} \beta_j \chi_{B_j}$ are representations (not necessarily canonical), then

$$(\phi + \psi)\chi_E = \sum_{k=1}^{n} \alpha_k \chi_{A_k} + \sum_{j=1}^{m} \beta_j \chi_{B_j}$$

is a representation. 👍

Note that the representation of $(\phi + \psi)\chi_E$ in the proof is almost necessarily non-canonical, so the proposition is necessary to establish linearity of the integral for simple functions.

XVIII.1.2.9. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and ϕ a simple function on $(X, \mathcal{A})$. If $D_1, \ldots, D_n \in \mathcal{A}$ are pairwise disjoint and $E = D_1 \cup \cdots \cup D_n$, then

$$\int_E \phi \, d\mu = \sum_{k=1}^{n} \left[\int_{D_k} \phi \, d\mu \right] \ .$$

Proof: We have $\phi\chi_E = \phi\chi_{D_1} + \cdots + \phi\chi_{D_n}$. 👍

Finite additivity can be extended to countable additivity (XVIII.1.2.12.).

We can also obtain monotonicity:

XVIII.1.2.10. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a measure space, and ϕ and ψ simple functions on $(X, \mathcal{A})$. If $E \in \mathcal{A}$ and $\phi \leq \psi$ a.e. on E, then

$$\int_E \phi \, d\mu \leq \int_E \psi \, d\mu \ .$$

PROOF: By modifying ϕ on a null set (cf. XVIII.1.2.5.(ii)), we may assume $\phi\chi_E \leq \psi\chi_E$ everywhere. Let $\phi\chi_E = \sum_{k=1}^n \alpha_k \chi_{A_k}$ and $\psi\chi_E = \sum_{j=1}^m \beta_j \chi_{B_j}$ be canonical representations. Then

$$\phi\chi_E = \sum_{j=1}^m \sum_{k=1}^n \alpha_k \chi_{A_k \cap B_j}$$

$$\psi\chi_E = \sum_{j=1}^m \sum_{k=1}^n \beta_j \chi_{A_k \cap B_j}$$

are semicanonical representations; and whenever $A_k \cap B_j \neq \emptyset$, we have $\alpha_k \leq \beta_j$ since $\phi\chi_E \leq \psi\chi_E$. Thus

$$\int_E \phi \, d\mu = \sum_{j=1}^m \sum_{k=1}^n \alpha_k \mu(A_k \cap B_j) \leq \sum_{j=1}^m \sum_{k=1}^n \beta_j \mu(A_k \cap B_j) = \int_E \psi \, d\mu \ .$$



We also have the following slight extension of the Baby Monotone Convergence Theorem (), which will be used to establish the full Monotone Convergence Theorem and other continuity properties of the integral:

XVIII.1.2.11. PROPOSITION. [JUNIOR MONOTONE CONVERGENCE THEOREM] Let $(X, \mathcal{A}, \mu)$ be a measure space, and ϕ a simple function on $(X, \mathcal{A})$. If $E_1, E_2, \ldots$ is an increasing sequence of sets in $\mathcal{A}$, and $E = \cup_{n=1}^\infty E_n$, then

$$\int_E \phi \, d\mu = \lim_{n\to\infty} \left[\int_{E_n} \phi \, d\mu \right] = \sup_n \left[\int_{E_n} \phi \, d\mu \right] \ .$$

PROOF: Let $\phi = \sum_{k=1}^m \alpha_k \chi(A_k)$ be a representation (not necessarily canonical). Fix k. Then we have $A_k \cap E_1 \subseteq A_k \cap E_2 \subseteq \cdots$ and $A_k \cap E = \cup_{n=1}^\infty [A_k \cap E_n]$, so by the Baby MCT we have

$$\alpha_k \mu(A_k \cap E) = \lim_{n\to\infty} \alpha_k \mu(A_k \cap E_n) = \sup_n \alpha_k \mu(A_k \cap E_n) \ .$$

This is true for all k, and since the sum is finite, we have

$$\sum_{k=1}^m \alpha_k \mu(A_k \cap E) = \lim_{n\to\infty} \left[\sum_{k=1}^m \alpha_k \mu(A_k \cap E_n) \right] = \sup_n \left[\sum_{k=1}^m \alpha_k \mu(A_k \cap E_n) \right] \ .$$



XVIII.1.2.12. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a measure space, and ϕ a simple function on $(X, \mathcal{A})$. If $D_1, \ldots, D_n, \ldots \in \mathcal{A}$ are pairwise disjoint and $E = \cup_{k=1}^{\infty} D_k$, then

$$\int_E \phi \, d\mu = \sum_{k=1}^{\infty} \left[\int_{D_k} \phi \, d\mu \right] .$$

PROOF: For each n set $E_n = D_1 \cup \cdots \cup D_n$. Then the E_n are increasing and $E = \cup_n E_n$, so by XVIII.1.2.11. and XVIII.1.2.9. we have

$$\int_E \phi \, d\mu = \lim_{n\to\infty} \left(\int_{E_n} \phi \, d\mu \right) = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \left[\int_{D_k} \phi \, d\mu \right] \right) = \sum_{k=1}^{\infty} \left[\int_{D_k} \phi \, d\mu \right] .$$



XVIII.1.3. The Integral of a Nonnegative Measurable Function

We can now define the integral of a general nonnegative measurable function:

XVIII.1.3.1. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. Define

$$\int_X f \, d\mu = \sup \left\{ \int_X \phi \, d\mu : \phi \text{ simple}, \, \phi \le f \right\} .$$

If $E \in \mathcal{A}$, define

$$\int_E f \, d\mu = \int_X f \chi_E \, d\mu .$$

Note that if f itself is a simple function, then this definition of $\int_E f \, d\mu$ agrees with the previous definition (cf. XVIII.1.2.10.); thus the definitions are consistent and unambiguous.

XVIII.1.3.2. The following properties of the integral are immediate:

(i) $0 \le \int_E f \, d\mu \le \infty$ for any f and E.

(ii) If $f \le g$ on E, then $\int_E f \, d\mu \le \int_E g \, d\mu$.

(iii) If $\alpha \ge 0$, then $\int_E \alpha f \, d\mu = \alpha \int_E f \, d\mu$ (using $0 \cdot \infty = 0$ and $\alpha \cdot \infty = \infty$ if $\alpha > 0$). In particular, $\int_E 0 \, d\mu = 0$ for any E.

(iv) If $\mu(E) = 0$, then $\int_E f \, d\mu = 0$ for any f (cf. XVIII.1.2.5.).

XVIII.1.3.3. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f and g nonnegative extended real-valued measurable functions on $(X, \mathcal{A})$. If $f = g$ a.e. (μ-a.e.) on E, then

$$\int_E f \, d\mu = \int_E g \, d\mu \ .$$

PROOF: If $f\chi_E = g\chi_E$ on $X \setminus C$, where $C \in \mathcal{A}$ and $\mu(C) = 0$, and ϕ is a simple function with $\phi \leq f\chi_E$, then $\psi = \phi\chi_{X\setminus C}$ is a simple function with $\psi \leq g\chi_E$, and $\psi = \phi$ a.e., so $\int_X \psi \, d\mu = \int_X \phi \, d\mu$ by XVIII.1.2.5.. Thus $\int_E f \, d\mu \leq \int_E g \, d\mu$. Symmetrically, $\int_E g \, d\mu \leq \int_E f \, d\mu$. 👍

Combining this with XVIII.1.3.2.(ii), we obtain:

XVIII.1.3.4. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f and g nonnegative extended real-valued measurable functions on $(X, \mathcal{A})$. If $f \leq g$ a.e. (μ-a.e.) on E, then

$$\int_E f \, d\mu \leq \int_E g \, d\mu \ .$$

XVIII.1.3.5. We have thus established all the desired properties of the integral (for nonnegative functions) from XVIII.1.1.1. except for the crucial properties of linearity, additivity, and continuity. These are trickier to show; we will first prove one of the important continuity properties, the Monotone Convergence Theorem, and use it to obtain linearity and additivity (and much more).

Integrable Functions

The integral of a nonnegative extended real-valued measurable function can be ∞ (even the integral of a simple function can be infinite). But functions with finite integral are the most interesting ones:

XVIII.1.3.6. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. Then f is *μ-integrable* if

$$\int_X f \, d\mu < \infty \ .$$

If $E \in \mathcal{A}$, then f is μ-integrable on E if $\int_E f \, d\mu < \infty$.

XVIII.1.3.7. Note that a μ-integrable function is by definition measurable. We often write "integrable" in place of "μ-integrable" when the μ is understood. But note that, although measurability of a function f depends only on $\mathcal{A}$ and not on μ, integrability of f depends on μ.

When μ is Lebesgue measure λ on $\mathbb{R}$ (or a subset), we will say f is *Lebesgue integrable* if it is λ-integrable. We usually, but not always, write "Lebesgue integrable" in this context to distinguish the notion from Riemann integrability (or other types of integrability on $\mathbb{R}$ such as H-K integrability).

XVIII.1.3.8. A nonnegative measurable function dominated by an integrable function is integrable. A nonnegative measurable function equal almost everywhere to an integrable function is integrable (XVIII.1.3.3.).

If μ is finite, any bounded nonnegative measurable function on $(X, \mathcal{A})$ is integrable; but if μ is infinite, even a (nonzero) constant function is not integrable.

An integrable function can be unbounded. It can even take the value ∞ on a null set, e.g. $f = \infty \cdot \chi_A$, where $\mu(A) = 0$. But an integrable function must be finite a.e., hence equal a.e. to a real-valued measurable function:

XVIII.1.3.9. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. Set $A = \{x \in X : f(x) = \infty\}$. If f is integrable, then $\mu(A) = 0$.

PROOF: If $\phi_n = n\chi_A$, then ϕ_n is a simple function and $\phi_n \leq f$. Thus

$$n\mu(A) = \int_X \phi_n \, d\mu \leq \int_X f \, d\mu$$

for all n, which implies $\mu(A) = 0$. 👍

Similarly, if f is integrable, the support () of f has "almost finite measure" (it need not actually have finite measure; cf. ()):

XVIII.1.3.10. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. Fix $\epsilon > 0$, and set $A_\epsilon = \{x \in X : f(x) \geq \epsilon\}$. Then

$$\mu(A_\epsilon) \leq \frac{1}{\epsilon} \int_X f \, d\mu \ .$$

In particular, if f is integrable, then $\mu(A_\epsilon) < \infty$ for all $\epsilon > 0$.

PROOF: Set $\phi_\epsilon = \epsilon\chi_{A_\epsilon}$. Then ϕ_ϵ is a simple function and $\phi_\epsilon \leq f$. Thus

$$\epsilon\mu(A_\epsilon) = \int_X \phi_\epsilon \, d\mu \leq \int_X f \, d\mu \ .$$

👍

XVIII.1.3.11. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. If $E \in \mathcal{A}$ and $\int_E f \, d\mu = 0$, then $f = 0$ a.e. on E.

PROOF: If $A = \{x \in E : f(x) > 0\}$, then $A = \cup_{n=1}^\infty A_n$, where $A_n = \left\{x \in E : f(x) \geq \frac{1}{n}\right\}$, and

$$\mu(A_n) \leq n \int_E f \, d\mu = 0 \ .$$

👍

XVIII.1.3.12. Proposition XVIII.1.3.10. is not only interesting for ϵ small; it is of interest for arbitrary positive ϵ. In fact, as $\epsilon \to \infty$ we obtain XVIII.1.3.9.. The proposition says that an integrable function is small except on a set of reasonable finite measure, and not too large except on a set of small measure.

XVIII.1.3.13. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative integrable function on $(X, \mathcal{A})$. Then the support of f is a set of σ-finite measure.

Proof: If $E = \{x \in X : f(x) > 0\}$ is the support of f, then $E = \cup_{n=1}^{\infty} A_{1/n}$. 👍

XVIII.1.3.14. Thus, although integration is defined on arbitrary measure spaces, it is only a nontrivial theory if there are sufficiently many sets of finite measure (e.g. if the measure is semifinite). If there are no sets of nonzero finite measure, then the only nonnegative integrable functions are ones equal to 0 almost everywhere.

Integration for Counting Measure

By far the most important example of the integral we have constructed is the Lebesgue integral on $\mathbb{R}$, discussed in Section XVIII.1.6.. But another illuminating example is counting measure on $\mathbb{N}$.

XVIII.1.3.15. Consider the measure space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$, where μ is counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. A real-valued function on $\mathbb{N}$ is just a sequence, and every such function is measurable since the σ-algebra is $\mathcal{P}(\mathbb{N})$. An extended real-valued function is one where some terms of the sequence are allowed to be ∞. A simple function is a sequence of nonnegative real numbers where only finitely many distinct numbers occur; such a function is integrable if and only if only finitely many terms are nonzero, and the integral is just the sum of the terms.

Thus if $x = (x_1, x_2, \dots)$ is a nonnegative sequence, then $\int_{\mathbb{N}} x \, d\mu$ is just the unordered sum

$$\sum_{n \in \mathbb{N}} x_n$$

which converges if and only if all terms are finite and the infinite series

$$\sum_{n=1}^{\infty} x_n$$

converges ().

We will see () that a general function (sequence) (x_n) is integrable with respect to μ if and only if the infinite series

$$\sum_{n=1}^{\infty} x_n$$

converges absolutely. Thus the theory of absolutely convergent infinite series is the special case of the integral for counting measure. In fact, it is worthwhile to view integration in general as some sort of "continuous" version of absolute summation.

XVIII.1.4. The Monotone Convergence Theorem

The Monotone Convergence Theorem (MCT) is one of the three main convergence or continuity theorems of integration theory, the others being Fatou's Lemma and the Dominated Convergence Theorem.

XVIII.1.4.1. THEOREM. [MONOTONE CONVERGENCE THEOREM] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (f_n) a nondecreasing sequence of nonnegative extended real-valued measurable functions converging pointwise a.e. to a (nonnegative) extended real-valued measurable function f. If $E \in \mathcal{A}$, then

$$\int_E f\,d\mu = \lim_{n\to\infty} \int_E f_n\,d\mu = \sup_n \int_E f_n\,d\mu\ .$$

Note that it is the *sequence* (f_n) which is nondecreasing, i.e. $f_1 \le f_2 \le \cdots$, not the individual functions which are nondecreasing. In fact, since the f_n are real-valued functions on a general set X, it does not even make sense to ask whether such a function is nondecreasing. Note also that the hypothesis that f is measurable is almost, but not quite, automatic (if the measure space is not complete, f might have to be modified on a subset of a null set to make it measurable).

PROOF: By modifying f on a null set, we may assume that $f_n \to f$ pointwise everywhere. Since $f_n \le f$ for all n, we have $\int_E f_n\,d\mu \le \int_E f\,d\mu$ for all n. Thus if

$$L := \lim_{n\to\infty} \int_E f_n\,d\mu = \sup_n \int_E f_n\,d\mu$$

(the limit exists since the sequence of integrals is nondecreasing), we have $L \le \int_E f\,d\mu$.

So we need to show $\int_E f\,d\mu \le L$. If $L = \infty$ there is nothing to show, so we may assume $L < \infty$.

Fix r, $0 < r < 1$. Let ϕ be a simple function with $\phi \le f\chi_E$, and set

$$E_n = \{x \in E : r\phi(x) \le f_n(x)\}\ .$$

Then E_n is measurable [$E_n = E \cap \{x \in X : f_n(x) - r\phi(x) \ge 0\}$ and $f_n - r\phi$ is a measurable function], and $E_n \subseteq E_{n+1}$ for all n since $f_n \le f_{n+1}$. If $x \in E$ and $f(x) > 0$, then $f_n(x) \to f(x) > r\phi(x)$, so there is an n such that $f_k(x) > r\phi(x)$ for all $k \ge n$; thus $x \in E_n$. And if $f(x) = 0$, then $\phi(x) = f_n(x) = 0$ for all n, so $x \in E_n$ for all n. Thus $\cup_n E_n = E$. And, for each n, $r\phi\chi_{E_n} \le f_n\chi_{E_n}$, so $\int_{E_n} r\phi\,d\mu \le \int_{E_n} f_n\,d\mu \le \int_E f_n\,d\mu \le L$ by monotonicity.

By the Junior Monotone Convergence Theorem (XVIII.1.2.11.),

$$\int_E r\phi\,d\mu = \sup_n \int_{E_n} r\phi\,d\mu \le \sup_n \int_{E_n} f_n\,d\mu \le L\ .$$

Thus

$$\int_E \phi\,d\mu \le \frac{1}{r} L\ .$$

This is true for all simple $\phi \le f\chi_E$, so we have

$$\int_E f\,d\mu = \sup_{\phi \le f\chi_E} \int_E \phi\,d\mu \le \frac{1}{r} L\ .$$

Since this is true for all $r < 1$, we conclude $\int_E f\,d\mu \le L$.

XVIII.1.4.2. In some references, the MCT or its series version XVIII.1.5.5. is called the *Beppo Levi Theorem*. While there is some historical justification for this, and we do like BEPPO'S name (perhaps the main reason mathematicians use this terminology?), we prefer the more descriptive MCT term.

XVIII.1.5. Consequences of the Monotone Convergence Theorem

The Monotone Convergence Theorem is the key to several more important properties of the integral.

Linearity

XVIII.1.5.1. THEOREM. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f and g nonnegative extended real-valued measurable functions functions on $(X, \mathcal{A})$. If $E \in \mathcal{A}$, then

$$\int_E (f+g)\,d\mu = \int_E f\,d\mu + \int_E g\,d\mu\ .$$

PROOF: Let (ϕ_n) and (ψ_n) be nondecreasing sequences of simple functions converging pointwise to $f\chi_E$ and $g\chi_E$ respectively (). Then $(\phi_n + \psi_n)$ is an increasing sequence of simple functions converging pointwise to $(f+g)\chi_E$, and by XVIII.1.2.8. we have

$$\int_E (\phi_n + \psi_n)\,d\mu = \int_E \phi_n\,d\mu + \int_E \psi_n\,d\mu$$

for all n. Thus by three applications of the MCT, along with (),

$$\begin{aligned}\int_E (f+g)\,d\mu &= \lim_{n\to\infty} \int_E (\phi_n + \psi_n)\,d\mu = \lim_{n\to\infty} \left[\int_E \phi_n\,d\mu + \int_E \psi_n\,d\mu\right] \\ &= \left[\lim_{n\to\infty} \int_E \phi_n\,d\mu\right] + \left[\lim_{n\to\infty} \int_E \psi_n\,d\mu\right] = \int_E f\,d\mu + \int_E g\,d\mu\ .\end{aligned}$$



Additivity

We next establish finite additivity:

XVIII.1.5.2. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. If $D_1, \ldots, D_n \in \mathcal{A}$ are pairwise disjoint and $E = D_1 \cup \cdots \cup D_n$, then

$$\int_E f\,d\mu = \sum_{k=1}^{n} \left[\int_{D_k} f\,d\mu\right]\ .$$

PROOF: We have $f\chi_E = f\chi_{D_1} + \cdots + f\chi_{D_n}$.



Before extending to countable additivity, we note the following special case of the MCT which extends the Junior MCT:

XVIII.1.5.3. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. If $E_1, E_2, \ldots$ is an increasing sequence of sets in $\mathcal{A}$, and $E = \cup_{n=1}^{\infty} E_n$, then

$$\int_E f\, d\mu = \lim_{n\to\infty} \left[\int_{E_n} f\, d\mu\right] = \sup_n \left[\int_{E_n} f\, d\mu\right] .$$

Proof: The sequence $(f\chi_{E_n})$ is nondecreasing and converges pointwise to $f\chi_E$. 👍

We then obtain countable additivity:

XVIII.1.5.4. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. If $D_1, \ldots, D_n, \ldots \in \mathcal{A}$ are pairwise disjoint and $E = \cup_{k=1}^{\infty} D_k$, then

$$\int_E f\, d\mu = \sum_{k=1}^{\infty} \left[\int_{D_k} f\, d\mu\right] .$$

Proof: For each n set $E_n = D_1 \cup \cdots \cup D_n$. Then the E_n are increasing and $E = \cup_n E_n$, so by XVIII.1.5.3. and XVIII.1.5.2. we have

$$\int_E f\, d\mu = \lim_{n\to\infty} \left(\int_{E_n} f\, d\mu\right) = \lim_{n\to\infty} \left(\sum_{k=1}^{n} \left[\int_{D_k} f\, d\mu\right]\right) = \sum_{k=1}^{\infty} \left[\int_{D_k} f\, d\mu\right] .$$

👍

The Monotone Convergence Theorem for Series

More generally, we have the following series version of the MCT:

XVIII.1.5.5. Theorem. [Monotone Convergence Theorem for Series] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (g_k) a sequence of nonnegative extended real-valued measurable functions on $(X, \mathcal{A})$. Set

$$f = \sum_{k=1}^{\infty} g_k$$

(the series converges pointwise to an extended real-valued function since the terms are nonnegative). If $E \in \mathcal{A}$, then

$$\int_E f\, d\mu = \sum_{k=1}^{\infty} \left[\int_E g_k\, d\mu\right] .$$

Proof: For each n set

$$f_n = \sum_{k=1}^{n} g_k .$$

Then (f_n) is a nondecreasing sequence of nonnegative extended real-valued measurable functions converging pointwise to f. By the MCT and XVIII.1.5.1.,

$$\int_E f\,d\mu = \lim_{n\to\infty}\int_E f_n\,d\mu = \lim_{n\to\infty}\left[\sum_{k=1}^{n}\int_E g_k\,d\mu\right] = \sum_{k=1}^{\infty}\left[\int_E g_k\,d\mu\right]\ .$$

The Downward MCT

If (f_n) is a nonincreasing sequence of nonnegative real-valued measurable functions on a measure space $(X,\mathcal{A},\mu)$, it is not necessarily true that $\int_X f_n\,d\mu \to \int_X f\,d\mu$, even if the f_n are indicator functions of measurable sets (XVIII.1.7.1.). However, if the f_n are integrable, downward continuity holds:

XVIII.1.5.6. Theorem. [Downward Monotone Convergence Theorem] Let $(X,\mathcal{A},\mu)$ be a measure space, and (f_n) a nonincreasing sequence of nonnegative extended real-valued measurable functions converging pointwise a.e. to a (nonnegative) extended real-valued measurable function f. If $E\in\mathcal{A}$, and $\int_E f_1\,d\mu < \infty$, then

$$\int_E f\,d\mu = \lim_{n\to\infty}\int_E f_n\,d\mu = \inf_n \int_E f_n\,d\mu\ .$$

Proof: If $A=\{x\in X : f_1(x)=\infty\}$, then $\mu(A\cap E)=0$ by XVIII.1.3.9.. Thus we may replace f_n by $f_n\chi_{A^c}$ for each n without affecting the integrals over E, i.e. we may assume the f_n are real-valued.

For each n, set $g_n = f_1 - f_n$ (this is well defined since the functions are real-valued). Then (g_n) is a nondecreasing sequence of nonnegative real-valued measurable functions converging pointwise a.e. to $f_1 - f$; thus by the MCT

$$\int_E (f_1 - f)\,d\mu = \lim_{n\to\infty}\int_E g_n\,d\mu = \lim_{n\to\infty}\int_E (f_1 - f_n)\,d\mu\ .$$

For each n we have

$$\int_E f_1\,d\mu = \int_E f_n\,d\mu + \int_E (f_1 - f_n)\,d\mu$$

and, since the integrals are finite,

$$\int_E (f_1 - f_n)\,d\mu = \int_E f_1\,d\mu - \int_E f_n\,d\mu\ .$$

Similarly,

$$\int_E (f_1 - f)\,d\mu = \int_E f_1\,d\mu - \int_E f\,d\mu\ .$$

Thus we have

$$\int_E f_1\,d\mu - \int_E f\,d\mu = \lim_{n\to\infty}\left[\int_E f_1\,d\mu - \int_E f_n\,d\mu\right] = \int_E f_1\,d\mu - \lim_{n\to\infty}\int_E f_n\,d\mu$$

and the result follows.

XVIII.1.5.7. The requirement that $\int_E f_1\,d\mu < \infty$ may be relaxed to $\int_E f_n\,d\mu < \infty$ for some n by passing to a subsequence. The Downward MCT is also an immediate corollary of the Dominated Convergence Theorem.

Approximation From Above

We now discuss when the integral of a nonnegative real-valued function can be approximated from above by integrals of simple functions. Fix a measure space $(X, \mathcal{A}, \mu)$.

XVIII.1.5.8. An unbounded nonnegative real-valued function cannot be dominated by a simple function, since simple functions are bounded. And if μ is infinite, an integrable f whose support has infinite measure cannot be dominated by an integrable simple function, since the support of an integrable simple function has finite measure. Thus the only functions for which there is a hope of approximating the integral from above by integrals of simple functions are bounded (nonnegative) functions whose support has finite measure.

XVIII.1.5.9. THEOREM. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a nonnegative extended real-valued measurable function on $(X, \mathcal{A})$. Let S be the support of f. If $E \in \mathcal{A}$ and $\mu(E \cap S) < \infty$, and f is bounded on E, then

$$\int_E f\,d\mu = \inf\left\{\int_E \psi\,d\mu : \psi \text{ simple, } f\chi_E \le \psi\right\}\ .$$

PROOF: Suppose $f \le M$ on E. Then $f\chi_E \le \theta := M\chi_{E\cap S}$, θ is a simple function, and $\int_E \theta\,d\mu < \infty$. We have

$$\int_E (\theta - f)\,d\mu = \sup\left\{\int_E \phi\,d\mu : \phi \text{ simple, } \phi \le \theta - f\chi_E\right\}\ .$$

If ϕ is a simple function with $\phi \le \theta - f\chi_E$, then $\psi = \theta - \phi$ is a simple function with $f\chi_E \le \psi \le \theta$, and conversely if $f\chi_E \le \psi \le \theta$; thus

$$\int_E (\theta - f)\,d\mu = \sup\left\{\int_E (\theta - \psi)\,d\mu : \phi \text{ simple, } f\chi_E \le \psi \le \theta\right\}\ .$$

If $\psi \le \theta$, we have

$$\int_E \theta\,d\mu = \int_E \psi\,d\mu + \int_E (\theta - \psi)\,d\mu$$

and, since the integrals are finite, we have

$$\int_E (\theta - \psi)\,d\mu = \int_E \theta\,d\mu - \int_E \psi\,d\mu\ .$$

Similarly, we have

$$\int_E f\,d\mu = \int_E \theta\,d\mu - \int_E (\theta - f)\,d\mu$$

$$= \int_E \theta\,d\mu - \sup\left\{\left[\int_E \theta\,d\mu - \int_E \psi\,d\mu\right] : \phi \text{ simple, } f\chi_E \le \psi \le \theta\right\}$$

$$= \int_E \theta \, d\mu - \left[\int_E \theta \, d\mu - \inf \left\{ \int_E \psi \, d\mu : \phi \text{ simple}, \, f\chi_E \leq \psi \leq \theta \right\} \right]$$
$$= \inf \left\{ \int_E \psi \, d\mu : \phi \text{ simple}, \, f\chi_E \leq \psi \leq \theta \right\} .$$



Here is a simpler version of XVIII.1.5.9. which captures its essence:

XVIII.1.5.10. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and f a bounded nonnegative real-valued measurable function on $(X, \mathcal{A})$. Then

$$\int_X f \, d\mu = \inf \left\{ \int_X \psi \, d\mu : \psi \text{ simple}, \, f \leq \psi \right\} .$$

In particular,

$$\sup \left\{ \int_X \phi \, d\mu : \phi \text{ simple}, \, \phi \leq f \right\} = \inf \left\{ \int_X \psi \, d\mu : \psi \text{ simple}, \, f \leq \psi \right\} .$$

XVIII.1.5.11. With more work, this theorem can be proved directly without use of the MCT or linearity of the integral. It then easily implies linearity of the integral for bounded functions with finite support (). From this one can obtain general additivity with some additional work (). One can then readily deduce Fatou's Lemma (), and obtain the MCT as a corollary of Fatou's Lemma (). This is the approach taken in [Roy88], for example. The approach we have taken, which is used e.g. in [MW99], seems a little better pedagogically.

The Role of Measurability

One can only reasonably define the integral directly for an indicator function or simple function if it is measurable. But the definition

$$\int_X f \, d\mu = \sup \left\{ \int_X \phi \, d\mu : \phi \text{ simple}, \, \phi \leq f \right\}$$

makes sense for any nonnegative extended real-valued function, measurable or not. Why not integrate arbitrary nonnegative functions this way? Since measurability is used in the MCT, and hence for linearity, it might be suspected that something goes wrong; in fact, the following converse of XVIII.1.5.10. shows that additivity must fail if we do not restrict to "almost measurable" functions.

XVIII.1.5.12. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and f a bounded nonnegative real-valued function on $(X, \mathcal{A})$. If

$$\sup\left\{\int_X \phi\, d\mu : \phi \text{ simple, } \phi \le f\right\} = \inf\left\{\int_X \psi\, d\mu : \psi \text{ simple, } f \le \psi\right\}$$

then f is equal a.e. to an $\mathcal{A}$-measurable function.

PROOF: The supremum/infimum I is finite. Choose a sequence (ϕ_n) of (measurable) simple functions with $\phi_n \le f$ for all n and $\int_X \phi_n\, d\mu \to I$. Replacing ϕ_n by $\max_{1\le k\le n} \phi_k$, we may assume (ϕ_n) is nondecreasing. Similarly, let (ψ_n) be a nonincreasing sequence of simple functions with $f \le \psi_n$ for all n and $\int_X \psi_n\, d\mu \to I$. Then (ϕ_n) and (ψ_n) converge pointwise to measurable functions g and h respectively, and $g \le f \le h$. By monotonicity,

$$\int_X g\, d\mu = \int_X h\, d\mu = I$$

so $\int_X (h-g)\, d\mu = 0$, and hence $h - g = 0$ a.e. (XVIII.1.3.11.). Thus $f = g$ a.e. 👍

XVIII.1.5.13. This result does not quite imply that f is measurable in general if the measure space is not complete (), only that f is equal a.e. to a measurable function. But if $(X, \mathcal{A}, \mu)$ is complete, e.g. if it is Lebesgue measure on $(X, \mathcal{L})$, then f must be measurable.

Restricting and Extending Measures

In this connection, we obtain a (hardly unexpected) relation between integrals with respect to restrictions of a measure, and in particular between a measure and its completion:

XVIII.1.5.14. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$, and $\nu = \mu|_{\mathcal{B}}$. If f is a nonnegative extended real-valued $\mathcal{B}$-measurable function, and $E \in \mathcal{B}$, then

$$\int_E f\, d\nu = \int_E f\, d\mu\ .$$

PROOF: This is a good example of bootstrapping. First suppose f is χ_B for some $B \in \mathcal{B}$. Then B and $B \cap E$ are in $\mathcal{A}$, so we have

$$\int_E \chi_B\, d\nu = \nu(B \cap E) = \mu(B \cap E) = \int_E \chi_B\, d\mu\ .$$

By linearity, if ϕ is a $\mathcal{B}$-measurable simple function, then $\int_E \phi\, d\nu = \int_E \phi\, d\mu$. Now, if f is a general nonnegative extended real-valued $\mathcal{B}$-measurable function, there is an increasing sequence (ϕ_n) of $\mathcal{B}$-measurable simple functions converging pointwise to f. Then by two applications of the MCT,

$$\int_E f\, d\nu = \sup_n \int_E \phi\, d\nu = \sup_n \int_E \phi\, d\mu = \int_E f\, d\mu\ .$$

👍

XVIII.1.5.15. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, with completion $(X, \bar{\mathcal{A}}, \bar{\mu})$. If f is a nonnegative extended real-valued $\mathcal{A}$-measurable function, and $E \in A$, then

$$\int_E f \, d\mu = \int_E f \, d\bar{\mu} \ .$$

XVIII.1.5.16. Thus when a measure is extended to a larger σ-algebra, more functions become measurable and hence potentially integrable, but the integrals of functions which were already measurable do not change.

XVIII.1.6. The Lebesgue Integral and Riemann Integral

We now examine the Lebesgue integral, which is by definition the integral on the measure space $(\mathbb{R}, \mathcal{L}, \lambda)$, and its relation with the Riemann integral. We only consider nonnegative functions for now; this is not a serious restriction for bounded functions on finite intervals since a constant function can be added to make them nonnegative. The Lebesgue integral for general functions is discussed in (). In this section, "measurable" will always mean "Lebesgue measurable."

XVIII.1.6.1. If E is a measurable subset of $\mathbb{R}$ and f a nonnegative extended real-valued measurable function on $\mathbb{R}$, the *Lebesgue integral* of f over E is

$$\int_E f\, d\lambda = \sup\left\{\int_E \phi\, d\lambda : \phi \text{ simple}, \phi \le f\right\}$$

where the integral of a simple function $\phi = \sum_{k=1}^n \alpha_k \chi_{A_k}$ is

$$\int_E \phi\, d\lambda = \sum_{k=1}^n \alpha_k \lambda(A_k \cap E)$$

as in XVIII.1.2.4.. If $E = [a, b]$ and f is Riemann-integrable on E, write the Riemann integral of f over E as

$$\int_a^b f(x)\, dx$$

to distinguish it notationally from the Lebesgue integral.

XVIII.1.6.2. If an extended real-valued measurable function is defined only on a measurable subset E of $\mathbb{R}$, we will always canonically extend it to a function on $\mathbb{R}$ by setting $f(x) = 0$ for $x \notin E$. The extended function is also measurable, and (if nonnegative) has the same integral over any (measurable) subset of E. Note the following fact which is used constantly:

XVIII.1.6.3. Proposition. Let I be an interval in $\mathbb{R}$, and f a nonnegative extended real-valued real-valued measurable function on $\mathbb{R}$ (or on a measurable subset of $\mathbb{R}$, extended canonically to $\mathbb{R}$). Then

$$\int_{\bar{I}} f\, d\lambda = \int_I f\, d\lambda\ .$$

Proof: We have $f\chi_{\bar{I}} = f\chi_I$ a.e. 👍

XVIII.1.6.4. More generally, if A and B are measurable subsets of $\mathbb{R}$ differing by a null set (i.e. $\lambda(A \triangle B) = 0$), then $f\chi_A = f\chi_B$ a.e., so $\int_A f\, d\lambda = \int_B f\, d\lambda$.

The following result shows that the Lebesgue integral agrees with the Riemann integral for Riemann-integrable functions, and thus the Lebesgue integral extends and generalizes the Riemann integral:

XVIII.1.6.5. THEOREM. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$, and $f : [a,b] \to \mathbb{R}$ a function. Extend f to $\mathbb{R}$ by setting $f(x) = 0$ for $x \notin [a,b]$. If f is Riemann-integrable on $[a,b]$, then f is Lebesgue-integrable on $\mathbb{R}$ and on $[a,b]$, and

$$\int_{\mathbb{R}} f\,d\lambda = \int_{[a,b]} f\,d\lambda = \int_a^b f(x)\,dx\ .$$

PROOF: It is almost transparent that the Riemann integral equals the Lebesgue integral for nonnegative step functions (cf. ()), which are nothing but piecewise-continuous simple functions. A general (necessarily bounded) nonnegative function f is Riemann integrable on $[a,b]$ if and only if there are an increasing sequence of (nonnegative) step functions (ϕ_n) (corresponding to lower Darboux sums) and a decreasing sequence of step functions (ψ_n) such that $\phi_n \le f \le \psi_n$ for all n and

$$\sup_n \left[\int_a^b \phi_n(x)\,dx\right] = \sup_n \left[\int_{[a,b]} \phi_n\,d\lambda\right] = \inf_n \left[\int_{[a,b]} \psi_n\,d\lambda\right] = \inf_n \left[\int_a^b \phi_n(x)\,dx\right]\ .$$

Thus by (), f is equal a.e. to a Lebesgue measurable function and, since Lebesgue measure is complete, f is Lebesgue measurable. By monotonicity, we have that

$$\int_{[a,b]} \phi_n\,d\lambda \le \int_{[a,b]} f\,d\lambda \le \int_{[a,b]} \psi_n\,d\lambda$$

$$\int_a^b \phi_n(x)\,dx \le \int_a^b f(x)\,dx \le \int_a^b \psi_n(x)\,dx$$

and since the suprema of the integrals on the left and the infima of the integrals on the right are all the same, it follows that

$$\int_{[a,b]} f\,d\lambda = \int_a^b f(x)\,dx\ .$$

XVIII.1.6.6. It follows easily from the Characterization Theorem () that a Riemann-integrable function is Lebesgue measurable. However, we did not need to use this theorem to obtain the result.

XVIII.1.6.7. **Notation.** If $a, b \in \mathbb{R}$, $a < b$, and f is a nonnegative Lebesgue measurable or a Lebesgue-integrable function on $[a,b]$, then

$$\int_{[a,b]} f\,d\lambda = \int_{(a,b)} f\,d\lambda = \int_{[a,b)} f\,d\lambda = \int_{(a,b]} f\,d\lambda\ .$$

we will write

$$\int_a^b f\,d\lambda$$

for the common value of these integrals. We can extend this notation by writing

$$\int_a^a f\,d\lambda = 0$$

for any a, and

$$\int_b^a f\,d\lambda = -\int_a^b f\,d\lambda\ .$$

With these notational conventions, if f is Lebesgue-integrable or nonnegative and Lebesgue measurable on an interval containing a, b, c, in any order and not necessarily distinct, then

$$\int_a^c f\,d\lambda = \int_a^b f\,d\lambda + \int_b^c f\,d\lambda\ .$$

If f is Riemann-integrable on $[a, b]$, then

$$\int_a^b f\,d\lambda = \int_a^b f(x)\,dx\ .$$

Note the distinction, however: the first expression denotes the Lebesgue integral and the second expression the Riemann integral.

Lebesgue Integrable Functions Which Are Not Riemann Integrable

There are many additional functions on $\mathbb{R}$ which are Lebesgue-integrable, as the following examples show.

XVIII.1.6.8. Examples. (i) The indicator function $\chi_{\mathbb{Q}}$ is not Riemann Integrable over any interval () because it is "too discontinuous." But it is Lebesgue integrable over $\mathbb{R}$, or any measurable subset E of $\mathbb{R}$ (in particular, over any interval), and

$$\int_E \chi_{\mathbb{Q}}\,d\lambda = 0$$

for any $E \in \mathcal{L}$.

(ii) Set

$$f(x) = \begin{cases} \frac{1}{\sqrt{x}} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \le 0 \end{cases}\ .$$

Then f is not Riemann integrable on $[0, 1]$ even though it is piecewise-continuous, since it is unbounded. But f is measurable, and Riemann integrable on $\left[\frac{1}{n}, 1\right]$ for any n, so

$$\int_{[0,1]} f\,d\lambda = \int_{(0,1]} f\,d\lambda = \sup_n \left[\int_{\left[\frac{1}{n},1\right]} f\,d\lambda\right]$$

$$= \sup_n \left[\int_{\frac{1}{n}}^1 \frac{1}{\sqrt{x}}\,dx\right] = \sup_n \left[2\sqrt{x}\Big|_{\frac{1}{n}}^1\right] = \sup_n \left[2 - \frac{2}{\sqrt{n}}\right] = 2$$

by XVIII.1.5.3., so f is Lebesgue integrable on $[0,1]$. In fact, f is Lebesgue integrable on any bounded interval (more generally, on any bounded measurable set) by a similar argument. Note that f is improperly Riemann integrable on $[0,1]$ (XVIII.1.6.9.).

(iii) Set

$$f(x) = \begin{cases} \frac{1}{x} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases} .$$

Then f is measurable, and Riemann integrable on $\left[\frac{1}{n}, 1\right]$ for any n, so

$$\int_{[0,1]} f \, d\lambda = \int_{(0,1]} f \, d\lambda = \sup_n \left[\int_{\left[\frac{1}{n},1\right]} f \, d\lambda \right]$$

$$= \sup_n \left[\int_{\frac{1}{n}}^1 \frac{1}{x} \, dx \right] = \sup_n \left[\log x \Big|_{\frac{1}{n}}^1 \right] = \sup_n \left[-\log \frac{1}{n} \right] = \infty$$

so f is not Lebesgue integrable over $[0,1]$. In fact, if

$$f(x) = \begin{cases} x^{-r} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases}$$

then f is Lebesgue integrable on $[0,1]$ if and only if $r < 1$. Note that these are exactly the power functions which are improperly Riemann integrable on $[0,1]$.

(iii) Set

$$f(x) = \begin{cases} \frac{1}{x^2} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases} .$$

Then f is measurable, but not Lebesgue integrable on $[0,1]$. It is Riemann integrable on $[1,b]$ for any $b > 1$, so

$$\int_{[1,\infty)} f \, d\lambda = \sup_n \left[\int_{[1,n]} f \, d\lambda \right]$$

$$= \sup_n \left[\int_1^n \frac{1}{x^2} \, dx \right] = \sup_n \left[-\frac{1}{x} \Big|_1^n \right] = \sup_n \left[1 - \frac{1}{n} \right] = 1$$

so f is Lebesgue integrable over $[1,\infty)$. In fact, if

$$f(x) = \begin{cases} x^{-r} & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x \leq 0 \end{cases}$$

then f is Lebesgue integrable on $[1,\infty)$ if and only if $r > 1$. Note that these are exactly the power functions which are improperly Riemann integrable on $[1,\infty)$.

The Lebesgue Integral and the Improper Riemann Integral

The fact that unbounded functions cannot be Riemann integrable is partially compensated by the various forms of improper Riemann integration.

XVIII.1.6.9. If f is a nonnegative function defined on $(a,b]$, and Riemann integrable on $[a+\epsilon,b]$ for every $\epsilon>0$, recall (IX.2.11.1.) that f is *improperly Riemann integrable* from a to b if

$$\lim_{\epsilon\to 0+}\left[\int_{a+\epsilon}^{b} f(x)\,dx\right]=\sup_{\epsilon>0}\left[\int_{a+\epsilon}^{b} f(x)\,dx\right]<\infty$$

and in this case, we define

$$\int_a^b f(x)\,dx=\lim_{\epsilon\to 0+}\left[\int_{a+\epsilon}^{b} f(x)\,dx\right]\ .$$

XVIII.1.6.10. Proposition. If f is a nonnegative function defined on $(a,b]$, and Riemann integrable on $[a+\epsilon,b]$ for every $\epsilon>0$, then f is improperly Riemann integrable from a to b if and only if f is Lebesgue integrable on $[a,b]$, and

$$\int_a^b f(x)\,dx=\int_{[a,b]} f\,d\lambda\ .$$

Proof: We have that $f\chi_{\left[a+\frac{1}{n},b\right]}$ is measurable for all n and $f\chi_{\left[a+\frac{1}{n},b\right]}\to f\chi_{(a,b]}$, so $f\chi_{(a,b]}$ is measurable. Thus

$$\int_{[a,b]} f\,d\lambda=\int_{(a,b]} f\,d\lambda=\sup_n\left[\int_{\left[a+\frac{1}{n},b\right]} f\,d\lambda\right]=\sup_n\left[\int_{a+\frac{1}{n}}^{b} f(x)\,dx\right]=\int_a^b f(x)\,dx$$

by XVIII.1.5.3., and in particular the suprema are either both finite or both infinite. 👍

XVIII.1.6.11. There are unbounded nonnegative functions which are continuous a.e. and Lebesgue integrable, but not improperly Riemann integrable in any reasonable sense. Let (r_k) be an enumeration of the left endpoints of the intervals removed in the construction of the Cantor set (), and

$$g(x)=\sum_{k=1}^{\infty}2^{-k}f(x-r_k)$$

where f is the function in XVIII.1.6.8.(ii). Then g is Lebesgue integrable on $[0,1]$ by XVIII.1.5.5., and continuous at all points not in the Cantor set, but not improperly Riemann integrable on $[0,1]$ in any of the usual senses.

XVIII.1.7. Continuity and Fatou's Lemma

To put the Monotone Convergence Theorem in perspective, we now note in more detail the failure of Continuity for integration (XVIII.1.1.1.) in general.

XVIII.1.7.1. EXAMPLES. (i) Let μ be counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. Let $E_n = \{k : k \geq n\}$. Then $\cap_n E_n = \emptyset$, so $\chi_{E_n} \to 0$ pointwise. But

$$\infty = \int_{\mathbb{N}} \chi_{E_n} \, d\mu \not\to \int_{\mathbb{N}} 0 \, d\mu = 0 \ .$$

Maybe the problem is just that the χ_{E_n} are not integrable. But:
(ii) Consider $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \mu)$ again. Let $f_n = \frac{1}{n}\chi_{\{1,\ldots,n\}}$. Then $f_n \to 0$, even uniformly; however,

$$1 = \int_{\mathbb{N}} f_n \, d\mu \not\to \int_{\mathbb{N}} 0 \, d\mu = 0 \ .$$

(iii) Consider Lebesgue measure on $[0, 1]$. Let f_n be the continuous function with the following graph:

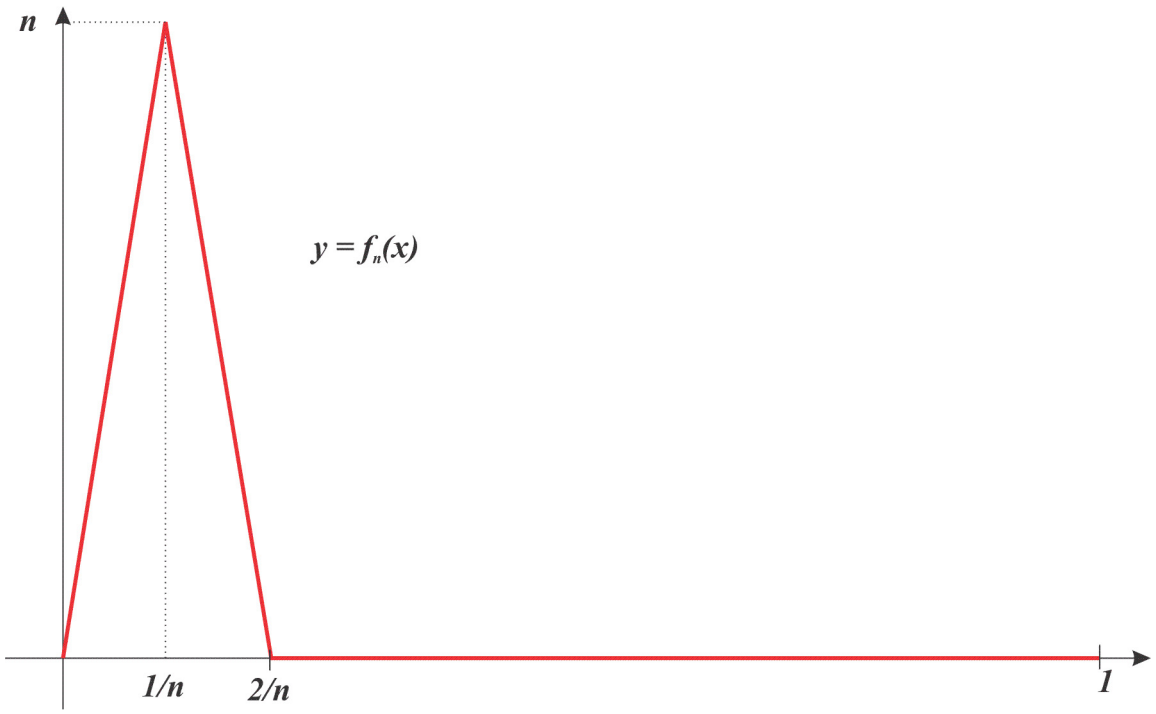


Figure XVIII.1: A continuous function

If $x \in (0,1]$, then for sufficiently large N we have $\frac{2}{N} \leq x$, so $f_n(x) = 0$ for all $n \geq N$. And $f_n(0) = 0$ for all n. So $f_n \to 0$ pointwise. However,

$$1 = \int_{[0,1]} f_n \, d\lambda \not\to \int_{[0,1]} 0 \, d\lambda = 0 \ .$$

Maybe the problem is that the functions are not uniformly bounded. But it is not quite so simple (cf. also (ii)):

(iv) Suppose we take f_n as in (iii), but the height of the triangle is n^r for fixed r. Then $f_n \to 0$ pointwise as before. If $r > 0$, the f_n are not uniformly bounded. But

$$\int_{[0,1]} f_n \, d\lambda = n^{r-1}$$

so $\int_{[0,1]} f_n \, d\lambda \to \int_{[0,1]} 0 \, d\lambda$ if and only if $r < 1$.

XVIII.1.7.2. One might note that in (iv), the difference between the case $r < 1$ and the case $r \geq 1$ is that in the case $r < 1$, there is an integrable function g such that $f_n \leq g$ for all n, namely $g(x) = x^{-r}$ (with $g(0) = 0$). (It may not be obvious that there is no such g if $r \geq 1$.) This is also the root of the difficulty in (ii). In fact, the Dominated Convergence Theorem (cf. XVIII.1.7.10.) states that existence of such a g is sufficient for convergence of the integrals.

Fatou's Lemma

XVIII.1.7.3. Another observation which might be made in these examples is that the integral of the limit is always less than or equal to the limit of the integrals. Actually this is a no-brainer in these examples since the functions are nonnegative and the integral of the limit function is zero; but experiments with less trivial similar examples suggest this is a general principle.

XVIII.1.7.4. Another motivation for why this should be so is that if (a_n) and (b_n) are (nonnegative) sequences of real numbers, then

$$\liminf_{n\to\infty} a_n + \liminf_{n\to\infty} b_n \leq \liminf_{n\to\infty} (a_n + b_n)$$

(), and equality does not hold in general, i.e. "the lim inf of a sum is greater than or equal to the sum of the lim inf's." Since integration is a continuous version of summation, we should have that "the integral of a lim inf is less than or equal to the lim inf of the integrals."

The second of the three main convergence theorems of measure theory asserts this:

XVIII.1.7.5. THEOREM. [FATOU'S LEMMA] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (f_n) a sequence of nonnegative extended real-valued measurable functions on $(X, \mathcal{A})$. Set

$$f(x) = \liminf_{n\to\infty} f_n(x) = \lim_{n\to\infty} \left[\inf_{k\geq n} f_k(x) \right] = \sup_n \left[\inf_{k\geq n} f_k(x) \right]$$

(note that f is a well defined nonnegative extended real-valued measurable function). If $E \in \mathcal{A}$, then

$$\int_E f\,d\mu \leq \liminf_{n\to\infty} \int_E f_n\,d\mu\ .$$

The following special case is the one most commonly used:

XVIII.1.7.6. Corollary. Let $(X, \mathcal{A}, \mu)$ be a measure space, and (f_n) a sequence of nonnegative extended real-valued measurable functions on $(X, \mathcal{A})$, converging pointwise a.e. to a measurable function f. If $E \in \mathcal{A}$, then

$$\int_E f\,d\mu \leq \liminf_{n\to\infty} \int_E f_n\,d\mu\ .$$

XVIII.1.7.7. Fatou's Lemma has the most general hypotheses of the three major convergence theorems, but also not surprisingly the weakest conclusion.

The proof of Fatou's Lemma is easy given the Monotone Convergence Theorem:

Proof: For each n set

$$g_n = \inf_{k\geq n} f_k\ .$$

Then (g_n) is a nondecreasing sequence of nonnegative extended real-valued measurable functions, and $g_n \to f$ pointwise. We also have $g_n \leq f_k$ for all $k \geq n$, and thus

$$\int_E g_n\,d\mu \leq \int_E f_k\,d\mu$$

for $k \geq n$ by monotonicity. Thus by the MCT we have

$$\int_E f\,d\mu = \sup_n \left[\int_E g_n\,d\mu\right] \leq \sup_n \left[\inf_{k\geq n} \int_E f_k\,d\mu\right] = \liminf_{n\to\infty} \int_E f_n\,d\mu\ .$$



XVIII.1.7.8. Note that the MCT also follows almost immediately from Fatou's Lemma: if the sequence (f_n) is nondecreasing, we get $\int_E f\,d\mu \geq \sup_n \int_E f_n\,d\mu$ by monotonicity and $\int_E f\,d\mu \leq \liminf_n \int_E f_n\,d\mu$ by Fatou, so $\int_E f\,d\mu = \lim_n \int_E f_n\,d\mu$.

Of course, we used the MCT to prove Fatou's Lemma; but there are approaches where Fatou's Lemma is proved first and the MCT obtained as a consequence (cf. ()).

XVIII.1.7.9. The examples of XVIII.1.7.1. show that the inequality in Fatou's Lemma cannot be replaced by an equality in general.

The Dominated Convergence Theorem for Nonnegative Functions

We can then obtain almost immediately the nonnegative version of the Dominated Convergence Theorem:

XVIII.1.7.10. THEOREM. [DOMINATED CONVERGENCE THEOREM, NONNEGATIVE VERSION] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (f_n) a sequence of nonnegative extended real-valued measurable functions on $(X, \mathcal{A})$, converging pointwise a.e. to a measurable function f. Suppose there is a nonnegative integrable function g with $f_n \leq g$ a.e. for all n. If $E \in \mathcal{A}$, then f is integrable on E and

$$\int_E f\, d\mu = \lim_{n\to\infty} \int_E f_n\, d\mu$$

(in particular, the limit exists).

PROOF: Note that f is automatically integrable since $f \leq g$ a.e. We have that $g - f_n$ is nonnegative a.e. for all n and $g - f_n \to g - f$ a.e. By two applications of Fatou's Lemma, we have

$$\int_E f\, d\mu \leq \liminf_{n\to\infty} \int_E f_n\, d\mu$$

$$\int_E g\, d\mu - \int_E f\, d\mu = \int_E (g - f)\, d\mu \leq \liminf_{n\to\infty} \int_E (g - f_n)\, d\mu$$

$$= \liminf_{n\to\infty} \left[\int_E g\, d\mu - \int_E f_n\, d\mu \right] = \int_E g\, d\mu - \limsup_{n\to\infty} \int_E f_n\, d\mu$$

(the integrals are finite so subtraction is valid). Rearranging the last inequality, we obtain

$$\int_E f\, d\mu \geq \limsup_{n\to\infty} \int_E f_n\, d\mu\ .$$



XVIII.1.7.11. The general DCT is essentially the same but applies to integrable functions which are not necessarily nonnegative. The g is called a *dominating function* for the sequence (f_n), and the reason for the name of the theorem is obvious.

XVIII.1.7.12. COROLLARY. [BOUNDED CONVERGENCE THEOREM, NONNEGATIVE VERSION] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (f_n) a uniformly bounded sequence of nonnegative real-valued measurable functions on $(X, \mathcal{A})$, converging pointwise a.e. to a measurable function f. If $E \in \mathcal{A}$ with $\mu(E) < \infty$, then f is integrable on E and

$$\int_E f\, d\mu = \lim_{n\to\infty} \int_E f_n\, d\mu$$

(in particular, the limit exists).

PROOF: If $f_n(x) \leq M$ for all n and all $x \in E$, then $M\chi_E$ is a dominating function for $(f_n\chi_E)$. 

XVIII.1.7.13. The restriction that $\mu(E) < \infty$ is necessary in general, even if the f_n and f are assumed to be integrable on E and the convergence is uniform (XVIII.1.7.1.(ii)).

XVIII.1.8. The Integral For General Functions

We can extend the integral to many measurable functions which take both positive and negative values, as long as we carefully avoid the $\infty - \infty$ problem. The simplest and most satisfactory way to proceed is:

XVIII.1.8.1. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a real-valued function on X. Then f is *μ-integrable* if $f = g - h$ for nonnegative real-valued integrable functions g and h. Define the *integral* of f to be

$$\int_X f\,d\mu = \int_X g\,d\mu - \int_X h\,d\mu\ .$$

If $E \in \mathcal{A}$, then $f\chi_E = g\chi_E - h\chi_E$ is also μ-integrable; define

$$\int_E f\,d\mu = \int_X f\chi_E\,d\mu = \int_E g\,d\mu - \int_E h\,d\mu\ .$$

XVIII.1.8.2. Note that we are here only considering real-valued functions, since if g and h are nonnegative extended real-valued functions, $g - h$ might not be well defined. We consider extended real-valued functions in ().

A μ-integrable function is a difference of two measurable functions, hence automatically measurable. We often write "integrable" for "μ-integrable" if the μ is understood.

We have defined the integral of an integrable function in the only possible way to obtain a linear integral; however, it is not entirely obvious that the integral is independent of the choice of g and h. But it is:

XVIII.1.8.3. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a real-valued integrable function on X. Then the integral of f is well defined.

PROOF: Suppose $f = g_1 - h_1 = g_2 - h_2$ with g_i, h_i nonnegative real-valued integrable functions. Then $g_1 + h_2 = g_2 + h_1$, so, for any $E \in \mathcal{A}$,

$$\int_E g_1\,d\mu + \int_E h_2\,d\mu = \int_E (g_1 + h_2)\,d\mu = \int_E (g_2 + h_1)\,d\mu = \int_E g_2\,d\mu + \int_E h_1\,d\mu$$

and since the integrals are finite we may rearrange the equation by subtraction to obtain

$$\int_E g_1\,d\mu - \int_E h_1\,d\mu = \int_E g_2\,d\mu - \int_E h_2\,d\mu\ .$$



XVIII.1.8.4. Although the definition in XVIII.1.8.1. is logically satisfactory in light of XVIII.1.8.3., it is perhaps uncomfortably indefinite because the g and h are (usually) not unique. It is more satisfying to have a canonical way to write a general integrable function as a difference of two nonnegative integrable functions. This is possible:

XVIII.1.8.5. DEFINITION. Let X be a set, and $f : X \to \mathbb{R}$ a function. The *positive part* of f is $f_+ = \max(f, 0)$ and the *negative part* of f is $f_- = -\min(f, 0) = |\min(f, 0)|$.

We have

$$f_+(x) = \begin{cases} f(x) & \text{if} \quad f(x) \geq 0 \\ 0 & \text{if} \quad f(x) < 0 \end{cases}$$

$$f_-(x) = \begin{cases} |f(x)| & \text{if} \quad f(x) \leq 0 \\ 0 & \text{if} \quad f(x) > 0 \end{cases}$$

so f_+ and f_- are nonnegative functions with the properties $f_+ - f_- = f$, $f_+ + f_- = |f|$, $f_+ f_- = 0$. Conversely, if g and h are nonnegative functions with $f = g - h$ and $gh = 0$, then necessarily $g = f_+$ and $h = f_-$.

If $(X, \mathcal{A})$ is a measurable space and f is measurable, then f_+ and f_- are also measurable. (And if $(X, \mathcal{T})$ is a topological space and f is continuous, then f_+ and f_- are also continuous.)

XVIII.1.8.6. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a real-valued measurable function on X. Then f is integrable if and only if f_+ and f_- are integrable. If f is integrable, then

$$\int_E f \, d\mu = \int_E f_+ \, d\mu - \int_E f_- \, d\mu$$

for any $E \in \mathcal{A}$.

PROOF: If f_+ and f_- are integrable, then f is integrable and by definition

$$\int_E f \, d\mu = \int_E f_+ \, d\mu - \int_E f_- \, d\mu$$

for any $E \in \mathcal{A}$. Conversely, if f is integrable, write $f = g - h$ for nonnegative integrable functions g and h. If $x \in X$ and $f(x) \geq 0$, then $f_+(x) = f(x) = g(x) - h(x) \leq g(x)$; if $f(x) < 0$, then $0 = f_+(x) \leq g(x)$. Thus $f_+ \leq g$. Similarly, $f_- \leq h$. Thus f_+ and f_- are integrable. 👍

XVIII.1.8.7. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a real-valued function on X. If f is integrable, then $|f|$ is integrable. If f is measurable and $|f|$ is integrable, then f is integrable.

PROOF: If f is integrable, then f_+ and f_- are integrable, so $|f| = f_+ + f_-$ is integrable. Conversely, if f is measurable and $|f|$ is integrable, then $f_+, f_- \leq |f|$ so f_+ and f_- are integrable. 👍

XVIII.1.8.8. Note that it is not quite correct to say that f is integrable if and only if $|f|$ is integrable: if $|f|$ is integrable, f might not be measurable. For example, if A is a subset of $[0, 1]$ which is not Lebesgue measurable, set ✋

$$f(x) = \begin{cases} 1 & \text{if} \quad x \in A \\ -1 & \text{if} \quad x \in [0,1] \setminus A \\ 0 & \text{if} \quad x \notin [0,1] \end{cases} .$$

Then $|f| = \chi_{[0,1]}$ is Lebesgue integrable, but f is not Lebesgue measurable.

It is correct to say that *if* f is measurable, then f is integrable if and only if $|f|$ is integrable. (Note that if f is measurable, then $|f|$ is also measurable.)

XVIII.1.8.9. Thus, except for the measurability problem, a function is integrable if and only if its absolute value is integrable. So the theory of abstract Lebesgue integration is an "absolute" integration theory, analogous to the theory of absolutely convergent infinite series. In fact:

XVIII.1.8.10. COROLLARY. Let μ be counting measure on $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. If $x = (x_1, x_2, \dots)$ is a sequence in $\mathbb{R}$ (a real-valued function on $\mathbb{N}$), then x is μ-integrable if and only if the infinite series

$$\sum_{k=1}^{\infty} x_k$$

converges absolutely. We then have

$$\int_{\mathbb{N}} x \, d\mu = \sum_{k=1}^{\infty} x_k \ .$$

Integration of Complex Functions

XVIII.1.8.11. If $(X, \mathcal{A}, \mu)$ is a measure space and f is a complex-valued $\mathcal{A}$-measurable function, we say that f is μ-integrable if and only if both $Re(f)$ and $Im(f)$ are μ-integrable, and set

$$\int_A f \, d\mu = \int_A Re(f) \, d\mu + i \int_A Im(f) \, d\mu$$

for $A \in \mathcal{A}$. We then again have that if f is measurable, then f is μ-integrable if and only if $|f|$ is μ-integrable, because of the inequalities

$$|Re(f)|, |Im(f)| \leq |f| \leq (|Re(f)| + |Im(f)| \ .$$

Properties of the Integral

XVIII.1.8.12. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space.

(i) If f and g are μ-integrable, then $f + g$ is μ-integrable and

$$\int_A (f + g) \, d\mu = \int_A f \, d\mu + \int_A g \, d\mu$$

for all $A \in A$.

(ii) If f is μ-integrable and $\alpha \in \mathbb{C}$, then αf is μ-integrable and

$$\int_A \alpha f \, d\mu = \alpha \int_A f \, d\mu$$

for all $A \in A$.

(iii) If f and g are μ-integrable and real-valued, and $f \leq g$ a.e., then

$$\int_A f \, d\mu \leq \int_A g \, d\mu$$

for all $A \in \mathcal{A}$.

(iv) If f is μ-integrable, then

$$\left| \int_A f \, d\mu \right| \leq \int_A |f| \, d\mu$$

for all $A \in \mathcal{A}$.

PROOF: (i)–(iii) are straightforward and left to the reader (for (iii), consider $g - f$). For (iv), approximate f by simple functions and use the triangle inequality in $\mathbb{C}$. 👍

A Vanishing Criterion

There is a convenient and useful criterion for an integrable function to vanish almost everywhere:

XVIII.1.8.13. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f an integrable extended real-valued function on X. If $\int_A f \, d\mu = 0$ for all $A \in \mathcal{A}$, then $f = 0$ a.e.

PROOF: Let $A = \{x \in X : f(x) \geq 0\}$ and $B = \{x \in X : f(x) \leq 0\}$. Since $\int_A f \, d\mu = \int_B f \, d\mu = 0$, we have that $f|_A$ and $f|_B$ are zero a.e. by XVIII.1.3.11.. 👍

Actually we only need to assume that $\int_A f \, d\mu$ for a set of A's which generate $\mathcal{A}$ as a σ-algebra, since it then follows that $\int_X f \, d\mu = 0$, and under this assumption it is easily checked that $\{B \in \mathcal{A} : \int_B f \, d\mu = 0\}$ is a σ-algebra. The set of such A even only needs to generate $\mathcal{A}$ up to null sets. The following special case is especially useful:

XVIII.1.8.14. COROLLARY. Let $[a, b]$ be an interval in $\mathbb{R}$, and f a Lebesgue integrable function on $[a, b]$. If

$$\int_a^x f \, d\lambda = 0$$

for all $x \in [a, b]$, then $f = 0$ a.e. on $[a, b]$.

XVIII.1.9. Exercises

XVIII.1.9.1. Prove the following converse to XVIII.1.8.12.(iv). Let $(X, \mathcal{A}, \mu)$ be a measure space, f a complex-valued μ-integrable function, and g a nonnegative $\mathcal{A}$-measurable function. If

$$\left| \int_A f \, d\mu \right| \leq \int_A g \, d\mu$$

for all $A \in \mathcal{A}$, then $|f| \leq g$ a.e. [If $|f| > g$ on a set A with $\mu(A) > 0$, find a set $B \in \mathcal{A}$, $B \subseteq A$, $\mu(B) > 0$, on which $arg(f)$ is almost constant, to obtain a contradiction.]

XVIII.1.9.2. Let $(X, \mathcal{A}, \mu)$ be a measure space, and g a nonnegative measurable function on $(X, \mathcal{A})$.
(a) For $A \in \mathcal{A}$, define

$$\nu(A) = \int_A g \, d\mu \ .$$

Show that ν is a measure on $(X, \mathcal{A})$, which is finite if and only if g is μ-integrable.
(b) Show that if f is $\mathcal{A}$-measurable, then f is ν-integrable if and only if fg is μ-integrable, and that

$$\int_A f \, d\nu = \int_A fg \, d\mu$$

for every $A \in \mathcal{A}$. [Hint: bootstrap.]

XVIII.2. The Dominated Convergence Theorem

Except for the Fundamental Theorem of Calculus () which is the basis for the actual calculation of elementary integrals, the Dominated Convergence Theorem is probably the most important theorem in integration theory in the sense that it is the result used most often in applications, although the Fubini-Tonelli Theorem is a close second.

XVIII.2.1. Applications of the DCT

Here is a general result about interchanging the order of differentiation and integration. It is phrased in terms of Lebesgue measure, but can be adapted to general measure spaces with the same proof (Exercise ()).

XVIII.2.1.1. THEOREM. Let I and J be intervals in $\mathbb{R}$, and let $F : I \times J \to \mathbb{R}$. Assume that

(a) For each $y \in J$, the function $f^{[y]} : I \to \mathbb{R}$ defined by $f^{[y]}(x) = F(x, y)$ is Lebesgue measurable.

(b) There is a $y_0 \in J$ such that $f^{[y_0]}$ is Lebesgue integrable on I.

(c) For almost all $x \in I$, $\frac{\partial F}{\partial y}(x, y)$ exists for all $y \in J$.

(d) There is a nonnegative Lebesgue integrable function g on I such that, for almost all $x \in I$,

$$\left|\frac{\partial F}{\partial y}(x, y)\right| \leq g(x)$$

for all $y \in J$.

Then we have

(i) $f^{[y]}$ is Lebesgue integrable on I for all $y \in J$.

(ii) For every $y \in J$, the function ϕ^y defined by $\phi^y(x) = \frac{\partial F}{\partial y}(x, y)$ (defined a.e.) is integrable on I.

(iii) For every $y_1 \in J$,

$$\frac{d}{dy}\left[\int_I F(x, y)\, d\lambda(x)\right](y_1) = \frac{d}{dy}\left[\int_I f^{[y]}(x)\, d\lambda(x)\right](y_1) \text{ exists and equals } \int_I \frac{\partial F}{\partial y}(x, y_1)\, d\lambda(x)\ .$$

PROOF: Let A be a subset of I such that (c) and (d) hold for all $x \in A$ and $\lambda(I \setminus A) = 0$.
(i): Fix $y \in J$, and set $\psi(x) = F(x, y) - F(x, y_0)$. Thus $f^{[y]} = f^{[y_0]} + \psi$, so to show that $f^{[y]}$ is Lebesgue integrable it suffices to show that ψ is Lebesgue integrable. Note that it is Lebesgue measurable by (a). For any $x \in A$, there is a c (depending on x) such that

$$\psi(x) = \frac{\partial F}{\partial y}(x, c)(y - y_0)$$

by the Mean Value Theorem (), and hence we have

$$|\psi(x)| \leq g(x)|y - y_0|$$

for almost all $x \in I$ by (d). Since $y - y_0$ is a constant, ψ is Lebesgue integrable.
(ii): Fix $y_1 \in J$. If (h_n) is any sequence converging to 0 such that $y_1 + h_n \in J$ for all n, set

$$\phi_n(x) = \frac{F(x, y_1 + h_n) - F(x, y_1)}{h_n}$$

for each n and note that each ϕ_n is Lebesgue measurable by (a). We have $\phi_n \to \phi^{y_1}$ pointwise a.e., so ϕ^{y_1} is Lebesgue measurable. Since $|\phi^{y_1}(x)| \leq g(x)$ for almost all x by (d), ϕ^{y_1} is Lebesgue integrable.
(iii): We use the sequential criterion. Fix $y_1 \in J$, and let

$$L = \int_I \frac{\partial F}{\partial y}(x, y_1)\, d\lambda(x) = \int_I \phi^{y_1}(x)\, d\lambda(x)$$

(note that this integral is defined by (ii)). To prove (iii), it suffices to show that whenever (h_n) is a sequence of nonzero numbers such that $y_1 + h_n \in J$ for all n, we have

$$\lim_{n\to\infty} \frac{\int_I F(x, y_1 + h_n)\, d\lambda(x) - \int_I F(x, y_1)\, d\lambda(x)}{h_n} = L\ .$$

We have

$$\frac{\int_I F(x, y_1 + h_n)\, d\lambda(x) - \int_I F(x, y_1)\, d\lambda(x)}{h_n} = \int_I \frac{F(x, y_1 + h_n) - F(x, y_1)}{h_n}\, d\lambda(x) = \int_I \phi_n(x)\, d\lambda(x)$$

where ϕ_n is as in the proof of (ii). For each n and $x \in A$, by the MVT there is a c_n (depending on x as well as n) between y_1 and $y_1 + h_n$ such that

$$\phi_n(x) = \frac{F(x, y_1 + h_n) - F(x, y_1)}{h_n} = \frac{\partial F}{\partial y}(x, c_n)$$

and thus $|\phi_n(x)| \leq g(x)$ for all n and almost all x by (d). Thus by the DCT,

$$\lim_{n\to\infty} \int_I \phi_n(x)\, d\lambda(x) = \int_I \phi^{y_1}(x)\, d\lambda(x) = L$$

since $\phi_n \to \phi^{y_1}$ pointwise a.e., and the result follows. 👍

XVIII.2.1.2. Hypothesis (a) is not automatic, as the following example shows. For (a) to hold, it suffices that F be a Borel measurable function from $I \times J$ to $\mathbb{R}$. It does not quite suffice for F to be Lebesgue measurable (), but in the presence of (c) Lebesgue measurability might be sufficient. Hypotheses (a) and (b) are obviously necessary to even make sense of the left side of the equation in (iii), and hypothesis (c) is necessary to make sense of the right side (technically what we need to define the right side is that for each y, $\frac{\partial F}{\partial y}(x, y)$ is defined for almost all x).

XVIII.2.1.3. Example. Let $I = J = [-1, 1]$. Let A be a nonmeasurable subset of I, and set

$$F(x, y) = \begin{cases} y & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases}\ .$$

Then F satisfies (b), (c), and (d), but not (a).

XVIII.2.2. Uniform Integrability and Vitali's Convergence Theorem

There is another convergence theorem, the Vitali Convergence Theorem, which gives a necessary and sufficient condition for convergence of integrals, although it is of limited practical use since the hypotheses are not usually easy to check. The result is most cleanly stated for finite measures, but there is a version for general measures.

Equiintegrability and Uniform Integrability

We first note the following property of integrals:

XVIII.2.2.1. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f an extended real-valued or complex-valued $\mathcal{A}$-measurable function on X. If f is μ-integrable, then for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $A \in \mathcal{A}$, $\mu(A) < \delta$,

$$\int_A |f|\, d\mu < \epsilon \ .$$

If μ is finite and has no atoms on which f takes an infinite value, the converse also holds: if for every $\epsilon > 0$ there is such a $\delta > 0$, then f is μ-integrable.

PROOF: There is a nonnegative simple function s with $s \le |f|$ and

$$\int_X s\, d\mu > \int_X |f|\, d\mu - \frac{\epsilon}{2} \ .$$

Then s is bounded, i.e. there is an $M > 0$ such that $|s(x)| \le M$ for all $x \in X$. Set $\delta = \frac{\epsilon}{2M}$. Then, if $A \in \mathcal{A}$, $\mu(A) < \delta$,

$$\int_A |f|\, d\mu = \int_A s\, d\mu + \int_A (|f| - s)\, d\mu \le M\mu(A) + \int_X (|f| - s)\, d\mu < M\frac{\epsilon}{2M} + \frac{\epsilon}{2} = \epsilon \ .$$

Conversely, suppose μ is finite and has no atoms on which f takes an infinite value. Fix $\delta > 0$ such that if $A \in \mathcal{A}$, $\mu(A) < \delta$, then $\int_A |f|\, d\mu < 1$. Then X can be written as a disjoint union

$$X = A_1 \cup \cdots \cup A_n \cup B_1 \cup \cdots \cup B_m$$

for some n with $\mu(A_k) < \delta$ for all k, where $B_1, \ldots, B_m$ are the atoms of μ of measure $\ge \delta$ (there are only finitely many since μ is finite). Note that f is constant and finite a.e. on each B_j, say $f = \beta_j$ a.e. on B_j. Then

$$\int_X |f|\, d\mu = \sum_{k=1}^{n} \int_{A_k} |f|\, d\mu + \sum_{j=1}^{m} \int_{B_j} |f|\, d\mu < n + \sum_{j=1}^{m} |\beta_j| \mu(B_j) < \infty$$

so f is integrable. 👍

This leads to the following definition:

XVIII.2.2.2. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\mathcal{F}$ a set of μ-integrable functions on X. Then $\mathcal{F}$ is *equiintegrable* (μ-*equiintegrable*) on X if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $A \in \mathcal{A}$ with $\mu(A) < \delta$,

$$\int_A |f|\, d\mu < \epsilon$$

for all $f \in \mathcal{F}$. If (f_n) is a sequence of μ-integrable functions, then (f_n) is equiintegrable if $\{f_n : n \in \mathbb{N}\}$ is equiintegrable.

There is a closely related notion, based on the next property:

XVIII.2.2.3. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f a μ-integrable function. Then for any $\epsilon > 0$ there is a λ such that

$$\int_{E_\lambda} |f|\, d\mu < \epsilon$$

where $E_\lambda = \{x \in X : |f(x)| > \lambda\}$.

PROOF: For $\lambda > 0$ set $D_\lambda = X \setminus E_\lambda = \{x \in X : |f(x)| \leq \lambda\}$. Then as $\lambda \to \infty$, $|f|\chi_{D_\lambda}$ increases to $|f|$ pointwise a.e. So by the MCT

$$\int_X |f|\, d\lambda = \sup_\lambda \int_X |f|\chi_{D_\lambda}\, d\mu = \int_{D_\lambda} |f|\, d\mu$$

so, as $\lambda \to \infty$,

$$\int_{E_\lambda} |f|\, d\mu = \int_X |f|\, d\mu - \int_{D_\lambda} |f|\, d\mu \to 0\ .$$



XVIII.2.2.4. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\mathcal{F}$ a set of μ-integrable functions on X. Then $\mathcal{F}$ is *uniformly integrable* (*uniformly* μ-*integrable*) on X if for every $\epsilon > 0$ there is a λ such that

$$\int_{E_{f,\lambda}} |f|\, d\mu < \epsilon$$

for all $f \in \mathcal{F}$, where $E_{f,\lambda} = \{x \in X : |f(x)| > \lambda\}$. If (f_n) is a sequence of μ-integrable functions, then (f_n) is uniformly integrable if $\{f_n : n \in \mathbb{N}\}$ is uniformly integrable.

XVIII.2.2.5. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and $\mathcal{F}$ a set of μ-integrable functions on X. Then $\mathcal{F}$ is uniformly integrable if and only if $\mathcal{F}$ is equiintegrable and there is a constant K such that

$$\int_X |f|\, d\mu \leq K$$

for all $f \in \mathcal{F}$ ($\mathcal{F}$ is L^1-*bounded*). If $(X, \mathcal{A}, \mu)$ has no atoms, then $\mathcal{F}$ is uniformly integrable if and only if it is equiintegrable.

PROOF: Suppose $\mathcal{F}$ is uniformly integrable. Let $\epsilon > 0$. Fix λ such that

$$\int_{E_{f,\lambda}} |f|\, d\mu < \frac{\epsilon}{2}$$

for all $f \in \mathcal{F}$. Set $\delta = \frac{\epsilon}{2\lambda}$. If $A \in \mathcal{A}$, $\mu(A) < \delta$, then for each $f \in \mathcal{F}$ we have

$$\begin{aligned}\int_A |f|\, d\mu &= \int_{A\cap D_{f,\lambda}} |f|\, d\mu + \int_{A\cap E_{f,\lambda}} |f|\, d\mu \\ &\leq \lambda\mu(A) + \int_{E_{f,\lambda}} |f|\, d\mu < \lambda\frac{\epsilon}{2\lambda} + \frac{\epsilon}{2} = \epsilon\ .\end{aligned}$$

Thus $\mathcal{F}$ is equiintegrable. We also have, for all $f \in \mathcal{F}$,

$$\int_X |f|\, d\mu = \int_{D_{f,\lambda}} |f|\, d\mu + \int_{E_{f,\lambda}} |f|\, d\mu \leq \lambda\mu(X) + \frac{\epsilon}{2}$$

so $\mathcal{F}$ is L^1-bounded.

Conversely, if μ has no atoms, suppose $\mathcal{F}$ is equiintegrable. Fix $\delta_0 > 0$ such that if $A \in \mathcal{A}$, $\mu(A) < \delta_0$, then $\int_A |f|\, d\mu < 1$ for all $f \in \mathcal{F}$. Since X has no atoms, X can be written as a disjoint union $X = A_1 \cup \cdots \cup A_n$ for some n, where $\mu(A_k) < \delta_0$ for all k (). Then, for all $f \in \mathcal{F}$

$$\int_X |f|\, d\mu = \sum_{k=1}^{n} \int_{A_k} |f|\, d\mu < n$$

so $\mathcal{F}$ is L^1-bounded.

Now let $\epsilon > 0$. For any $\lambda > 0$ we have, for any $f \in \mathcal{F}$,

$$\mu(E_{f,\lambda}) \leq \frac{1}{\lambda}\int_X |f|\, d\mu < \frac{n}{\lambda}\ .$$

If $\delta > 0$ is fixed so that $\int_A |f|\, d\mu < \epsilon$ for all $A \in \mathcal{A}$, $\mu(A) < \delta$, and all $f \in \mathcal{F}$, and λ is fixed so that $\frac{n}{\lambda} < \delta$, we have, for all $f \in \mathcal{F}$,

$$\int_{E_{f,\lambda}} |f|\, d\mu < \epsilon$$

so $\mathcal{F}$ is uniformly integrable. 👍

XVIII.2.2.6. Thus the only difference between equiintegrability and uniform integrability on finite measure spaces is uniform boundedness on each atom. This is not automatic: if X is a one-point space with $\mu(X) = 1$, let $\mathcal{F}$ be the set of all real-valued (constant) functions on X. Then $\mathcal{F}$ is vacuously equiintegrable (each f is integrable, and for any $\epsilon > 0$ just take $\delta = \frac{1}{2}$), but $\mathcal{F}$ is not uniformly integrable.

XVIII.2.2.7. There are various definitions of equiintegrability and uniform integrability in the literature, all of which agree for finite measure spaces without atoms. In some references, equiintegrability (by our definition) is called "uniform integrability" or "uniformly absolutely continuous integrals." ✋

Vitali's Convergence Theorem, Finite Case

XVIII.2.2.8. THEOREM. [VITALI CONVERGENCE THEOREM] Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and (f_n) a sequence of μ-integrable functions converging pointwise a.e. on X to a measurable function f. If (f_n) is equiintegrable, then (f_n) is uniformly integrable, f is μ-integrable, and

$$\lim_{n\to\infty} \int_X |f_n - f|\, d\mu = 0 \ .$$

In particular,

$$\int_E f_n \, d\mu \to \int_E f \, d\mu$$

for all $E \in \mathcal{A}$.

Conversely, if $f_n \to f$ pointwise a.e., f and the f_n are integrable, and

$$\lim_{n\to\infty} \int_X |f_n - f|\, d\mu = 0$$

then (f_n) is uniformly integrable.

PROOF: Since each f_n takes infinite values only on a null set and $f_n \to f$ a.e., f is finite a.e. Thus, modifying f_n and f on null sets, we may assume that f and the f_n are finite everywhere and $f_n \to f$ everywhere.

If B is an atom of μ, then each f_n takes a constant value β_n a.e. on B, f takes a constant value β a.e. on B, and $\beta_n \to \beta$. Thus

$$\int_B f_n \, d\mu = \beta_n \mu(B) \to \beta\mu(B) = \int_B f \, d\mu \ .$$

In particular, f is integrable over B.

Suppose (f_n) is equiintegrable. Fix δ_0 such that, whenever $\mu(A) < \delta_0$, $\int_A |f_n|\, d\mu < 1$. As in the proof of XVIII.2.2.1., write

$$X = A_1 \cup \cdots \cup A_r \cup B_1 \cup \cdots \cup B_m$$

for some r with $\mu(A_k) < \delta_0$ for all k, where $B_1, \ldots, B_m$ are the atoms of μ of measure $\geq \delta_0$. If β_{nj} is the a.e. constant value of f_n on B_j, and β_j the a.e. constant value of f on B_j, then $\beta_n j \to \beta_j$ for all j. Thus the β_{nj} are uniformly bounded, i.e. there is a c with $|\beta_{nj}| \leq c$ for all n, j. For each n we have

$$\int_X |f_n|\, d\mu \leq \sum_{k=1}^{r} \int_{A_k} |f_n|\, d\mu + \sum_{j=1}^{m} \int_{B_j} |f_n|\, d\mu < r + c \sum_{j=1}^{m} \mu(B_j) =: M$$

so (f_n) is L^1-bounded, and hence (f_n) is uniformly integrable by XVIII.2.2.5.. Also, since $|f_n| \to |f|$, by Fatou's Lemma () we also have $\int_X |f|\, d\mu \leq M$. Thus f is integrable.

Let $\epsilon > 0$, and fix $\delta > 0$ such that whenever $A \in \mathcal{A}$, $\mu(A) < \delta$, $\int_A |f_n|\, d\mu < \frac{\epsilon}{3}$ for all n. We may also choose δ so that $\int_A |f|\, d\mu < \frac{\epsilon}{3}$ for any such A (actually this is automatic by Fatou's Lemma, but we could just apply XVIII.2.2.1. instead). By Egorov's Theorem (), there is a $Y \in \mathcal{A}$ with $\mu(Y) < \delta$ such that $f_n \to f$ uniformly on $X \setminus Y$. Fix N so that

$$|f_n(x) - f(x)| < \frac{\epsilon}{3\mu(X)}$$

for all $x \in X \setminus Y$ and all $n \geq N$. We then have, for $n \geq N$,

$$\int_X |f_n - f|\, d\mu = \int_Y |f_n - f|\, d\mu + \int_{X\setminus Y} |f_n - f|\, d\mu \leq \int_Y (|f_n| + |f|)\, d\mu + \int_{X\setminus Y} |f_n - f|\, d\mu$$

$$= \int_Y |f_n|\, d\mu + \int_Y |f|\, d\mu + \int_{X\setminus Y} |f_n - f|\, d\mu < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3\mu(X)}\mu(X \setminus Y) \leq \epsilon\ .$$

For the last statement of the forward part, if $E \in \mathcal{A}$, for any n

$$\left|\int_E f_n\, d\mu - \int_E f\, d\mu\right| = \left|\int_E (f_n - f)\, d\mu\right| \leq \int_E |f_n - f|\, d\mu \leq \int_X |f_n - f|\, d\mu\ .$$

Conversely, suppose each f_n and f are integrable and

$$\lim_{n\to\infty} \int_X |f_n - f|\, d\mu = 0\ .$$

Let $\epsilon > 0$. Choose N such that $\int_X |f_n - f|\, d\mu < \frac{\epsilon}{2}$ for all $n \geq N$, and choose δ_0 so that $A \in \mathcal{A}$, $\mu(A) < \delta_0$ implies $\int_A |f|\, d\mu < \frac{\epsilon}{2}$ by XVIII.2.2.1.. If $A \in \mathcal{A}$, $\mu(A) < \delta_0$, then for all $n \geq N$,

$$\int_A |f_n|\, d\mu = \int_A |f_n - f + f|\, d\mu \leq \int_A |f_n - f|\, d\mu + \int_A |f|\, d\mu \leq \int_X |f_n - f|\, d\mu + \int_A |f|\, d\mu < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\ .$$

If $1 \leq n < N$, there is a δ_n such that if $A \in \mathcal{A}$, $\mu(A) < \delta_n$, then $\int_A |f_n|\, d\mu < \epsilon$. Set $\delta = \min(\delta_0, \delta_1, \ldots, \delta_{N-1})$. Then if $A \in \mathcal{A}$, $\mu(A) < \delta$, then $\int_A |f_n|\, d\mu < \epsilon$ for all n, so (f_n) is equiintegrable. 👍

XVIII.2.2.9. There is a stronger form of the converse: if $f_n \to f$ pointwise a.e., f and the f_n are integrable, and

$$\int_E f_n\, d\mu \to \int_E f\, d\mu$$

for all $E \in \mathcal{A}$, then (f_n) is uniformly integrable [Bog07, 4.5.6].

Tightness

XVIII.2.2.10. Vitali's Convergence Theorem fails in general if μ is not finite. For example, note that any uniformly bounded set of integrable functions is uniformly integrable. In $(\mathbb{R}, \mathcal{L}, \lambda)$, let $f_n = \chi_{[n,n+1]}$. Then the f_n are uniformly bounded, hence uniformly integrable, and converge pointwise to 0, but $\int_{\mathbb{R}} f_n\, d\lambda = 1 \not\to 0$. And if $f_n = \chi_{[0,n]}$, then (f_n) is uniformly bounded, hence uniformly integrable, and $f_n \to \chi_{[0,+\infty)}$, which is not integrable.

The additional ingredient needed is motivated by the following:

XVIII.2.2.11. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f an integrable extended real-valued or complex-valued function. Then for any $\epsilon > 0$ there is an $A \in \mathcal{A}$ with $\mu(A) < \infty$ and

$$\int_{X \setminus A} |f| \, d\mu < \epsilon \ .$$

PROOF: Let $\epsilon > 0$. There is a simple function s with $s \leq |f|$ and

$$\int_X s \, d\mu > \int_X |f| \, d\mu - \epsilon \ .$$

There is an $A \in \mathcal{A}$ with $\mu(A) < \infty$ and s identically 0 outside A. Then

$$\int_{X \setminus A} |f| \, d\mu = \int_{X \setminus A} (|f| - s) \, d\mu \leq \int_X (|f| - s) \, d\mu < \epsilon \ .$$



XVIII.2.2.12. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space and $\mathcal{F}$ a set of μ-integrable functions on X. Then $\mathcal{F}$ is *tight* (μ-*tight*) if for every $\epsilon > 0$ there is an $A \in \mathcal{A}$ with $\mu(A) < \infty$ and

$$\int_{X \setminus A} |f| \, d\mu < \epsilon$$

for all $f \in \mathcal{F}$. A sequence (f_n) of μ-integrable functions is tight if $\{f_n : n \in \mathbb{N}\}$ is tight.

If μ is finite, then any set of integrable functions is tight.

Vitali's Convergence Theorem, General Case

XVIII.2.2.13. THEOREM. [GENERAL VITALI CONVERGENCE THEOREM] Let $(X, \mathcal{A}, \mu)$ be a measure space, and (f_n) a sequence of μ-integrable functions converging pointwise a.e. on X to a measurable function f. If (f_n) is equiintegrable and tight, then f is μ-integrable and

$$\lim_{n \to \infty} \int_X |f_n - f| \, d\mu = 0 \ .$$

In particular,

$$\int_E f_n \, d\mu \to \int_E f \, d\mu$$

for all $E \in \mathcal{A}$.

Conversely, if $f_n \to f$ pointwise a.e., f and the f_n are integrable, and

$$\lim_{n \to \infty} \int_X |f_n - f| \, d\mu = 0$$

then (f_n) is equiintegrable and tight.

PROOF: Let $\epsilon > 0$. Fix $A \in \mathcal{A}$, $\mu(A) < \infty$, such that

$$\int_{X\setminus A} |f_n|\, d\mu < \frac{\epsilon}{3}$$

for all n. Then by Fatou's Lemma we also have $\int_{X\setminus A} |f|\, d\mu \le \epsilon$ and, in particular, f is integrable over $X \setminus A$. On A, the f_n and f satisfy the hypotheses of XVIII.2.2.8., so f is integrable on A and

$$\int_A |f_n - f|\, d\mu \to 0 \ .$$

Thus f is integrable on X. Fix N such that

$$\int_A |f_n - f|\, d\mu < \frac{\epsilon}{3}$$

for all $n \ge N$. Then, if $n \ge N$,

$$\int_X |f_n - f|\, d\mu = \int_A |f_n - f|\, d\mu + \int_{X\setminus A} |f_n - f|\, d\mu \le \int_A |f_n - f|\, d\mu + \int_{X\setminus A} |f_n|\, d\mu + \int_{X\setminus A} |f|\, d\mu < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon \ .$$

Conversely, suppose f_n, f are integrable and $\int_X |f_n - f|\, d\mu \to 0$. Let $\epsilon > 0$. Fix N such that

$$\int_X |f_n - f|\, d\mu < \frac{\epsilon}{3}$$

for all $n \ge N$. By XVIII.2.2.11. there is a $D_0 \in \mathcal{A}$ with $\mu(D_0) < \infty$ and $\int_{E_0} |f|\, d\mu < \frac{\epsilon}{3}$, where $E_0 = X \setminus D_0$. If $n \ge N$, then

$$\int_{E_0} |f_n|\, d\mu = \int_{E_0} |f_n - f + f|\, d\mu \le \int_{E_0} |f_n - f|\, d\mu + \int_{E_0} |f|\, d\mu \le \int_X |f_n - f|\, d\mu + \int_{E_0} |f|\, d\mu < \frac{\epsilon}{3} + \frac{\epsilon}{3} = \frac{2\epsilon}{3} < \epsilon \ .$$

If $1 \le n < N$, then by XVIII.2.2.11. there is a set $D_n \in \mathcal{A}$, $\mu(D_n) < \infty$, such that $\int_{E_n} |f_n|\, d\mu < \frac{2\epsilon}{3}$, where $E_n = X \setminus D_n$. Set $D = D_0 \cup D_1 \cup \cdots \cup D_{N-1}$, and $E = X \setminus D = E_0 \cap \cdots \cap E_{N-1}$. Then $\mu(D) < \infty$ and

$$\int_E |f_n|\, d\mu \le \int_{E_n} |f_n|\, d\mu < \frac{2\epsilon}{3} \text{ for } 1 \le n < N$$

$$\int_E |f_n|\, d\mu \le \int_{E_0} |f_n|\, d\mu < \frac{2\epsilon}{3} \text{ for } n \ge N \ .$$

Thus (f_n) is tight. And f and the f_n satisfy the hypotheses of the converse part of XVIII.2.2.8. on D, so there is a $\delta > 0$ such that if $A \in \mathcal{A}$, $A \subseteq D$, $\mu(A) < \delta$, then $\int_A |f_n|\, d\mu < \frac{\epsilon}{3}$. So if $A \in \mathcal{A}$, $\mu(A) < \delta$, then

$$\int_A |f_n|\, d\mu = \int_{A\cap D} |f_n|\, d\mu + \int_{A\cap E} |f_n|\, d\mu \le \int_{A\cap D} |f_n|\, d\mu + \int_E |f_n|\, d\mu < \frac{\epsilon}{3} + \frac{2\epsilon}{3} = \epsilon$$

for all n. Thus (f_n) is equiintegrable. 👍

Bounded Mean-Square Convergence Theorem

A nice application of Vitali's Convergence Theorem is to sequences which are mean-square bounded.

XVIII.2.2.14. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and f an extended real-valued or complex-valued $\mathcal{A}$-measurable function. Then f is *square-integrable* (*μ-square-integrable*) if $|f|^2$ is integrable, i.e.

$$\int_X |f|^2 \, d\mu < \infty$$

(the absolute values are needed if f is complex-valued).

A set $\mathcal{F}$ of measurable functions is *mean-square bounded* if each function in $\mathcal{F}$ is square-integrable and there is a constant K such that

$$\int_X |f|^2 \, d\lambda \leq K$$

for all $f \in \mathcal{F}$. A sequence (f_n) is mean-square bounded if $\{f_n : n \in \mathbb{N}\}$ is mean-square bounded.

XVIII.2.2.15. The term "mean-square" should properly only be used if μ is finite and f is real-valued: the mean-square of f is the average value of f^2, i.e.

$$\frac{1}{\mu(X)} \int_X f^2 \, d\mu \ .$$

The square root of this average ("root-mean-square" or RMS) is often used. If μ is a probability measure, this is just the L^2-norm of f (), and is closely related to the standard deviation ().

XVIII.2.2.16. If μ is finite, a square-integrable function on X is integrable. But this is not necessarily true if μ is not finite.

XVIII.2.2.17. THEOREM. [BOUNDED MEAN-SQUARE CONVERGENCE THEOREM] Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and (f_n) a sequence of extended real-valued or complex-valued $\mathcal{A}$-measurable functions on X converging pointwise a.e. to an $\mathcal{A}$-measurable function f. If (f_n) is mean-square bounded, then f is square-integrable, hence integrable, and

$$\lim_{n\to\infty} \int_X |f_n - f| \, d\mu = 0 \ .$$

In particular,

$$\int_E f_n \, d\mu \to \int_E f \, d\mu$$

for all $E \in \mathcal{A}$.

PROOF: By assumption, the f_n are square-integrable, hence integrable. Fix K so that

$$\int_X |f_n|^2 \, d\lambda \leq K$$

for all n. By Fatou's Lemma (),

$$\int_X |f|^2 \, d\lambda \leq K$$

so f is also square-integrable.

We show that (f_n) is uniformly integrable. Let $\epsilon > 0$. If $\lambda > 0$ and $n \in \mathbb{N}$, set

$$E_{n,\lambda} = \{x \in X : |f_n(x)| > \lambda\}$$

as before. For each n, we have

$$\int_{E_{n,\lambda}} |f_n| \, d\mu \leq \int_{E_{n,\lambda}} \frac{|f_n|^2}{\lambda} \, d\mu \leq \frac{1}{\lambda} \int_X |f_n|^2 \, d\mu \leq \frac{K}{\lambda} \ .$$

Choose λ large enough that $\frac{K}{\lambda} < \epsilon$. Thus (f_n) is uniformly integrable, so the conclusion follows from XVIII.2.2.5. and XVIII.2.2.8.. 👍

XVIII.2.2.18. This result is somewhat reminiscent of IV.2.12.10., at least in spirit. See [Edg98, 2.2.12] for a proof using the DCT instead of Vitali's Convergence Theorem.

XVIII.2.3. Exercises

XVIII.2.3.1. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $\delta > 0$. Show that X can be written in the form of the proof of XVIII.2.2.1..

XVIII.2.3.2. Let $(X, \mathcal{A}, \mu)$ be a measure space, and (f_n) a sequence of $\mathcal{A}$-measurable functions on X. If there is a nonnegative integrable function g on X such that $|f_n| \leq g$ a.e. for all n, show that (f_n) is equiintegrable and tight. [Use XVIII.2.2.1. and XVIII.2.2.11..] Thus the General Vitali Convergence Theorem implies the Dominated Convergence Theorem. Show that a stronger result holds: if $(f_n), f$ satisfy the hypotheses of the DCT, then

$$\lim_{n\to\infty} \int_X |f_n - f| \, d\mu = 0 \ .$$

XVIII.2.3.3. Let $(X, \mathcal{A}, \mu)$ be a measure space, (f_n) a sequence of μ-integrable functions on X, and f an integrable function with $\int_X |f_n - f| \, d\mu \to 0$. Show that (f_n) (which is equiintegrable and tight by XVIII.2.2.13.) is uniformly integrable. [Use an argument similar to the proof of XVIII.2.2.5..]

XVIII.3. Integration on Product Spaces

XVIII.3.1. The Fubini-Tonelli Theorems

XVIII.3.1.1. The Fubini-Tonelli Theorems concern calculation of integrals over a product space by iterated integrals, and/or interchange of order of integration in iterated integrals. There are two distinct but closely related results, one concerning integration of nonnegative measurable functions and the other concerning integration of integrable functions which are not necessarily nonnegative (in line with the development of Lebesgue integration itself, first done for nonnegative functions and then for general integrable functions).

It is conventional to call the nonnegative version "Tonelli's Theorem" and the integrable version "Fubini's Theorem" (the name "Fubini's Theorem" is often used generically for both). But the actual history is more complicated, and these names are not quite accurate historically: both FUBINI and TONELLI (who worked roughly simultaneously but independently) were partially responsible for both results. We will nonetheless follow convention and use the names "Tonelli's Theorem" and "Fubini's Theorem" since it is convenient to have distinct names for the two results, with just the warning that the names are not well justified historically.

Fubini's Theorem is one of the two most important theorems of integration theory, along with the Dominated Convergence Theorem, in the sense that these two theorems are by far the results used most often in applications.

XVIII.3.1.2. Some versions of Fubini's Theorem were previously "known" (if not proved); cf. IX.3.4.34.. But it is conventional to call all versions "Fubini's Theorem" and we will follow the convention.

Preliminaries

XVIII.3.1.3. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, $C \in \mathcal{A} \otimes \mathcal{B}$. Then C is of σ-finite measure if and only if there are $A \in \mathcal{A}$, $B \in \mathcal{B}$ of σ-finite measure, and $M \in \mathcal{A}$, $N \in \mathcal{B}$ of measure 0, such that $C \subseteq (A \times B) \cup (M \times Y) \cup (X \times N)$.

PROOF: It is obvious that if there are such A, B, M, and N, then C is σ-finite (see Exercise ().) Conversely, let $C = \cup_{n=1}^{\infty} C_n$, with $(\mu \times \nu)(C_n) < \infty$. For each n there is a sequence $R_{nk} = A_{nk} \times B_{nk}$ of rectangles with $A_{nk} \in \mathcal{A}$, $B_{nk} \in \mathcal{B}$, $C_n \subseteq \cup_k R_{nk}$, and $\sum_k \mu(A_{nk})\nu(B_{nk}) < (\mu \times \nu)(C_n) + 1 < \infty$. For each k, $\mu(A_{nk}) = \infty$ implies $\nu(B_{nk}) = 0$ and $\nu(B_{nk}) = \infty$ implies $\mu(A_{nk}) = 0$. Let M_n be the union of the A_{nk} with $\mu(A_{nk}) = 0$, and N_n the union of the B_{nk} with $\nu(B_{nk}) = 0$; renumbering, we may assume $0 < \mu(A_{nk}), \nu(A_{nk}) < \infty$ for all k and $C_n \subseteq (M_n \times Y) \cup (X \times N_n) \cup \cup_k R_{nk}$. Set $A = \cup_{n,k} A_{nk}$, $B = \cup_{n,k} B_{n,k}$, $M = \cup_n M_n$, $N = \cup_n N_n$. Then $\mu(M) = \nu(N) = 0$, A and B have σ-finite measure, and $C \subseteq (A \times B) \cup (M \times Y) \cup (X \times N)$.

XVIII.3.1.4. Note: it is not true that a set of σ-finite measure in $\mathcal{A} \otimes \mathcal{B}$ is contained in a product of sets of σ-finite measure in general. For example, if ν is not σ-finite, and $x \in X$ is not an atom, then $\{x\} \times Y$ has $\mu \times \nu$-measure zero, hence is σ-finite, but is not contained in a rectangle $A \times B$ with B of σ-finite measure.

We will use the following notation, which was partially introduced in ():

XVIII.3.1.5. DEFINITION. Let X and Y be sets.
(i) If $C \subseteq X \times Y$, for each $x \in X$ the *vertical cross section* of C at x is

$$C_x = \{y \in Y : (x, y) \in C\}$$

and, for each $y \in Y$, the *horizontal cross section* of C at Y is

$$C^y = \{x \in X : (x, y) \in C\} .$$

(ii) If $f : X \times Y \to \mathbb{R}$ (or to $\mathbb{C}$) is a function, for each $x \in X$ define $f_{[x]} : Y \to \mathbb{R}$ by

$$f_{[x]}(y) = f(x, y)$$

and for each $y \in Y$, define $f^{[y]} : X \to \mathbb{R}$ by

$$f^{[y]}(x) = f(x, y) .$$

The functions $f_{[x]}$ and $f^{[y]}$ are called the *marginal functions* of f.

Cross sections and marginal functions are well-behaved under set-theoretic and algebraic operations:

XVIII.3.1.6. PROPOSITION. Let X and Y be sets.

(i) If $C \subseteq X \times Y$, then $[(X \times Y) \setminus C]_x = Y \setminus C_x$ for all $x \in X$ and $[(X \times Y) \setminus C]^y = X \setminus C^y$ for all $y \in Y$ (informally, $(C^c)_x = (C_x)^c$ and $(C^c)^y = (C^y)^c$ for all x and y).

(ii) If $\{C_i : i \in I\}$ is a collection of subsets of $X \times Y$, then $(\cup_{i \in I} C_i)_x = \cup_{i \in I} (C_i)_x$ for all $x \in X$ and $(\cup_{i \in I} C_i)^y = \cup_{i \in I} (C_i)^y$ for all $y \in Y$.

(iii) If f and g are functions from $X \times Y$ to $\mathbb{R}$, and $\alpha \in \mathbb{R}$, then $(f + g)_{[x]} = f_{[x]} + g_{[x]}$ and $(\alpha f)_{[x]} = \alpha f_{[x]}$ for all $x \in X$ and $(f + g)^{[y]} = f^{[y]} + g^{[y]}$ and $(\alpha f)^{[y]} = \alpha f^{[y]}$ for all $y \in Y$.

(iv) If $C \subseteq X \times Y$, then $(\chi_C)_x = \chi_{C_x}$ and $(\chi_C)^y = \chi_{C^y}$ for all $x \in X$, $y \in Y$.

The straightforward proof is left to the reader.

XVIII.3.1.7. Now let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces. Form the product σ-algebra $\mathcal{A} \otimes \mathcal{B}$ () and the product measure $\mu \times \nu$ on $\mathcal{A} \otimes \mathcal{B}$. We need several technical results guaranteeing that various sets and functions arising in iterated integrals are measurable. These results might be expected, but it turns out that some σ-finite hypotheses are necessary.

The first results are that cross sections of measurable sets are measurable, and margins of measurable functions are measurable. No measures are needed for these results, and in particular no σ-finite hypotheses are needed (or even make sense!)

XVIII.3.1.8. LEMMA. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and $C \in \mathcal{A} \otimes \mathcal{B}$. Then $C_x \in \mathcal{B}$ for all $x \in X$ and $C^y \in \mathcal{A}$ for all $y \in Y$.

PROOF: Let $\mathcal{C}$ be the collection of all $C \in \mathcal{A} \otimes \mathcal{B}$ for which the conclusion holds. If $C = A \times B \in \mathcal{A} \times \mathcal{B}$, then

$$C_x = \begin{cases} B & \text{if} \quad x \in A \\ \emptyset & \text{if} \quad x \notin A \end{cases}$$

and thus $C_x \in \mathcal{B}$ for any x, and similarly $C^y \in \mathcal{A}$ for any y. Thus $C \in \mathcal{C}$, i.e. $\mathcal{A} \times \mathcal{B} \subseteq \mathcal{C}$. In particular, $\emptyset$ and $X \times Y$ are in $\mathcal{C}$. If $C \in \mathcal{C}$, then $(C^c)_x = (C_x)^c \in \mathcal{B}$ for any $x \in X$ since $\mathcal{B}$ is closed under complements, and similarly $(C^c)^y \in \mathcal{A}$ for all y, so $C^c \in \mathcal{C}$, i.e. $\mathcal{C}$ is closed under complements.

If (C_n) is a sequence in $\mathcal{C}$, then $(\cup_n C_n)_x = \cup_n C_x \in \mathcal{B}$ for all x since $\mathcal{B}$ is closed under countable unions, and similarly $(\cup_n C_n)^y \in \mathcal{A}$ for all y, so $\cup_n C_n \in \mathcal{C}$ and $\mathcal{C}$ is a σ-algebra containing $\mathcal{A} \times \mathcal{B}$, and thus $\mathcal{C}$ contains (and therefore equals) $\mathcal{A} \otimes \mathcal{B}$. 👍

XVIII.3.1.9. LEMMA. Let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces, and $f : X \times Y \to \mathbb{R}$. If f is $(\mathcal{A} \otimes \mathcal{B})$-measurable, then $f_{[x]}$ is $\mathcal{B}$-measurable for all $x \in X$ and $f^{[y]}$ is $\mathcal{A}$-measurable for all $y \in Y$.

PROOF: We bootstrap. If $f = \chi_C$ for $C \in \mathcal{A} \otimes \mathcal{B}$, then f is measurable by XVIII.3.1.6.(iv) and XVIII.3.1.8.. If f is a simple function, then f is measurable by XVIII.3.1.6.(iii). And if f is a general measurable function, there is a sequence (s_n) of simple functions converging pointwise to f (). For each x the sequence $((s_n)_{[x]})$ converges pointwise to $f_{[x]}$, and thus $f_{[x]}$ is measurable. Similarly, $f^{[y]}$ is measurable. 👍

We now need measures and a σ-finite hypothesis. The next result is the most delicate part of the proof of the Fubini-Tonelli Theorems.

XVIII.3.1.10. LEMMA. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces, and $C \in \mathcal{A} \otimes \mathcal{B}$. For $x \in X$ set $g(x) = \nu(C_x)$ and for $y \in Y$ set $h(y) = \mu(C^y)$. Then g is $\mathcal{A}$-measurable and h is $\mathcal{B}$-measurable.

PROOF: Let $\mathcal{D}$ be the set of all C in $\mathcal{A} \otimes \mathcal{B}$ for which the functions g and h are measurable. We will show that $\mathcal{D}$ equals $\mathcal{A} \otimes \mathcal{B}$ by the Monotone Class Theorem ().

First note that if $C = A \times B \in \mathcal{A} \times \mathcal{B}$, then $g = \nu(B)\chi_A$ and $h = \mu(A)\chi_B$, so g and h are measurable and $C \in \mathcal{D}$, i.e. $\mathcal{A} \times \mathcal{B} \subseteq \mathcal{D}$. Now note that $\mathcal{E}$ is the set of finite disjoint unions of sets in $\mathcal{A} \times \mathcal{B}$. If C is the disjoint union of $A_k \times B_k$, $1 \leq k \leq n$, then

$$g = \sum_{k=1}^{n} \nu(B_k)\chi_{A_k}, \quad h = \sum_{k=1}^{n} \mu(A_k)\chi_{B_k}$$

so g and h are again measurable and $C \in \mathcal{D}$, i.e. $\mathcal{E} \subseteq \mathcal{D}$.

Now suppose (C_n) is an increasing sequence in $\mathcal{D}$ with union C. Let g_n, h_n be the functions corresponding to C_n, and g and h the functions corresponding to C. For each $x \in X$, we have that $((C_n)_x)$ is an increasing sequence in $\mathcal{B}$ with union C_x, and thus $g_n(x) = \nu((C_n)_x)$ increases to $g(x) = \nu(C_x)$ by the Baby Monotone Convergence Theorem (), i.e. $g_n \to g$ pointwise. Since $C_n \in \mathcal{D}$ for all n, the g_n are measurable, hence g is

also measurable. By an almost identical argument, h is also measurable, so $C \in \mathcal{D}$, i.e. $\mathcal{D}$ is closed under increasing sequential unions.

Next we handle decreasing intersections. This is the trickiest part, and we must use the hypothesis that μ and ν are σ-finite. Let $R_m = A_m \times B_m$ be an increasing sequence of rectangles with $A_m \in \mathcal{A}$, $\cup_{m=1}^{\infty} A_m = X$, $B_m \in \mathcal{B}$, $\cup_{m=1}^{\infty} B_m = Y$, $\mu(A_m) < \infty$, $\nu(B_m) < \infty$ for all m. Set

$$\mathcal{F} = \{C \in \mathcal{A} \otimes \mathcal{B} : C \cap (A_m \times B_m) \in \mathcal{D} \text{ for all } m\} \ .$$

By arguments identical to the ones above (or by just applying them to the product space obtained by replacing X and Y by A_m and B_m), we have that $\mathcal{E} \subseteq \mathcal{F}$ and $\mathcal{F}$ is closed under increasing sequential unions. To show that $\mathcal{F}$ is closed under decreasing intersections, let (F_n) be a decreasing sequence in $\mathcal{F}$ with intersection F. Fix m, and let g_n and h_n be the functions corresponding to $Z_n := F_n \cap (A_m \times B_m)$, and g and h the functions corresponding to $Z := F \cap (A_m \times B_m)$. Then (Z_n) is a decreasing sequence in $\mathcal{D}$ with intersection Z. If $x \in X$, then $((Z_n)_x)$ is a decreasing sequence in Y with intersection Z_x, and since each $(Z_n)_x$ is contained in B_m and thus has finite ν-measure, we have $g_n(x) = \nu((Z_n)_x) \to \nu(Z_x) = g(x)$ by downward continuity (), i.e. $g_n \to g$ pointwise. Since each Z_n is in $\mathcal{D}$, g_n is measurable, so g is measurable. Similarly, h is measurable. So $Z \in \mathcal{D}$. This is true for every m, so $F \in \mathcal{F}$. Thus $\mathcal{F}$ is a monotone class containing $\mathcal{E}$, so it contains (and equals) $\sigma(\mathcal{E}) = \mathcal{A} \otimes \mathcal{B}$ by the Monotone Class Theorem.

Finally, let $C \in \mathcal{A} \otimes \mathcal{B}$. Set $C_m = C \cap (A_m \times B_m)$. Then (C_m) is an increasing sequence with union C. We have that $C_m \in \mathcal{D}$ for all m since $C \in \mathcal{F}$. Since $\mathcal{D}$ is closed under increasing sequential unions, $C \in \mathcal{D}$. Thus $\mathcal{D} = \mathcal{A} \otimes \mathcal{B}$. 👍

XVIII.3.1.11. LEMMA. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces, and $f : X \times Y \to \mathbb{R}$ a nonnegative extended real-valued $(\mathcal{A} \otimes \mathcal{B})$-measurable function. Set

$$g(x) = \int_Y f_{[x]} \, d\nu = \int_Y f(x, y) \, d\nu(y)$$

$$h(y) = \int_X f^{[y]} \, d\mu = \int_X f(x, y) \, d\mu(x) \ .$$

Then g is $\mathcal{A}$-measurable and h is $\mathcal{B}$-measurable.

PROOF: Bootstrap. If $f = \chi_C$ for $C \in \mathcal{A} \otimes \mathcal{B}$, then $f_{[x]} = \chi_{C_x}$ and $f^{[y]} = \chi_{C^y}$ for $x \in X$, $y \in Y$, so

$$g(x) = \int_Y f_{[x]} \, d\nu = \int_Y \chi_{C_x} \, d\nu = \nu(C_x)$$

$$h(y) = \int_X f^{[y]} \, d\mu = \int_X \chi_{C^y} \, d\mu = \mu(C^y)$$

so g and h are measurable by Lemma XVIII.3.1.10.. Note also that if $f_1, \ldots, f_n$ are nonnegative extended real-valued measurable functions with corresponding $g_1, \ldots, g_n$ and $h_1, \ldots, h_n$, then $f = f_1 + \cdots + f_n$ has corresponding functions $g = g_1 + \cdots + g_n$ and $h = h_1 + \cdots h_n$, so if all g_k and h_k are measurable, so are g and h. Similarly, if f has measurable g and h, and $\alpha > 0$, then αf has measurable αg and αh. In particular, if f is a simple function, then the corresponding g and h are measurable.

For a general f with corresponding functions g and h, let (s_n) be an increasing sequence of simple functions converging pointwise to f, with corresponding functions g_n and h_n. Then, for each $x \in X$, $((s_n)_{[x]})$ is an increasing sequence of $\mathcal{B}$-measurable functions converging pointwise to $f_{[x]}$, so

$$g_n(x) = \int_Y (s_n)_{[x]}\, d\nu \to \int_Y f_{[x]}\, d\nu = g(x)$$

by the Monotone Convergence Theorem, i.e. $g_n \to g$ pointwise. Each g_n is measurable, so g is measurable. Similarly, h is measurable. 👍

The Junior Tonelli Theorem

The next result is called the *Junior Tonelli Theorem* since it is Tonelli's theorem for characteristic or indicator functions. It is a measure-theoretic version of *Cavalieri's Principle*: to find the volume of a solid, integrate the cross-sectional area over the length of the solid (cf. [].)

XVIII.3.1.12. Theorem. [Junior Tonelli Theorem] Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces, and $C \in \mathcal{A} \otimes \mathcal{B}$. Then $C_x \in \mathcal{B}$ for each $x \in X$, $C^y \in \mathcal{A}$ for all $y \in A$, $g(x) = \nu(C_x)$ is μ-measurable, $h(y) = \mu(C^y)$ is ν-measurable, and

$$(\mu \times \nu)(C) = \int_X \nu(C_x)\, d\mu(x) = \int_Y \mu(C^y)\, d\nu(y)\,.$$

Proof: If $C \in \mathcal{A} \otimes \mathcal{B}$, the first four assertions hold by Lemmas XVIII.3.1.8. and XVIII.3.1.10., so the integrals make sense. For $C \in \mathcal{A} \otimes \mathcal{B}$, define

$$\theta(C) = \int_X \nu(C_x)\, d\mu = \int_X g\, d\mu\ .$$

We show θ is a measure. Obviously $\theta(\emptyset) = 0$. If (C_n) is a sequence of pairwise disjoint sets in $\mathcal{A} \otimes \mathcal{B}$ with union C, then for each $x \in X$, $((C_n)_x)$ is a sequence of pairwise disjoint sets in $\mathcal{B}$, so we have

$$\theta(C) = \theta\left(\bigcup_{n=1}^{\infty} C_n\right) = \int_X \nu\left(\left[\bigcup_{n=1}^{\infty} C_n\right]_x\right) d\mu = \int_X \nu\left(\bigcup_{n=1}^{\infty} (C_n)_x\right) d\mu = \int_X \left[\sum_{n=1}^{\infty} \nu((C_n)_x)\right] d\mu$$

$$= \sum_{n=1}^{\infty} \left[\int_X \nu((C_n)_x)\, d\mu\right] = \sum_{n=1}^{\infty} \theta(C_n)$$

where the interchange of integral and sum is justified by the Monotone Convergence Theorem. So θ is a measure. If $A \times B \in \mathcal{A} \times \mathcal{B}$, then

$$\theta(A \times B) = \int_X \nu((A \times B)_x)\, d\mu = \int_X \nu(B)\chi_A\, d\mu = \mu(A)\nu(B) = (\mu \times \nu)(A \times B)\ .$$

Thus by the uniqueness part of the Extension Theorem in the σ-finite case, $\theta = \mu \times \nu$ on $\mathcal{A} \otimes \mathcal{B}$, and the first equality is proved. The proof that $(\mu \times \nu)(C) = \int_Y \mu(C^y)\, d\nu(y)$ for $C \in \mathcal{A} \otimes \mathcal{B}$ is analogous. 👍

Tonelli's Theorem

Tonelli's Theorem concerns calculating the integral of a nonnegative measurable function over a product space by iterated integrals, and/or interchanging the order of integration in iterated integrals of nonnegative measurable functions.

XVIII.3.1.13. THEOREM. [TONELLI'S THEOREM] Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces, and f a nonnegative extended real-valued $(\mathcal{A} \otimes \mathcal{B})$-measurable function on $X \times Y$, with marginal functions $f_{[x]}$ and $f^{[y]}$. Then

(i) $f_{[x]}$ is $\mathcal{B}$-measurable for all $x \in X$.

(i′) $f^{[y]}$ is $\mathcal{A}$-measurable for all $y \in Y$.

(ii) g defined by $g(x) = \int_Y f_{[x]}\, d\nu$ is $\mathcal{A}$-measurable.

(ii′) h defined by $h(y) = \int_X f^{[y]}\, d\mu$ is $\mathcal{B}$-measurable.

(iii) We have

$$\int_{X\times Y} f\, d(\mu\times\nu) = \int_X g\, d\mu = \int_X \left[\int_Y f_{[x]}\, d\nu\right] d\mu = \int_X \left[\int_Y f(x,y)\, d\nu(y)\right] d\mu(x)$$
$$= \int_Y h\, d\nu = \int_Y \left[\int_X f^{[y]}\, d\mu\right] d\nu = \int_Y \left[\int_X f(x,y)\, d\mu(x)\right] d\nu(y)\ .$$

In particular, the iterated integrals

$$\int_X \left[\int_Y f(x,y)\, d\nu(y)\right] d\mu(x) = \int_Y \left[\int_X f(x,y)\, d\mu(x)\right] d\nu(y)$$

exist and are equal.

PROOF: We have (i) and (i′) by Lemma XVIII.3.1.9., and (ii) and (ii′) by Lemma XVIII.3.1.11.. For (iii), we bootstrap. If $f = \chi_C$, $C \in \mathcal{A} \otimes \mathcal{B}$, the equalities hold by the Junior Tonelli Theorem XVIII.3.1.12.. It is also clear from the (nonnegative) linearity of integrals that if the equalities hold for $f_1, \ldots, f_n$, and if $\alpha_1, \ldots, \alpha_n \geq 0$, then the equalities hold for $\alpha_1 f_1 + \cdots + \alpha_n f_n$. In particular, they hold if f is a simple function.

Now let f be a general nonnegative $(\mathcal{A}\otimes\mathcal{B})$-measurable function. Let (s_n) be a sequence of simple functions increasing pointwise to f. Then $(s_n)_{[x]}$ increases pointwise to $f_{[x]}$ for all $x \in X$, so by the Monotone Convergence Theorem we have

$$g_n(x) := \int_Y (s_n)_{[x]}\, d\nu \to \int_Y f_{[x]}\, d\nu = g(x)$$

for all $x \in X$, i.e. g_n increases pointwise to g. Thus, by two more applications of the MCT,

$$\int_{X\times Y} f\, d(\mu\times\nu) = \lim_{n\to\infty}\left[\int_{X\times Y} s_n\, d(\mu\times\nu)\right] = \lim_{n\to\infty}\left[\int_Y g_n\, d\mu\right] = \int_Y g\, d\mu$$

and the first equality is proved for f. The proof of the second equality is analogous. 👍

Fubini's Theorem

Fubini's Theorem is analogous to Tonelli's Theorem, but instead of nonnegative measurable functions it deals with integrable functions which are not necessarily nonnegative.

XVIII.3.1.14. THEOREM. [FUBINI'S THEOREM] Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces, and f a complex-valued $(\mathcal{A} \otimes \mathcal{B})$-measurable function with marginal functions $f_{[x]}$ and $f^{[y]}$ for $x \in X$, $y \in Y$. Assume at least one of the following integrals exists and is finite:

$$\int_{X \times Y} |f| \, d(\mu \times \nu), \quad \int_Y \left[\int_X |f(x,y)| \, d\mu(x) \right] d\nu(y), \quad \int_X \left[\int_Y |f(x,y)| \, d\nu(y) \right] d\mu(x) \ .$$

Then all three integrals exist and are finite and equal, hence f is $(\mu \times \nu)$-integrable, and

(i) $f_{[x]}$ is ν-integrable for μ-almost all $x \in X$.

(i′) $f^{[y]}$ is μ-integrable for ν-almost all $y \in Y$.

(ii) g defined by $g(x) = \int_Y f_{[x]} \, d\nu$ (which is defined μ-a.e.) is equal μ-a.e. to a μ-integrable function.

(ii′) h defined by $h(y) = \int_X f^{[y]} \, d\mu$ (which is defined ν-a.e.) is equal ν-a.e. to a ν-integrable function.

(iii) We have

$$\int_{X \times Y} f \, d(\mu \times \nu) = \int_X g \, d\mu = \int_X \left[\int_Y f_{[x]} \, d\nu \right] d\mu = \int_X \left[\int_Y f(x,y) \, d\nu(y) \right] d\mu(x)$$

$$= \int_Y h \, d\nu = \int_Y \left[\int_X f^{[y]} \, d\mu \right] d\nu = \int_Y \left[\int_X f(x,y) \, d\mu(x) \right] d\nu(y) \ .$$

In particular, the iterated integrals

$$\int_X \left[\int_Y f(x,y) \, d\nu(y) \right] d\mu(x) = \int_Y \left[\int_X f(x,y) \, d\mu(x) \right] d\nu(y)$$

exist and are equal.

PROOF: We prove the theorem for f real-valued; the complex-valued case follows by taking real and imaginary parts.

The first integral in the hypotheses automatically exists, and by Tonelli's Theorem the other two also exist and equal the first one; thus if one is finite, they all are, and are equal. For each $x \in X$, we have that $f_{[x]}$ is $\mathcal{B}$-measurable by Lemma XVIII.3.1.9., and $|f_{[x]}| = |f|_{[x]}$; and since

$$\int_{X \times Y} |f| \, d(\mu \times \nu) = \int_Y \left[\int_X |f|_{[x]} \, d\nu \right] d\mu = \int_Y \left[\int_X |f_{[x]}| \, d\nu \right] d\mu < \infty$$

it follows that

$$\int_X |f_{[x]}| \, d\nu < \infty$$

for almost all x (), i.e. $f_{[x]}$ is ν-integrable for almost all $x \in X$. Similarly, $f^{[y]}$ is μ-integrable for almost all $y \in Y$.

Let f_+ and f_- be the positive and negative parts of f; then $f_+, f_- \leq |f|$, so are $(\mu \times \nu)$-integrable. We have $(f_+)_{[x]} = (f_{[x]})_+$ and $(f_-)_{[x]} = (f_{[x]})_-$ for all $x \in X$, and if

$$g_+(x) = \int_Y (f_+)_{[x]}\, d\nu = \int_Y (f_{[x]})_+\, d\nu$$

$$g_-(x) = \int_Y (f_-)_{[x]}\, d\nu = \int_Y (f_{[x]})_-\, d\nu$$

we have that

$$g_+(x) \leq \bar{g}(x) := \int_Y |f_{[x]}|\, d\nu = \int_Y |f|_{[x]}\, d\nu$$

and since

$$\int_X \bar{g}\, d\mu = \int_{X\times Y} |f|\, d(\mu \times \nu) < \infty$$

by Tonelli's Theorem, g_+ is μ-integrable. Similarly, g_- is μ-integrable. In particular, g_+ and g_- are finite μ-a.e., and if they are both finite at x, then $f_{[x]} = (f_{[x]})_+ - (f_{[x]})_-$ and

$$g(x) = \int_Y f_{[x]}\, d\mu = \int_Y [(f_{[x]})_+ - (f_{[x]})_-]\, d\mu = \int_Y (f_{[x]})_+\, d\mu - \int_Y (f_{[x]})_-\, d\mu = g_+(x) - g_-(x)\ .$$

Thus g is defined μ-a.e. and is equal μ-a.e. to a μ-integrable function (note though that g_+ and g_- are not necessarily the positive and negative parts of g). We also have

$$\int_X g\, d\mu = \int_X g_+\, d\mu - \int_X g_-\, d\mu = \int_X \left[\int_Y (f_+)_{[x]}\, d\nu\right]\, d\mu - \int_X \left[\int_Y (f_-)_{[x]}\, d\nu\right]\, d\mu$$

$$= \int_{X\times Y} f_+\, d(\mu \times \nu) - \int_{X\times Y} f_-\, d(\mu \times \nu) = \int_{X\times Y} f\, d(\mu \times \nu)$$

by Tonelli's Theorem applied to f_+ and f_-. The proof of the second equality is analogous. 👍

Examples where σ-finite hypothesis is necessary

The Complete Case

Extension to Sets of σ-Finite Measure

Fubini-Tonelli

We have the following corollary of the Fubini-Tonelli Theorem, which can be used to justify the usual iterated integrals encountered in vector calculus problems:

XVIII.3.1.15. COROLLARY. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, $C \in \mathcal{A} \otimes \mathcal{B}$ of σ-finite measure, and $f : C \to \mathbb{R}$ a function which is either nonnegative and $(\mu \times \nu)$-measurable, or integrable over C (i.e. $f\chi_C$ is integrable over $X \times Y$.) Then

(i) The function $g(x) = \int_{C_x} f(x,y)\, d\nu(y)$ is $\mathcal{A}$-measurable, and is μ-integrable for almost all x if f is $(\mu \times \nu)$-integrable over C. *****FIX

(ii) The function $h(y) = \int_{C^y} f(x,y)\, d\mu(x)$ is $\mathcal{B}$-measurable, and is ν-integrable for almost all x if f is $(\mu \times \nu)$-integrable over C. *****FIX

(iii) We have that

$$\int_C f\, d(\mu \times \nu) = \int_X \left[\int_{C_x} f(x,y)\, d\nu(y) \right] d\mu(x) = \int_Y \left[\int_{C^y} f(x,y)\, d\mu(x) \right] d\nu(y)\,.$$

PROOF: Let $A, M \in \mathcal{A}$, $B, N \in \mathcal{B}$ be as in the proposition. Then

$$\int_C f\, d(\mu \times \nu) = \int_{X \times Y} f\chi_C\, d(\mu \times \nu) = \int_{A \times B} f\chi_C\, d(\mu \times \nu)$$

(since $f\chi_C = f\chi_C\chi_{A\times B}$ a.e.) Then, by the Fubini-Tonelli Theorem applied to $f\chi_C$ on $A \times B$, $g_0(x) = \int_B f(x,y)\chi_C(x,y)\, d\nu(y) = \int_{B\cap C_x} f(x,y)\, d\nu(y)$ is $\mathcal{A}$-measurable, and integrable for almost all x if f is integrable. Since $C_x \setminus B \subseteq N$, we have $g_0(x) = \int_{C_x} f(x,y)\, d\nu(y)$, i.e. $g(x) = g_0(x)$ for $x \in A$. Also, $g(x) = 0$ for $x \notin A \cup M$ since $C_x \subseteq N$, so $g(x) = 0$ a.e. on $X \setminus A$. Then, by Fubini-Tonelli on $A \times B$,

$$\int_{A\times B} f\chi_C\, d(\mu \times \nu) = \int_A \left[\int_B f(x,y)\chi_C(x,y)\, d\nu(y) \right] d\mu(x)$$

$$= \int_A \left[\int_B f(x,y)\chi_{C_x}(x,y)\, d\nu(y) \right] d\mu(x) = \int_A \left[\int_{C_x} f(x,y)\, d\nu(y) \right] d\mu(x)$$

$$= \int_A \left[\int_{C_x} f(x,y)\, d\nu(y) \right] d\mu(x) + \int_M \left[\int_{C_x} f(x,y)\, d\nu(y) \right] d\mu(x) +$$

$$+ \int_{X\setminus(A\cup M)} \left[\int_{C_x} f(x,y)\, d\nu(y) \right] d\mu(x) = \int_X \left[\int_{C_x} f(x,y)\, d\nu(y) \right] d\mu(x)$$

since $\mu(M) = 0$ and $\nu(C_x) = 0$ for $x \notin A \cup M$. The proof for the iterated integral in the opposite order is nearly identical. 👍

XVIII.3.1.16. As a sample application, suppose C is a region in $\mathbb{R}^2$ of the form

$$\{(x,y) : g(x) \le y \le h(x), a \le x \le b\}$$

where g, h are continuous functions with $g(x) \le h(x)$ for $a \le x \le b$, and suppose f is a Lebesgue measurable function from C to $\mathbb{R}$ which is either nonnegative or integrable over C (e.g. a continuous function.) Then $C_x = [g(x), h(x)]$ for each $x \in [a,b]$, and so

$$\int_C f\, d\lambda^{(2)} = \int_{[a,b]} \left[\int_{[g(x),h(x)]} f(x,y)\, d\lambda(y) \right] d\lambda(x) = \int_a^b \left[\int_{g(x)}^{h(x)} f(x,y)\, dy \right] dx\,.$$

XVIII.3.2. Lebesgue Integration on $\mathbb{R}^n$

XVIII.3.3. Exercises

XVIII.3.3.1. Another Nonmeasurable Set. [?] Assume the Continuum Hypothesis, and let $\precsim$ be a well-ordering on $[0,1]$ according to the first uncountable ordinal, i.e. such that the set of predecessors of any element is countable. Set

$$E = \{(x,y) \in [0,1]^2 : x \precsim y\} \ .$$

(a) For fixed $y \in [0,1]$, the set $E^y = \{x \in [0,1] : x \precsim y\}$ is countable, hence Lebesgue measurable, and $\lambda(E^y) = 0$. Thus, if E is Lebesgue measurable, by the Junior Tonelli Theorem (XVIII.3.1.12.)

$$\lambda^{(2)}(E) = \int_0^1 \lambda(E^y)\, d\lambda(y) = 0 \ .$$

(b) For fixed $x \in [0,1]$, the set $E_x = \{y \in [0,1] : x \precsim y\}$ is cocountable, hence Lebesgue measurable, and $\lambda(E_x) = 1$. Thus, if E is Lebesgue measurable, by the Junior Tonelli Theorem

$$\lambda^{(2)}(E) = \int_0^1 \lambda(E_x)\, d\lambda(x) = 1 \ .$$

(c) Conclude that E is not a Lebesgue measurable subset of $[0,1]^2$.

This suggests why the CH cannot be proved in ZF: there can be no "measurable" bijection from $\mathbb{R}$ to ω_1.

XVIII.3.3.2. [?] Let X be the set of ordinals less than the first uncountable ordinal, $\mathcal{A}$ the σ-algebra of countable-cocountable subsets of X, and μ the measure on $(X,\mathcal{A})$ with $\mu(A) = 0$ if A is countable and $\mu(A) = 1$ if A is cocountable. Let $E = \{(\alpha,\beta) : \alpha \leq \beta\} \subseteq X \times X$.

(a) Show that the iterated integrals

$$\int_X \int_X \chi_E(\alpha,\beta)\, d\mu(\alpha)\, d\mu(\beta) \text{ and } \int_X \int_X \chi_E(\alpha,\beta)\, d\mu(\beta)\, d\mu(\alpha)$$

both exist but are not equal.

(b) Conclude that E is not $(\mathcal{A}\bar{\otimes}\mathcal{A})$-measurable, i.e. χ_E is not $\overline{(\mu \times \mu)}$-integrable. (It is not hard to see directly that $E \notin \mathcal{A} \otimes \mathcal{A}$.)

This is a variation of XVIII.3.3.1. which does not use the Continuum Hypothesis or the AC.

XVIII.3.3.3. (a) Let $f : \mathbb{R} \to \mathbb{R}$ be a function such that the graph $\Gamma(f)$ is a Lebesgue measurable subset of $\mathbb{R}^2$. Show that $\lambda^{(2)}(\Gamma(f)) = 0$. [Use the Junior Tonelli Theorem ().]

(b) Characterize the functions whose graphs are Lebesgue measurable.

(c) Generalize to functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, and between more general measure spaces.

Note that if f is continuous, there is an elementary proof (IX.3.2.17.). Compare also with **??**.

XVIII.3.3.4. Fubini's Theorem for Signed and Complex Measures. (a) Let μ and ν be finite signed measures on measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ respectively. If $\mu \times \nu$ is defined as in XVII.9.5.7., show that the statement of Fubini's Theorem (XVIII.3.1.14.) remains valid if "σ-finite" is replaced by "finite signed."

(b) Let μ and ν be complex measures on measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ respectively. If $\mu \times \nu$ is defined as in XVII.9.5.8., show that the statement of Fubini's Theorem remains valid if "σ-finite" is replaced by "complex."

XVIII.4. The Henstock-Kurzweil Integral

The Henstock-Kurzweil (HK) integral on $\mathbb{R}$ is a beautiful and rather simple extension of the Riemann integral. The only modification which needs to be made is to consider a more general and flexible notion of "fineness" of a partition than its mesh. Once some preliminary technicalities are handled, the theory proceeds essentially identically to the Riemann integral case. One ends up with an integral which not only extends the Riemann integral, but even (slightly) the Lebesgue integral, with nice properties such as an unrestricted Fundamental Theorem of Calculus.

XVIII.4.1. Definition and Basic Properties

Recall the definition of gauges and fine partitions ().

XVIII.4.1.1. DEFINITION. Let $[a,b]$ be a closed bounded interval and $f : [a,b] \to \mathbb{R}$ a function. Then f is *HK-integrable* on $[a,b]$, with integral I, if for every $\epsilon > 0$ there is a gauge δ on I such that $|\mathcal{R}(f,\dot{\mathcal{P}}) - I| < \epsilon$ for every δ-fine partition $\dot{\mathcal{P}}$ of $[a,b]$.

It is obvious that the integral I is unique if it exists.

We first observe that the HK-integral extends the Riemann integral:

XVIII.4.1.2. PROPOSITION. Let $f : [a,b] \to \mathbb{R}$ be Riemann integrable on $[a,b]$. Then f is HK-integrable on $[a,b]$, with integral $I = \displaystyle\int_a^b f(t)\,dt$.

PROOF: Let $\epsilon > 0$, and choose $\delta > 0$ such that whenever $\dot{\mathcal{P}}$ is a tagged partition with $\|\mathcal{P}\| < \delta$, we have $|\mathcal{R}(f,\dot{\mathcal{P}}) - I| < \delta$. Then the constant gauge δ works in the definition of HK-integration. 👍

XVIII.4.1.3. Thus we may justifiably use the notation $\displaystyle\int_a^b f(t)\,dt$ for the HK-integral of an HK-integrable function.

Conversely, if we can always choose a constant gauge for any ϵ, then f is Riemann integrable. Since there are many nonconstant gauges with varying properties which can be used in the definition of HK-integration, it is much easier to satisfy the HK definition for a given f and one might expect that many more functions are HK-integrable. The next result illustrates this dramatically:

XVIII.4.1.4. THEOREM. Let $[a,b]$ be a closed bounded interval, and A a Lebesgue measurable subset of $[a,b]$. Then χ_A is HK-integrable over $[a,b]$ with integral $\lambda(A)$.

PROOF: Set $B = [a,b] \setminus A$. Let $\epsilon > 0$. There is a sequence (I_n) of open intervals such that $A \subseteq \cup_n I_n$ and $\sum_n \ell(I_n) < \lambda(A) + \epsilon$, and a sequence (J_n) of open intervals such that $B \subseteq \cup_n J_n$ and $\sum_n \ell(J_n) < \lambda(B) + \epsilon$. If $t \in A$, choose an n (say, the smallest n) for which $t \in I_n$, and set $\delta(t)$ equal to the distance from t to I_n^c, and similarly if $t \in B$ choose an n for which $t \in J_n$ and set $\delta(t)$ equal to the distance from t to J_n^c. Then δ is a gauge on $[a,b]$.

Let $\dot{\mathcal{P}} = \{a = x_0, \ldots, x_m = b; t_1, \ldots, t_m\}$ be a δ-fine partition of $[a, b]$. Then each t_k is either in A or B. Let N_1 be the set of k for which $t_k \in A$, and N_2 the set of k for which $t_k \in B$. For each $k \in N_1$, the interval (x_{k-1}, x_k) is contained in I_n for some n; thus we have

$$\sum_{k \in N_1} \ell([x_{k-1}, x_k]) \leq \sum_n \ell(I_n) < \lambda(A) + \epsilon \ .$$

Similarly, we have

$$\sum_{k \in N_2} \ell([x_{k-1}, x_k]) \leq \sum_n \ell(J_n) < \lambda(B) + \epsilon = (b - a) - [\lambda(A) - \epsilon] \ .$$

Since

$$\sum_{k \in N_1} \ell([x_{k-1}, x_k]) + \sum_{k \in N_2} \ell([x_{k-1}, x_k]) = \sum_{k=1}^{m} \ell([x_{k-1}, x_k]) = b - a$$

we have

$$\lambda(A) - \epsilon < \mathcal{R}(\chi_A, \dot{\mathcal{P}}) = \sum_{k \in N_1} \ell([x_{k-1}, x_k]) < \lambda(A) + \epsilon \ .$$



XVIII.4.1.5. Unbounded functions cannot be Riemann integrable, but they can be HK-integrable because gauges can be made to force certain points to be tags. See XVIII.4.2.4. for examples.

XVIII.4.2. The Fundamental Theorem of Calculus

The following version of the Fundamental Theorem of Calculus is one of the most attractive features of HK-integration. It fails for Riemann integration () and even for Lebesgue integration (). The proof is quite simple but uses a clever choice of gauge.

XVIII.4.2.1. Theorem. [Fundamental Theorem of Calculus, HK Version] Let $[a, b]$ be a closed bounded interval, and f a differentiable function on $[a, b]$. Then f' is HK-integrable on $[a, b]$ and

$$\int_a^b f'(t)\, dt = f(b) - f(a) \ .$$

Proof: Let $\epsilon > 0$. For each $t \in [a, b]$ there is a $\delta(t) > 0$ such that

$$\left| \frac{f(x) - f(t)}{x - t} - f'(t) \right| < \frac{\epsilon}{b - a}$$

whenever $x \in [a, b]$, $0 < |x - t| < \delta(t)$. Then δ is a gauge on $[a, b]$.
Let $\dot{\mathcal{P}} = \{a = x_0, \ldots, x_n = b; t_1, \ldots, t_n\}$ be a δ-fine partition of $[a, b]$. For each k, since $x_{k-1} \leq t_k \leq x_k$ we have

$$|f(x_k) - f(x_{k-1}) - f'(t_k)(x_k - x_{k-1})| = |f(x_k) - f(t_k) - f'(t_k)(x_k - t_k) + f(t_k) - f(x_{k-1}) - f'(t_k)(t_k - x_{k-1})|$$

$$\leq |f(x_k) - f(t_k) - f'(t_k)(x_k - t_k)| + |f(t_k) - f(x_{k-1}) - f'(t_k)(t_k - x_{k-1})|$$
$$< \frac{\epsilon}{b-a}(x_k - t_k) + \frac{\epsilon}{b-a}(t_k - x_{k-1}) = \frac{\epsilon}{b-a}(x_k - x_{k-1})$$

since $x_k - t_k, t_k - x_{k-1} < \delta(t_k)$, and thus

$$|f(b) - f(a) - \mathcal{R}(f', \dot{\mathcal{P}})| = \left|\sum_{k=1}^{n}[f(x_k) - f(x_{k-1}) - f'(t_k)(x_k - x_{k-1})]\right|$$

$$\leq \sum_{k=1}^{n} |f(x_k) - f(x_{k-1}) - f'(t_k)(x_k - x_{k-1})| \leq \sum_{k=1}^{n} \frac{\epsilon}{b-a}(x_k - x_{k-1}) = \frac{\epsilon}{b-a}(b-a) = \epsilon \ .$$



We can extend this result to the case where f is not quite differentiable everywhere:

XVIII.4.2.2. THEOREM. [FUNDAMENTAL THEOREM OF CALCULUS, GENERAL HK VERSION] Let $[a, b]$ be a closed bounded interval, and f a continuous function on $[a, b]$ which is differentiable except at countably many points. If g is a function on $[a, b]$ which equals f' a.e. on $[a, b]$, then g is HK-integrable on $[a, b]$ and

$$\int_a^b g(t)\, dt = f(b) - f(a) \ .$$

XVIII.4.2.3. In XVIII.4.2.2., we cannot relax the hypothesis that f be differentiable except at countably many points to simply require that f be differentiable almost everywhere. If g is a nonnegative measurable function on $[a, b]$ which is not Lebesgue integrable on $[a, b]$, hence not HK-integrable on $[a, b]$ (), there is a continuous and a.e. differentiable function f on $[a, b]$ with $f' = g$ a.e. (XIX.5.2.26.).



XVIII.4.2.4. EXAMPLES. (i) Let $f(x) = x^2 \sin \frac{1}{x^2}$ (with $f(0) = 0$). Then f is differentiable everywhere, and

$$f'(x) = 2x \sin \frac{1}{x^2} - \frac{2}{x} \cos \frac{1}{x^2}$$

($f'(0) = 0$). But f' is not Lebesgue integrable on any interval containing 0; however, it is HK-integrable on any (closed bounded) interval. This is an example of an unbounded function which is HK-integrable.

(ii) For a simpler example, $f(x) = 2\sqrt{x}$ is continuous on $[0, 1]$ and differentiable except at 0. Then $f'(x) = x^{-1/2}$ is HK-integrable on $[0, 1]$ even though it is unbounded. We have

$$\int_0^1 t^{-1/2}\, dt = 2\sqrt{1} - 2\sqrt{0} = 2$$

i.e. the HK-integral equals the improper Riemann integral. This holds generally ().

XVIII.5. Convolution

The operation of convolution of functions is valuable in a great many contexts. It is a "multiplication" of functions quite different from ordinary pointwise product, which bears some resemblance to matrix multiplication.

An informal motivation for the operation of convolution comes from multiplying series. We give a completely heuristic version which ignores all questions of convergence. Suppose

$$\sum_{k=-\infty}^{\infty} a_k x^k \quad \text{and} \quad \sum_{k=-\infty}^{\infty} b_k x^k$$

are two-way infinite summations. We want to "multiply" these to get a new two-way infinite summation

$$\left[\sum_{k=-\infty}^{\infty} a_k x^k\right]\left[\sum_{k=-\infty}^{\infty} b_k x^k\right] = \sum_{k=-\infty}^{\infty} c_k x^k$$

and the "formula" for c_k is obtained by formally multiplying the two expressions and collecting together like powers of x to obtain

$$c_k = \sum_{i+j=k} a_j b_i = \sum_{j=-\infty}^{\infty} a_j b_{k-j} \ .$$

Such a computation is meaningless as done informally here, but with suitable tweaking and interpretation becomes meaningful and useful.

We will not trace the full convoluted history of the development of this subject here.

XVIII.5.1. Convolution of Functions on $\mathbb{R}$, Continuous Case

In this subsection, we will discuss properties of convolution which do not require a knowledge of Lebesgue integration; all integrals in this subsection will be Riemann integrals or improper Riemann integrals. We will consider only continuous functions, and mostly only functions with *compact support*, i.e. functions which vanish (are identically zero) outside a bounded interval.

XVIII.5.1.1. Definition. Let f and g be continuous functions from $\mathbb{R}$ to $\mathbb{R}$. Suppose for every $x \in \mathbb{R}$ the integral

$$\int_{-\infty}^{+\infty} f(y)g(x-y)\,dy = \int_{\mathbb{R}} f(y)g(x-y)\,dy$$

exists as an improper Riemann integral and is finite. The function $f * g : \mathbb{R} \to \mathbb{R}$ defined by

$$(f*g)(x) = \int_{-\infty}^{+\infty} f(y)g(x-y)\,dy$$

is called the *convolution* of f and g.

XVIII.5.1.2. PROPOSITION. Let f and g be continuous functions on $\mathbb{R}$. If $f * g$ is defined, then $g * f$ is also defined and $g * f = f * g$.

PROOF: Fix $x \in \mathbb{R}$. Then

$$(f * g)(x) = \int_{-\infty}^{+\infty} f(y)g(x-y)\,dy = \lim_{a\to-\infty, b\to+\infty} \int_a^b f(y)g(x-y)\,dy\ .$$

Make the change of variable $z = x - y$. Then, for $a, b \in \mathbb{R}$,

$$\int_a^b f(y)g(x-y)\,dy = \int_{x-a}^{x-b} f(x-z)g(z)\,(-1)dz = \int_{x-b}^{x-a} f(x-z)g(z)\,dz$$

and since $c = x - b \to -\infty$ and $d = x - a \to +\infty$ as $a \to -\infty$ and $b \to +\infty$, we have that

$$\int_{-\infty}^{+\infty} g(z)f(x-z)\,dz = \lim_{c\to-\infty, d\to+\infty} \int_c^d g(z)f(x-z)\,dz = \lim_{a\to-\infty, b\to+\infty} \int_a^b f(y)g(x-y)\,dy$$

exists as an improper Riemann integral and equals $(f * g)(x)$. 👍

XVIII.5.1.3. One can be fooled into thinking that $g * f = -f * g$ since the substitution $z = x - y$ introduces a minus sign into the integrand. But the direction of integration is also reversed, canceling the negation. It is more straightforward to use the change-of-variables formula (); cf. (). To see that $g * f$ must be $+f * g$ and not $-f * g$, consider the case where f and g are nonnegative; then $f * g$ and $g * f$ are also nonnegative.

The Compact Support Case

XVIII.5.1.4. Suppose f and g are continuous functions, and f has compact support. Then $f * g$ is defined. For if f vanishes outside $[a, b]$, then for each x we have

$$(f * g)(x) = \int_{-\infty}^{+\infty} f(y)g(x-y)\,dy = \int_a^b f(y)g(x-y)\,dy$$

as an ordinary Riemann integral. The convolution $g * f$ is also defined by XVIII.5.1.2.. So for existence of the convolution it suffices that one of the functions have compact support.

Continuity of the Convolution

To discuss continuity of $f * g$, we need the following fact:

XVIII.5.1.5. PROPOSITION. Let f be a continuous function on $\mathbb{R}$. For $x \in \mathbb{R}$, define $f_x(y) = f(x + y)$. If $x_n \to x$ in $\mathbb{R}$, then $f_{x_n} \to f_x$ u.c.

PROOF: Let $\epsilon > 0$, and $[a, b]$ a bounded interval in $\mathbb{R}$. Since f is uniformly continuous on bounded intervals (), there is a $\delta > 0$ such that if $y, z \in [a + x - 1, b + x + 1]$ and $|y - z| < \delta$, then $|f(y) - f(z)| < \epsilon$. Fix N such that $|x_n - x| < \min(\delta, 1)$ for all $n \geq N$. Then, if $y \in [a, b]$ and $n \geq N$, we have

$$|f_{x_n}(y) - f_x(y)| = |f(x_n + y) - f(x + y)| < \epsilon\ .$$

👍

XVIII.5.1.6. PROPOSITION. Let f and g be continuous functions, with f of compact support. Then $f * g$ is continuous.

PROOF: Let $x_n \to x$ in $\mathbb{R}$. Set $h_n(y) = f(y)g(x_n - y) = f(y)g_{x_n}(-y)$ and $h(y) = f(y)g(x-y) = f(y)g_x(-y)$. Since $g_{x_n} \to g_x$ u.c. by XVIII.5.1.5. and f has compact support, $h_n \to h$ uniformly on $\mathbb{R}$; and the h_n and h all vanish outside a finite interval $[a, b]$ (the support of f). Thus

$$(f * g)(x_n) = \int_{-\infty}^{+\infty} f(y)g(x_n - y)\,dy = \int_a^b h_n(y)\,dy \to \int_a^b h(y)\,dy = \int_{-\infty}^{+\infty} f(y)g(x-y)\,dy = (f * g)(x)$$

so $f * g$ is continuous by the Sequential Criterion. 👍

XVIII.5.1.7. PROPOSITION. Let f and g be continuous functions of compact support contained in $[a, b]$ and $[c, d]$ respectively. Then $f * g$ has compact support contained in $[a + c, b + d]$.

PROOF: If $x < a + c$, then for every y either $y < a$ or $x - y < c$, so $f(y)g(x - y) = 0$ for all y, and similarly for $x > b + d$. 👍

Differentiability of the Convolution

The convolution of two functions f and g has the pleasant property of being least as smooth as the smoother of f and g. We next show the continuous version of this principle.

XVIII.5.1.8. THEOREM. Let f be a continuous function and g a $\mathcal{C}^r$ function on $\mathbb{R}$ $(0 \le r \le \infty)$. Suppose either f or g has compact support. Then $f * g$ is $\mathcal{C}^r$, and $(f * g)^{(k)} = f * (g^{(k)})$ for $1 \le k \le r$.

PROOF: The case $r = 0$ is XVIII.5.1.6.. Suppose g is $\mathcal{C}^1$. Then the support of g' is contained in the support of g, so $f * (g')$ is defined.

Set $H(x, y) = f(y)g(x - y)$. Then $\frac{\partial H}{\partial x}(x, y) = f(y)g'(x - y)$, which is continuous by assumption.

If f has compact support contained in $[a, b]$, then by () we have

$$(f * g)'(x) = \frac{d}{dx}\left(\int_{-\infty}^{+\infty} f(y)g(x-y)\,dy\right)(x) = \frac{d}{dx}\left(\int_a^b H(x, y)\,dy\right)(x)$$

$$= \int_a^b \frac{\partial H}{\partial x}(x, y)\,dy = \int_{-\infty}^{+\infty} f(y)g'(x-y)\,dy = (f * (g'))(x)\ .$$

The argument if g has compact support is similar, but slightly more complicated. Suppose the support of g is contained in $[a, b]$. Fix x_0, and fix an interval $[c, d]$ with $x_0 \in (c, d)$. Then, restricting x to $[c, d]$,

$$(f * g)'(x_0) = \frac{d}{dx}\left(\int_{-\infty}^{+\infty} f(y)g(x-y)\,dy\right)(x_0) = \frac{d}{dx}\left(\int_{c-b}^{d-a} H(x, y)\,dy\right)(x_0)$$

$$= \int_{c-b}^{d-a} \frac{\partial H}{\partial x}(x_0, y)\,dy = \int_{-\infty}^{+\infty} f(y)g'(x_0 - y)\,dy = (f * (g'))(x_0)\ .$$

This proves the result for $r = 1$. For larger r apply the $r = 1$ case inductively to f and $g^{(k)}$. 👍

XVIII.5.1.9. Of course, by XVIII.5.1.2., the result also holds if f is $\mathcal{C}^r$, g is continuous, and at least one has compact support; in this case $(f * g)^{(k)} = (f^{(k)}) * g$. If f and g are both $\mathcal{C}^1$, and at least one has compact support, then $(f') * g = f * (g')$. If f and g are $\mathcal{C}^2$, then $(f'') * g = (f') * (g') = f * (g'')$, etc.

In fact, we have:

XVIII.5.1.10. Corollary. Let f be a $\mathcal{C}^r$ function and g a $\mathcal{C}^s$ function on $\mathbb{R}$ $(0 \leq r, s < \infty)$. Suppose either f or g has compact support. Then $f * g$ is $\mathcal{C}^{r+s}$, and $(f * g)^{(r+s)} = (f^{(r)}) * (g^{(s)})$.

Proof: The convolution $f * g$ is $\mathcal{C}^r$ and $(f * g)^{(r)} = (f^{(r)}) * g$; and $(f^{(r)}) * g$ is $\mathcal{C}^s$ with $[(f^{(r)}) * g]^{(s)} = (f^{(r)}) * (g^{(s)})$. 👍

Convolutions on Intervals

XVIII.5.1.11. If f has compact support contained in $[a, b]$, and g is any continuous function on $\mathbb{R}$, then for any interval $[c, d]$ the values of $f * g$ on $[c, d]$ depend only on the values of g on the interval $[c - b, d - a]$. This leads to two important observations:

(1) The function g can be modified arbitrarily outside the interval $[c - b, d - a]$ without affecting the values of $f * g$ on $[c, d]$. In particular, there is a continuous function h of compact support agreeing with g on an arbitrarily large interval such that $f * h$ agrees with $f * g$ on $[c, d]$. This observation is useful in studying local properties of $f * g$; in particular, it can be used to give an alternate proof of the "harder" half of XVIII.5.1.8..

(2) If g is any continuous function on $[c - b, d - a]$, then we can define the convolution of f and g as a function on the interval $[c, d]$.

Algebraic Properties of Convolution

Convolution satisfies all the algebraic properties of a reasonable "multiplication". We have already established the commutative property $f * g = g * f$ (XVIII.5.1.2.). Another simple property whose proof is left to the reader (Exercise ()) is the distributive or bilinear property:

XVIII.5.1.12. Proposition. Let f_1, f_2, g be continuous functions on $\mathbb{R}$, and $\alpha \in \mathbb{R}$.

(i) If $f_1 * g$ and $f_2 * g$ are defined, then $(f_1 + f_2) * g$ is defined and $(f_1 + f_2) * g = f_1 * g + f_2 * g$.

(ii) If $f_1 * g$ is defined, then $(\alpha f_1) * g$ is defined and $(\alpha f_1) * g = \alpha(f_1 * g)$.

In other words, for fixed g, convolution $f * g$ is linear in f. Of course, by commutativity, $f * g$ is also linear in g for fixed f. Thus convolution is *bilinear*.

The other important algebraic property, associativity, also holds; the proof requires Fubini's Theorem for vector calculus ().

XVIII.5.1.13. PROPOSITION. Let f, g, h be continuous functions on $\mathbb{R}$. Suppose at least two of the functions have compact support. Then $(f * g) * h = f * (g * h)$.

PROOF: If $x_0 \in \mathbb{R}$, then the function which does not have compact support can be replaced by a function of compact support without changing either $(f*g)*h$ or $f*(g*h)$ in an interval around x_0, by two applications of XVIII.5.1.11.(1). Thus to prove the result we may assume all three functions have compact support. We have

$$[(f * g) * h](x_0) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} f(z)g(y - z)\, dz \right] h(x_0 - y)\, dy$$

$$[f * (g * h)](x_0) = \int_{-\infty}^{+\infty} f(z) \left[\int_{-\infty}^{+\infty} g(y)h(x_0 - z - y)\, dy \right] dz$$

where the integrals are actually over a sufficiently large bounded region (rectangle) in $\mathbb{R}^2$. By Fubini's Theorem () the integrals are equal to the double integrals

$$\iint_{\mathbb{R}^2} f(z)g(y - z)h(x_0 - y)\, dy\, dz$$

$$\iint_{\mathbb{R}^2} f(z)g(y)h(x_0 - z - y)\, dy\, dz$$

respectively, where the integrals are again really over a sufficiently large bounded region.

Make the change of variables $\phi(u, v) = (u + v, v)$. Writing $y = u + v$, $z = v$, we have $x_0 - y = x_0 - u - v$ and $y - z = u$. We have $\Delta_\phi(u, v) \equiv 1$, so by the change-of-variables formula for double integrals () we have

$$\iint_{\mathbb{R}^2} f(z)g(y-z)h(x_0-y)\, dy\, dz = \iint_{\mathbb{R}^2} f(v)g(u)h(x_0-u-v)|\Delta_\phi(u, v)|\, du\, dv = \iint_{\mathbb{R}^2} f(v)g(u)h(x_0-u-v)\, du\, dv$$

(where the integrals are over large bounded regions), so the double integrals are equal. 👍

Integrable Functions

We can give various other conditions insuring the existence of the convolution. If neither function has compact support, then in general restrictions on both functions are necessary.

XVIII.5.1.14. DEFINITION. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then f is *integrable* (over $\mathbb{R}$) if

$$\|f\|_1 := \int_{-\infty}^{+\infty} |f(x)|\, dx < \infty$$

(i.e. the improper integral converges).

XVIII.5.1.15. The set $CL^1(\mathbb{R})$ of integrable continuous functions on $\mathbb{R}$ is a real vector space, and $\|\cdot\|_1$ is a norm () on $CL^1(\mathbb{R})$. Every continuous function of compact support is in $CL^1(\mathbb{R})$, but $CL^1(\mathbb{R})$ is much larger than the set $C_c(\mathbb{R})$ of continuous functions of compact support: for example, an integrable continuous function need not even be bounded (Example XVIII.5.1.19.). The set $BCL^1(\mathbb{R})$ of bounded continuous integrable functions on $\mathbb{R}$ is a vector space properly between $C_c(\mathbb{R})$ and $CL^1(\mathbb{R})$.

XVIII.5.1.16. PROPOSITION. Let f and g be continuous functions on $\mathbb{R}$, with f integrable and g bounded. Then $f * g$ is defined and bounded, and $\|f * g\|_\infty \leq \|f\|_1 \|g\|_\infty$.

PROOF: We have that, for each x,

$$\int_{-\infty}^{+\infty} |f(x-y)g(y)|\,dy \leq \int_{-\infty}^{+\infty} |f(x-y)|\|g\|_\infty\,dy = \|g\|_\infty \int_{-\infty}^{+\infty} |f(x-y)|\,dy = \|f\|_1\|g\|_\infty$$

so by () the improper integral

$$\int_{-\infty}^{+\infty} f(x-y)g(y)\,dy$$

converges and

$$|(f*g)(x)| = \left|\int_{-\infty}^{+\infty} f(x-y)g(y)\,dy\right| \leq \int_{-\infty}^{+\infty} |f(x-y)g(y)|\,dy \leq \|f\|_1\|g\|_\infty \ .$$

👍

XVIII.5.1.17. PROPOSITION. Let f and g be continuous functions on $\mathbb{R}$, with f integrable and g bounded (or vice versa). Then $f * g$ is continuous.

PROOF: Let (f_n) be a sequence of continuous functions of compact support with $\|f_n - f\|_1 \to 0$ (). Then

$$\|f_n * g - f * g\|_\infty = \|(f_n - f) * g\|_\infty \leq \|f_n - f\|_1\|g\|_\infty \to 0$$

so $f_n * g \to f * g$ uniformly. By XVIII.5.1.6., each $f_n * g$ is continuous; hence $f * g$ is continuous by (). 👍

XVIII.5.1.18. PROPOSITION. Let f and g be integrable continuous functions on $\mathbb{R}$, with at least one of f and g bounded. Then $f * g$ is defined and in $BCL^1(\mathbb{R})$, and

$$\|f * g\|_1 \leq \|f\|_1\|g\|_1 \ .$$

PROOF: By XVIII.5.1.16. and XVIII.5.1.17., $f * g$ is defined, and is bounded and continuous. To show it is integrable, consider the integral

$$\iint_{\mathbb{R}^2} |f(y)||g(x-y)|\,dx\,dy \ .$$

We can change variables in this integral using $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $\phi(z,y) = (z+y, y)$. Then $x = z + y$, $z = x - y$, and we have $\Delta_\phi(z,y) \equiv 1$, so

$$\iint_{\mathbb{R}^2} |f(y)||g(x-y)|\,dx\,dy = \iint_{\mathbb{R}^2} |f(y)||g(z)||\Delta_\phi(z,y)|\,dz\,dy = \iint_{\mathbb{R}^2} |f(y)||g(z)|\,dz\,dy \ .$$

By the Fubini-Tonelli Theorem () this integral can be computed as an iterated integral

$$\int_{\mathbb{R}}\int_{\mathbb{R}} |f(y)||g(z)|\,dz\,dy = \int_{\mathbb{R}} |f(y)| \left[\int_{\mathbb{R}} |g(z)|\,dz\right] dy = \left[\int_{\mathbb{R}} |f(y)|\,dy\right]\left[\int_{\mathbb{R}} |g(z)|\,dz\right] = \|f\|_1\|g\|_1 \ .$$

👍

XVIII.5.1.19. EXAMPLE. Here is an example showing that the boundedness hypothesis in XVIII.5.1.18. cannot be deleted. Define a nonnegative continuous function f on $\mathbb{R}$ as follows:

$$f(x) = \begin{cases} 0 & \text{if} \quad x \in \left[0, \frac{15}{8}\right] \\ n & \text{if} \quad x \in \left[n - \frac{1}{2n^3}, n + \frac{1}{2n^3}\right], n \geq 2 \\ \text{linear} & \text{if} \quad x \in \left[n - \frac{1}{n^3}, n - \frac{1}{2n^3}\right], n \geq 2 \\ \text{linear} & \text{if} \quad x \in \left[n + \frac{1}{2n^3}, n + \frac{1}{n^3}\right], n \geq 2 \\ 0 & \text{if} \quad x \in \left[n + \frac{1}{n^3}, n + 1 - \frac{1}{(n+1)^3}\right], n \geq 2 \\ f(-x) & \text{if} \quad x < 0 \end{cases} .$$

The graph of f is identically 0 except for small "blunt spikes" around integers; see Figure (). We have

$$\int_{-\infty}^{+\infty} f(x)\, dx < \infty$$

$$\int_{-\infty}^{+\infty} [f(x)]^2\, dx = \infty$$

i.e. f is integrable but not square-integrable.

Consider $f * f$. We have

$$\int_{-\infty}^{+\infty} f(y) f(-y)\, dy = \int_{-\infty}^{+\infty} [f(y)]^2\, dy = \infty$$

so $f * f(0)$ is not defined. In fact, $f * f(x)$ is defined if and only if x is not an integer.

XVIII.5.1.20. It turns out that using Lebesgue measure and integration we can still make a reasonable definition of the convolution $f * f$ of XVIII.5.1.19. (or of any two integrable functions), but the convolution is not continuous and only defined "almost everywhere." This is in fact one of the historical motivations for developing a more general theory of measure and integration. See section ().

XVIII.5.1.21. Thus convolution gives a binary operation of $BCL^1(\mathbb{R})$. Convolution on $BCL^1(\mathbb{R})$ retains nice algebraic properties: associativity (), commutativity (XVIII.5.1.2.), and bilinearity ().

XVIII.5.1.22. Using Hölder's inequality (), XVIII.5.1.16. can be extended. If $1 \leq p < \infty$, a continuous function f on $\mathbb{R}$ is *p-integrable* if

$$\|f\|_p^p := \int_{-\infty}^{+\infty} |f(x)|^p\, dx < \infty .$$

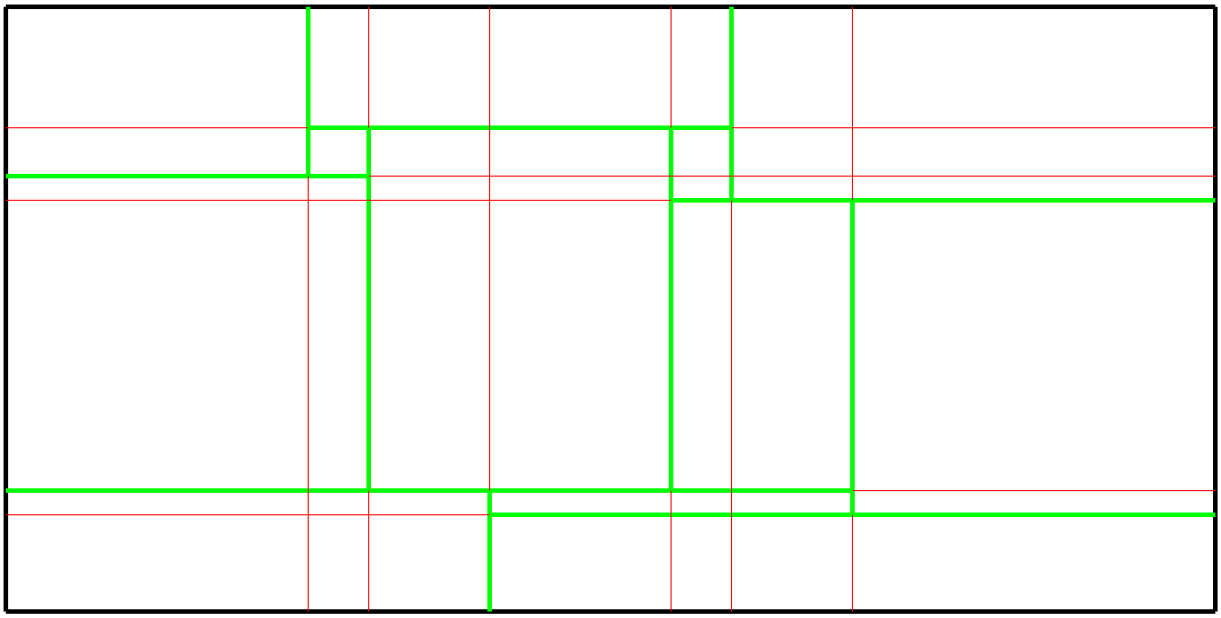
XVIII.5.1.23. THEOREM. Let $1 < p < \infty$, and $q = \frac{p}{p-1}$ the conjugate exponent () of p. Let f and g be continuous functions which are p-integrable and q-integrable respectively. Then $f * g$ is defined, continuous, and bounded, and $\|f * g\|_\infty \leq \|f\|_p \|g\|_q$.

See () for the proof, which actually gives a more general statement not requiring f and g to be continuous. Using this, another version of XVIII.5.1.18. can be obtained with an identical proof:

XVIII.5.1.24. COROLLARY. Let f and g be integrable continuous functions on $\mathbb{R}$. Suppose there is a p, $1 < p < \infty$, such that f is p-integrable and g is q-integrable, where $q = \frac{p}{p-1}$ is the conjugate exponent to p. Then $f * g$ is defined and in $BCL^1(\mathbb{R})$, and

$$\|f * g\|_1 \leq \|f\|_1 \|g\|_1 \ .$$

In particular, if f and g are continuous, integrable, and 2-integrable (square-integrable), then $f * g$ is defined, continuous, and bounded.



XVIII.6. The Radon-Nikodým Theorem

The Radon-Nikodým Theorem is one of the important results of the theory of measure and integration, with many varied applications. It describes a precise way that two measures on the same measurable space can be compared.

XVIII.6.1. Density Functions

XVIII.6.1.1. Let μ be a measure on a measurable space $(X, \mathcal{A})$, and ϕ a nonnegative extended real-valued $\mathcal{A}$-measurable function. For $A \in \mathcal{A}$, define

$$\nu(A) = \int_A \phi \, d\mu \ .$$

Then ν is a measure on $(X, \mathcal{A})$ (XVIII.1.9.2.). The function ϕ is a *density function* for ν with respect to μ (μ-*density function for* ν).

XVIII.6.1.2. If μ is a fixed measure on $(X, \mathcal{A})$, which measures ν on $(X, \mathcal{A})$ have μ-density functions? There is a simple necessary condition: ν must be absolutely continuous () with respect to μ, i.e. $\nu(A) = 0$ whenever $A \in \mathcal{A}$ and $\mu(A) = 0$.

XVIII.6.1.3. Proposition. Let μ and ν be measures on $(X, \mathcal{A})$. If ν has a μ-density function, then $\nu \ll \mu$.

Proof: Let ϕ be a μ-density function for ν. If $A \in \mathcal{A}$ and $\mu(A) = 0$, then

$$\nu(A) = \int_A \phi \, d\mu = 0$$

by XVIII.1.3.2.(iv). 👍

XVIII.6.1.4. The Radon-Nikodým Theorem asserts the converse if μ is σ-finite: if $\nu \ll \mu$, then there is a μ-density function for ν.

XVIII.6.1.5. Example. Let λ be Lebesgue measure on $([0,1], \mathcal{L})$, and μ counting measure on $([0,1], \mathcal{L})$. Then $\lambda \ll \mu$ (in fact, any measure on $([0,1], \mathcal{L})$ is absolutely continuous with respect to μ since only $\emptyset$ has μ-measure 0). But there cannot be a Lebesgue-measurable function ϕ such that $\lambda(A) = \int_A \phi \, d\mu$ for all $A \in \mathcal{L}$: for any such ϕ, we would have

$$0 = \lambda(\{x\}) = \int_{\{x\}} \phi \, d\mu = \phi(x)$$

for every $x \in [0,1]$. Note that λ is a finite measure on $([0,1], \mathcal{L})$, but that μ is not σ-finite, although it is decomposable.

Borel Measures on the Line

XVIII.6.1.6. The case where μ is Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ is especially important (we write μ instead of λ since we are working with $\mathcal{B}$ instead of $\mathcal{L}$, i.e. $\mu = \lambda|_{\mathcal{B}}$). Suppose ν is a finite Borel measure on $\mathbb{R}$, with cdf F_ν (). Then, as shown in ():

XVIII.6.1.7. THEOREM. Let ν be a finite Borel measure on $\mathbb{R}$ with cdf F_ν. The following are equivalent:

(i) $\nu \ll \mu$.

(ii) F_ν is absolutely continuous.

(iii) $\nu(B) = \int_B F'_\nu \, d\mu$ for every Borel set B.

(iv) ν has a μ-density function.

The equivalence of (i) and (iv) is the Radon-Nikodým Theorem in this special case. The equivalence of (i) and (ii) is motivation for the term "absolutely continuous" for measures.

XVIII.6.1.8. In fact, in the case of finite Borel measures on $\mathbb{R}$, if $\nu \ll \mu$, then F'_ν is a μ-density function for ν. Thus the μ-density function for ν is a "derivative." In the general setting, there is no direct analog of a cdf, but the μ-density function for ν, if it exists, is often called the *Radon-Nikodým derivative* of ν with respect to μ and written $\frac{d\nu}{d\mu}$ (essential uniqueness of the density function will be discussed in XVIII.6.1.10.). The idea that $\frac{d\nu}{d\mu}$ is a "derivative" is reflected in the formula

$$\nu(A) = \int_A \frac{d\nu}{d\mu} \, d\mu$$

for $A \in \mathcal{A}$, which is reminiscent of the Fundamental Theorem of Calculus.

Essential Uniqueness of Density Functions

The following observation, a special case of (), implies that if μ is σ-finite, a μ-density function for a measure ν, if it exists, is essentially unique, and is also the key idea in using the theory of signed measures to prove the Radon-Nikodým Theorem.

XVIII.6.1.9. PROPOSITION. Let μ be a finite measure on $(X, \mathcal{A})$, and ν a measure on $(X, \mathcal{A})$ with a density function ϕ. For each $r \in \mathbb{R}$, set

$$P_r = \{x \in X : \phi(x) \geq r\} \, .$$

Then $(P_r, X \setminus P_r)$ is a Hahn decomposition of X for the signed measure $\nu - r\mu$, and in particular P_r is a positive set and $X \setminus P_r$ a negative set for $\nu - r\mu$.

XVIII.6.1.10. PROPOSITION. Let μ be a σ-finite measure on a measurable space $(X, \mathcal{A})$, and ν a measure on $(X, \mathcal{A})$. If ϕ and ψ are μ-density functions for ν, then $\phi = \psi$ μ-a.e.

PROOF: First suppose μ is finite. Set $A = \{x \in X : \phi(x) < \psi(x)\}$, and for each $r, s \in \mathbb{Q}$, $r < s$,

$$A_{r,s} = \{x \in X : \phi(x) < r \text{ and } \psi(x) > s\} .$$

We then have

$$A = \cup\{A_{r,s} : r, s \in \mathbb{Q}, r < s\}$$

and since this is a countable union, to show that A is μ-null it suffices to show that $A_{r,s}$ is μ-null for each $r, s \in \mathbb{Q}$, $r < s$. Fix r and s. Using the notation of XVIII.6.1.9. (where P_s uses ψ and P_r uses ϕ), we have that $A_{r,s} \subseteq P_s \cap (X \setminus P_r)$. In particular, $A_{r,s}$ is a positive set for $\nu - s\mu$ and a negative set for $\nu - r\mu$. We thus have

$$(\nu - r\mu)(A_{r,s}) \leq 0, \quad \nu(A_{r,s}) \leq r\mu(A_{r,s})$$

and in particular $\nu(A_{r,s}) < \infty$ since $\mu(A_{r,s}) < \infty$. We also have

$$(\nu - s\mu)(A_{r,s}) \geq 0, \quad \nu(A_{r,s}) \geq s\mu(A_{r,s})$$

and combining the two, we get

$$s\mu(A_{r,s}) \leq r\mu(A_{r,s})$$

and since $r < s$ and $\mu(A_{r,s}) \geq 0$, we conclude that $\mu(A_{r,s}) = 0$. Thus $\mu(A) = 0$. Similarly,

$$\mu(\{x \in X : \psi(x) < \phi(x)\}) = 0 .$$

Thus $\phi = \psi$ μ-a.e.

If μ is only σ-finite, partition X into a countable collection (X_n) of measurable sets with $\mu(X_n) < \infty$ for all n, and set $\mu_n = \mu|_{X_n}$, $\nu_n = \nu|_{X_n}$, $\phi_n = \phi|_{X_n}$, $\psi_n = \psi|_{X_n}$. Then, for each n, ϕ_n and ψ_n are μ_n-density functions for ν_n, so $\phi_n = \psi_n$ μ_n-a.e. and therefore $\phi = \psi$ μ-a.e.

XVIII.6.1.11. Thus, in the σ-finite case, density functions, if they exist, are essentially unique, so by slight abuse of terminology we may refer to *the* μ-density function of ν (if there is one). But it should be kept in mind that the density function is only determined μ-a.e. and is technically not unique; any nonnegative measurable function agreeing a.e. with a μ-density function for ν is also a μ-density function for ν. In particular, the Radon-Nikodým derivative $\frac{d\nu}{d\mu}$ (XVIII.6.1.8.) is technically not defined as a specific function, but only as an equivalence class of functions, and thus some authors write $\left[\frac{d\nu}{d\mu}\right]$ for the Radon-Nikodým derivative to emphasize the nonuniqueness of definition, although this notation seems to be passing out of favor. (The distinction rarely causes any difficulties.)

Actually, it is even a slight abuse of language to say that $\frac{d\nu}{d\mu}$ is uniquely determined a.e.; it is technically not uniquely determined anywhere in most cases, in the sense that $\frac{d\nu}{d\mu}(x)$ is not uniquely determined for any specific x. We will nonetheless sometimes use this language for brevity.

XVIII.6.2. Statement and Proof of the Theorem

There are several proofs known of the Radon-Nikodým Theorem. The one we give is not the shortest, but it gives the most general result, and we think it gives the best insight into why the theorem is true. There are some slicker proofs of the case where ν, as well as μ, is σ-finite (admittedly by far the most important case), cf. (), (). Our proof is adapted from [Roy88] and [MW99] (note that in [**?**] this proof has been replaced by the proof in ()).

XVIII.6.2.1. THEOREM. [RADON-NIKODÝM THEOREM] Let $(X, \mathcal{A})$ be a measurable space, and μ and ν measures on $(X, \mathcal{A})$, with μ σ-finite. If $\nu \ll \mu$, then there is a μ-density function for ν, i.e. there is a nonnegative extended real-valued $\mathcal{A}$-measurable function ϕ such that

$$\nu(A) = \int_A \phi \, d\mu$$

for all $A \in \mathcal{A}$. The function ϕ is uniquely determined a.e. and is called the *Radon-Nikodým derivative* of ν with respect to μ, denoted $\frac{d\nu}{d\mu}$.

PROOF: First assume μ is finite. Using XVIII.6.1.9. as motivation, for each nonnegative *rational* number r we take a Hahn decomposition (P_r, N_r) for the signed measure $\nu - r\mu$. The set

$$Q_r = [\cup_{s \geq r} P_s] \cup [\cap_{s < r} P_s]$$

(since r and s only range over the countable set $\mathbb{Q}_+$, this set is $\mathcal{A}$-measurable) is a positive set for $\nu - r\mu$ containing P_r, so $(Q_r, X \setminus Q_r)$ is also a Hahn decomposition for $\nu - r\mu$. We have $Q_s \subseteq Q_r$ if $r < s$. Also take $Q_0 = X$. (It simplifies the argument to use the Q_r.)

Now set

$$\phi(x) = \sup\{r : x \in Q_r\}$$

for $x \in X$. We claim ϕ is a μ-density function for ν. Note that $Q_r = \{x : \phi(x) \geq r\}$.

First note that $\phi = \sup_{r \in \mathbb{Q}_+} \phi_r$, where $\phi_r = r\chi_{Q_r}$. Since each ϕ_r is $\mathcal{A}$-measurable, ϕ is $\mathcal{A}$-measurable, and obviously nonnegative.

We now fix $A \in \mathcal{A}$. Let $A_\infty = \{x \in A : \phi(x) = \infty\}$. Then $A_\infty \subseteq Q_r$ for all r, so is a positive set for $\nu - r\mu$, i.e. $\nu(A_\infty) \geq r\mu(A_\infty)$ for all $r \geq 0$. If $\mu(A_\infty) > 0$, we have $\nu(A_\infty) = \infty$; if $\mu(A_\infty) = 0$, then $\nu(A_\infty) = 0$ since $\nu \ll \mu$. In either case, we have

$$\nu(A_\infty) = \int_{A_\infty} \phi \, d\mu \ .$$

In particular, if $\mu(A_\infty) > 0$, the density formula holds for ϕ and A: both $\nu(A)$ and $\int_A \phi \, d\mu$ are ∞.

Now assume $\mu(A_\infty) = 0$. Fix $n \in \mathbb{N}$, and for each $k \in \mathbb{N}$ set

$$A_{n,k} = \left\{ x \in A : \frac{k-1}{n} \leq \phi(x) < \frac{k}{n} \right\} \ .$$

Then A is the disjoint union of A_∞ and the $A_{n,k}$ (and these sets are all $\mathcal{A}$-measurable). Then

$$\frac{k-1}{n}\mu(A_{n,k}) \leq \int_{A_{n,k}} \phi \, d\mu \leq \frac{k}{n}\mu(A_{n,k}) \ .$$

We have

$$A_{n,k} \subseteq Q_{(k-1)/n} \setminus Q_{k/n}$$

so we have

$$\frac{k-1}{n}\mu(A_{n,k}) \leq \nu(A_{n,k}) \leq \frac{k}{n}\mu(A_{n,k}) \ .$$

Thus

$$\int_{A_{n,k}} \phi \, d\mu - \frac{1}{n}\mu(A_{n,k}) \leq \nu(A_{n,k}) \leq \int_{A_{n,k}} \phi \, d\mu + \frac{1}{n}\mu(A_{n,k}) \ .$$

Since A is the disjoint union of A_∞ and the $A_{n,k}$, and $\mu(A_\infty) = 0$ (so $\nu(A_\infty) = 0$), summing over k we have

$$\int_A \phi \, d\mu - \frac{1}{n}\mu(A) \leq \nu(A) \leq \int_A \phi \, d\mu + \frac{1}{n}\mu(A)$$

(if one of $\int_A \phi \, d\mu$ and $\nu(A)$ is ∞, so is the other). If $\int_A \phi \, d\mu$ and $\nu(A)$ are finite, this is true for all $n \in \mathbb{N}$, so

$$\int_A \phi \, d\mu = \nu(A) \ .$$

This proves the theorem if μ is finite.

Now suppose μ is just σ-finite. Partition X into a countable collection (X_n) of measurable sets with $\mu(X_n) < \infty$ for all n, and set $\mu_n = \mu|_{X_n}$ and $\nu_n = \nu|_{X_n}$ for each n. Then there is a μ_n-density function ϕ_n for each n. Let ϕ be the function on X whose restriction to X_n is ϕ_n for each n. Then ϕ is a nonnegative $\mathcal{A}$-measurable function on X. If $A \in \mathcal{A}$, we have

$$\int_A \phi \, d\mu = \sum_{n=1}^{\infty} \int_{A \cap X_n} \phi \, d\mu = \sum_{n=1}^{\infty} \int_{A \cap X_n} \phi_n \, d\mu_n = \sum_{n=1}^{\infty} \nu_n(A \cap X_n) = \sum_{n=1}^{\infty} \nu(A \cap X_n) = \nu(A)$$

so ϕ is a μ-density function for ν.

The last assertion is XVIII.6.1.10.. 👍

Note that the proof uses the Countable AC (where?)

XVIII.6.2.2. Without the σ-finite hypothesis on μ, the result can fail. A nice standard example is XVIII.6.1.5.. There is even an extremely simple and rather stupid counterexample: let $X = \{x\}$ be a singleton set, and define μ and ν on $(X, \mathcal{P}(X))$ by $\mu(X) = \infty$ and $\nu(X) = 1$. Then $\nu \ll \mu$, but there is no density function. There is an extension of the theorem to the case where μ is decomposable, but it must be carefully stated; cf. [HS75, 19.27].

Using (), the Radon-Nikodým Theorem can be extended:

XVIII.6.2.3. Corollary. Let μ and ν be measures on a measurable space $(X, \mathcal{A})$, with μ σ-finite and $\nu \ll \mu$. Then there is a nonnegative extended real-valued $\mathcal{A}$-measurable function $\frac{d\nu}{d\mu}$, uniquely determined a.e., with the following properties:

(i) If f is a nonnegative extended real-valued $\mathcal{A}$-measurable function, then

$$\int_A f\,d\nu = \int_A f\frac{d\nu}{d\mu}\,d\mu$$

for all $A \in \mathcal{A}$.
(ii) If f is a ν-integrable function, then $f\frac{d\nu}{d\mu}$ is μ-integrable and

$$\int_A f\,d\nu = \int_A f\frac{d\nu}{d\mu}\,d\mu$$

for all $A \in \mathcal{A}$.

Finiteness of the Radon-Nikodým Derivative

XVIII.6.2.4. Let μ and ν be measures on $(X, \mathcal{A})$, with μ σ-finite and $\nu \ll \mu$. Set

$$X_\infty = \left\{x \in X : \frac{d\nu}{d\mu}(x) = \infty\right\}$$

(where $\frac{d\nu}{d\mu}$ is some fixed Radon-Nikodým derivative). If $A \in \mathcal{A}$, $A \subseteq X_\infty$, then $\nu(A)$ is either 0 (if $\mu(A) = 0$) or ∞ (if $\mu(A) > 0$). Thus if ν is σ-finite (or, more generally, semifinite), then $\mu(X_\infty) = 0$, i.e. $\frac{d\nu}{d\mu}$ is finite a.e. It can thus be modified on a set of measure 0 to be finite everywhere. Conversely, if $\frac{d\nu}{d\mu}$ is finite a.e. then ν must be σ-finite. (This has the interesting consequence that if μ is σ-finite and $\nu \ll \mu$ is semifinite, then ν is actually σ-finite.)

History

XVIII.6.2.5. Theorem XVIII.6.1.7. was proved by Lebesgue. Beppo Levi criticized the rigor of Lebesgue's proof and suggested fixes, and Lebesgue wrote two additional papers about the result. According to [Bog07], "Large portions of [these two papers] are occupied by Lebesgue's polemics with B. Levi with respect to the critical remarks of the latter and the rigor of his arguments." The result for Borel measures on $\mathbb{R}^n$ was proved by Radon, and the general theorem by Nikodým. Some authors call the theorem the "Lebesgue-Radon-Nikodým Theorem," sometimes incorporating the Lebesgue Decomposition Theorem () into the result. Although the attributions may not be clearcut, we have used the most common names for these results.

XVIII.6.3. The Lebesgue Decomposition Theorem

XVIII.6.4. Signed and Complex Measures

XVIII.6.5. Conditional Expectations

One of the most striking applications of the Radon-Nikodým Theorem is to the existence of conditional expectations. Although these can be defined somewhat more generally, we will restrict attention to probability measures, both for simplicity and for the fact that conditional expectations are particularly important in probability.

XVIII.6.5.1. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a probability measure space, and f a μ-integrable function (real- or complex-valued). If $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$, a *conditional expectation of f given $\mathcal{B}$* is a $\mathcal{B}$-measurable function ϕ satisfying

$$\int_B \phi \, d\mu = \int_B f \, d\mu$$

for all $B \in \mathcal{B}$.

XVIII.6.5.2. A conditional expectation, if it exists, is essentially unique in the sense that any two conditional expectations for f given $\mathcal{B}$ are equal a.e. Conversely, any $\mathcal{B}$-measurable function equal a.e. to a conditional expectation is also a conditional expectation. By slight abuse of terminology, we will denote a conditional expectation for f given $\mathcal{B}$ by $\mathbf{E}(f|\mathcal{B})$.

> "[I]n almost every discussion involving conditional expectations, *almost everywhere* or *almost surely*, or their abbreviations *a.e.*, *a.s.*, will be sure to appear almost everywhere."
>
> *J. Doob*[1]

XVIII.6.5.3. What is the content of this concept, i.e. how does $\mathbf{E}(f|\mathcal{B})$ differ from f? The key is that $\mathbf{E}(f|\mathcal{B})$ must be $\mathcal{B}$-measurable, whereas f is merely $\mathcal{A}$-measurable (and integrable). Recall that the smaller the σ-algebra, the harder it is for a function to be measurable. (Of course, if f happens to be $\mathcal{B}$-measurable, we can indeed take $\mathbf{E}(f|\mathcal{B}) = f$, but this situation is generally exceptional.)

XVIII.6.5.4. EXAMPLES. (i) Suppose $\mathcal{B} = \{\emptyset, X\}$. Then the only $\mathcal{B}$-measurable functions are constant functions. If f is $\mathcal{A}$-measurable and μ-integrable, $\mathbf{E}(f|\mathcal{B})$ must be a constant function α with the property that

$$\int_X \alpha \, d\mu = \int_X f \, d\mu \ .$$

Since μ is a probability measure, the left side is just α, so

$$\mathbf{E}(f|\mathcal{B}) = \int_X f \, d\mu$$

where the right side is regarded as a constant function. Recall that in probability the right side is called the *expectation* of f. Thus the conditional expectation of f given the trivial σ-algebra is just the expectation of f (regarded as a constant function).

(ii) For a slightly less trivial example, fix a set $A \in \mathcal{A}$ with $0 < \mu(A) < 1$ (so $\mu(A^c) > 0$). Let $\mathcal{B} = \{\emptyset, A, A^c, X\}$ be the σ-algebra generated by A. The $\mathcal{B}$-measurable functions are precisely the functions which are constant on A and on A^c. If f is $\mathcal{A}$-measurable and μ-integrable, and $\mathbf{E}(f|\mathcal{B})$ is a conditional expectation, then the constant value α that $\mathbf{E}(f|\mathcal{B})$ takes on A must satisfy

$$\int_A \alpha \, d\mu = \int_A f \, d\mu$$

[1] [Doo94, p. 180].

and since the left side is $\alpha\mu(A)$, we obtain

$$\alpha = \frac{1}{\mu(A)} \int_A f \, d\mu$$

which can be interpreted as the average value of f on A. Probabilists call this number the *conditional expectation of f given A*. Similarly, the constant value $\mathbf{E}(f|\mathcal{B})$ must take on A^c is

$$\beta = \frac{1}{\mu(A^c)} \int_{A^c} f \, d\mu$$

which is the conditional expectation of f given A^c. If $f = \chi_B$, then we have

$$\alpha = \frac{1}{\mu(A)} \int_A \chi_B \, d\mu = \frac{\mu(A \cap B)}{\mu(A)}$$

which probabilists call the *conditional probability of B given A*, and

$$\beta = \frac{\mu(A^c \cap B)}{\mu(A^c)}$$

which is the conditional probability of B given A^c.

If $\mathbf{E}(f|\mathcal{B})$ is defined this way, we automatically have

$$\int_X \mathbf{E}(f|\mathcal{B}) \, d\mu = \int_A \mathbf{E}(f|\mathcal{B}) \, d\mu + \int_{A^c} \mathbf{E}(f|\mathcal{B}) \, d\mu = \int_A f \, d\mu + \int_{A^c} f \, d\mu = \int_X f \, d\mu$$

so $\mathbf{E}(f|\mathcal{B})$ is indeed a conditional expectation for f given $\mathcal{B}$. This example is the origin of the name "conditional expectation" for the concept in XVIII.6.5.1..

(iii) Let $(X, \mathcal{A}_1, \mu)$ and $(Y, \mathcal{A}_2, \nu)$ be probability measure spaces, and form $(X \times Y, \mathcal{A}_1 \otimes \mathcal{A}_2, \mu \times \nu)$, which is again a probability measure space. Let

$$\mathcal{B} = \{A \times Y : A \in \mathcal{A}_1\} \, .$$

Then $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}_1 \otimes \mathcal{A}_2$. A function f on $X \times Y$ is $\mathcal{B}$-measurable if and only it is of the form $f(x, y) = g(x)$ for an $\mathcal{A}_1$-measurable function g, i.e. if it is constant on vertical cross sections. If f is an $(\mathcal{A}_1 \times \mathcal{A}_2)$-measurable function which is $(\mu \times \nu)$-integrable, a conditional expectation $\mathbf{E}(f|\mathcal{B})$ must be constant on vertical sections and satisfy

$$\int_{A \times Y} \mathbf{E}(f|\mathcal{B}) \, d(\mu \times \nu) = \int_{A \times Y} f \, d(\mu \times \nu)$$

for all $A \in \mathcal{A}_1$. By Fubini's Theorem, we may (and must) take the constant value on the vertical cross section $\{x\} \times Y$ to be

$$E(f|\mathcal{B})(x, z) = \int_Y f(x, y) d\nu(y)$$

for all $z \in Y$.

XVIII.6.5.5. If $\mathcal{B}$ is a complicated σ-algebra, it is highly unclear how a conditional expectation for an f can be defined or even whether one exists. But the Radon-Nikodým Theorem comes to the rescue in a quite simple way:

XVIII.6.5.6. THEOREM. Let $(X, \mathcal{A}, \mu)$ be a probability measure space, and f an $\mathcal{A}$-measurable, μ-integrable complex-valued function on X. If $\mathcal{B}$ is any sub-σ-algebra of $\mathcal{A}$, then there exists a conditional expectation for f given $\mathcal{B}$.

PROOF: On the measurable space $(X, \mathcal{B})$, for $B \in \mathcal{B}$, define

$$\nu(B) = \int_B f \, d\mu \ .$$

Then ν is a complex measure on $(X, \mathcal{B})$, and $\nu \ll \mu|_{\mathcal{B}}$. Thus there is a $\mathcal{B}$-measurable function $\phi = \frac{d\nu}{d(\mu|_{\mathcal{B}})}$ such that

$$\nu(B) = \int_B \phi \, d(\mu|_{\mathcal{B}}) = \int_B \phi \, d\mu$$

for all $B \in \mathcal{B}$ (). Thus we can set $\mathbf{E}(f|\mathcal{B}) = \phi$. 👍

Properties of Conditional Expectations

Conditional expectations have most of the nice properties one would expect or hope for, although the lack of exactness in their definition makes precise statements annoyingly complicated.

XVIII.6.5.7. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a probability measure space, and $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$.
(i) If f and g are $\mathcal{A}$-measurable and μ-integrable functions, then

$$\mathbf{E}(f + g|\mathcal{B}) = \mathbf{E}(f|\mathcal{B}) + \mathbf{E}(g|\mathcal{B})$$

in the sense that $\mathbf{E}(f|\mathcal{B}) + \mathbf{E}(g|\mathcal{B})$ is a conditional expectation for $f + g$ given $\mathcal{B}$.
(ii) If f is an $\mathcal{A}$-measurable and μ-integrable function and $\alpha \in \mathbb{C}$, then

$$\mathbf{E}(\alpha f|\mathcal{B}) = \alpha \mathbf{E}(f|\mathcal{B})$$

in the sense that $\alpha\mathbf{E}(f|\mathcal{B})$ is a conditional expectation for αf given $\mathcal{B}$.
(iii) If f and g are $\mathcal{A}$-measurable and μ-integrable functions, and $f = g$ a.e., then $\mathbf{E}(f|\mathcal{B}) = \mathbf{E}(g|\mathcal{B})$ a.e.
(iv) If f is an $\mathcal{A}$-measurable and μ-integrable function which is nonnegative a.e., then $\mathbf{E}(f|\mathcal{B})$ is nonnegative a.e.
(v) If f and g are $\mathcal{A}$-measurable and μ-integrable functions, and $f \leq g$ a.e., then $\mathbf{E}(f|\mathcal{B}) \leq \mathbf{E}(g|\mathcal{B})$ a.e.
(vi) If f is an $\mathcal{A}$-measurable and μ-integrable function, then $\mathbf{E}(f|\mathcal{B}) = f$ a.e. if and only if f is equal a.e. to a $\mathcal{B}$-measurable function.
(vii) If f is an $\mathcal{A}$-measurable and μ-integrable function, then $|\mathbf{E}(f|\mathcal{B})| \leq \mathbf{E}(|f||\mathcal{B})$ a.e.
(viii) If (f_n) is a nondecreasing sequence of nonnegative $\mathcal{A}$-measurable functions converging pointwise a.e. to f and f is μ-integrable, then $\mathbf{E}(f_n|\mathcal{B}) \to \mathbf{E}(f, \mathcal{B})$ a.e.

PROOF: (i)–(iii) and (vi) are straightforward and left to the reader. For (iv), if $\mathbf{E}(f|\mathcal{B})$ is negative on a set B of positive measure, then $\int_B \mathbf{E}(f|\mathcal{B})\,d\mu = \int_B f\,d\mu < 0$ (), contradicting that $f \geq 0$ a.e. (v) then follows from (iv), (i), and (ii). For (vii), we apply XVIII.1.9.1., since for any $B \in \mathcal{B}$,

$$\left|\int_B \mathbf{E}(f|\mathcal{B})\,d\mu\right| = \left|\int_B f\,d\mu\right| \leq \int_B |f|\,d\mu = \int_B \mathbf{E}(|f||\mathcal{B})\,d\mu\ .$$

For (viii), we have $\mathbf{E}(f_n|\mathcal{B}) \leq \mathbf{E}(f_{n+1}|\mathcal{B}) \leq \mathbf{E}(f,\mathcal{B})$ a.e. for each n by (v). Set $g = \sup_n \mathbf{E}(f_n|\mathcal{B})$. Then $g \leq \mathbf{E}(f|\mathcal{B})$ a.e. By the MCT, for each $B \in \mathcal{B}$ we have

$$\int_B \mathbf{E}(f_n|\mathcal{B})\,d\mu = \int_B f_n\,d\mu \to \int_B f\,d\mu = \int_B \mathbf{E}(f|\mathcal{B})\,d\mu$$

and since

$$\int_B E(f_n|\mathcal{B})\,d\mu \leq \int_B g\,d\mu \leq \int_B \mathbf{E}(f|\mathcal{B})\,d\mu$$

for each n and B, we conclude that

$$\int_B g\,d\mu = \int_B \mathbf{E}(f|\mathcal{B})\,d\mu\ , \quad \int_B [\mathbf{E}(f|\mathcal{B}) - g]\,d\mu = 0$$

for each B, and thus $\mathbf{E}(f|\mathcal{B}) = g$ a.e. 👍

XVIII.6.5.8. It is not true that $\mathbf{E}(fg|\mathcal{B}) = \mathbf{E}(f|\mathcal{B})\mathbf{E}(g|\mathcal{B})$ for general f and g, even if fg is integrable (which is not automatic). But we have an important partial version:

XVIII.6.5.9. PROPOSITION. Let $(X,\mathcal{A},\mu)$ be a probability measure space, and $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$. Let f be an $\mathcal{A}$-measurable and μ-integrable complex-valued function and g a $\mathcal{B}$-measurable complex-valued function such that fg is μ-integrable (this will be automatic if g is bounded). Then $\mathbf{E}(fg|\mathcal{B}) = g\mathbf{E}(f|\mathcal{B})$ a.e.

PROOF: If for $B \in \mathcal{B}$ we define

$$\nu(B) = \int_B f\,d\mu = \int_B \mathbf{E}(f|\mathcal{B})\,d\mu$$

then g is ν-integrable by XVII.9.5.4., and we have

$$\int_B g\,d\nu = \int_B fg\,d\mu$$

for every $B \in \mathcal{B}$. By another application of XVII.9.5.4., $g\mathbf{E}(f|\mathcal{B})$ is μ-integrable and

$$\int_B g\,d\nu = \int_B g\mathbf{E}(f|\mathcal{B})\,d\mu$$

for all $B \in \mathcal{B}$. Putting these together, we get

$$\int_B fg\,d\mu = \int_B g\mathbf{E}(f|\mathcal{B})\,d\mu$$

for all $B \in \mathcal{B}$. This means $g\mathbf{E}(f|\mathcal{B})$ is a conditional expectation for fg given $\mathcal{B}$. 👍

Successive Conditional Expectations

XVIII.6.5.10. Suppose $(X, \mathcal{A}, \mu)$ is a probability measure space, $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$, and $\mathcal{D}$ a sub-σ-algebra of $\mathcal{B}$. Then we have conditional expectations of $\mathcal{A}$-measurable and μ-integrable functions given both $\mathcal{B}$ and $\mathcal{D}$, and for $\mathcal{B}$-measurable and $\mu|_{\mathcal{B}}$-integrable functions given $\mathcal{D}$. How are these related? The relation turns out to be as nice as could be hoped:

XVIII.6.5.11. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a probability measure space, $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$, and $\mathcal{D}$ a sub-σ-algebra of $\mathcal{B}$. Then, for any $\mathcal{A}$-measurable and μ-integrable function f, we have

$$\mathbf{E}(f, \mathcal{D}) = \mathbf{E}(\mathbf{E}(f|\mathcal{B})|\mathcal{D})) \text{ a.e.}$$

PROOF: If $D \in \mathcal{D}$, we have

$$\int_D \mathbf{E}(\mathbf{E}(f|\mathcal{B})|\mathcal{D}))\, d\mu = \int_D \mathbf{E}(f|\mathcal{B})\, d\mu = \int_D f\, d\mu$$

since $\mathcal{D} \subseteq \mathcal{B}$, so $\mathbf{E}(\mathbf{E}(f|\mathcal{B})|\mathcal{D}))$ is a conditional expectation for f given $\mathcal{D}$. 👍

Reduction to Equivalence Classes

XVIII.6.5.12. The lack of uniqueness in conditional expectations can be nicely remedied. Recall that $\mathcal{L}^1_{\mathbb{R}}(X, \mathcal{A}, \mu)$ and $\mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \mu)$ are the spaces of $\mathcal{A}$-measurable and μ-integrable real- and complex-valued functions respectively. On each of these, define an equivalence relation by setting $f \sim g$ if $f = g$ a.e. (It is easy to check this is an equivalence relation.) Let $L^1_{\mathbb{R}}(X, \mathcal{A}, \mu)$ and $L^1_{\mathbb{C}}(X, \mathcal{A}, \mu)$ be the sets of equivalence classes. There are well-defined operations of addition and scalar multiplication on these sets, defined in the obvious way: $[f] + [g] = [f + g]$ and $\alpha[f] = [\alpha f]$, where $[f]$ denotes the equivalence class of f. These operations make $L^1_{\mathbb{R}}(X, \mathcal{A}, \mu)$ and $L^1_{\mathbb{C}}(X, \mathcal{A}, \mu)$ into real and complex vector spaces respectively. Although strictly speaking $L^1_{\mathbb{R}}(X, \mathcal{A}, \mu)$ is not a subset of $L^1_{\mathbb{C}}(X, \mathcal{A}, \mu)$, there is an obvious natural identification making $L^1_{\mathbb{R}}(X, \mathcal{A}, \mu)$ a real vector subspace of $L^1_{\mathbb{C}}(X, \mathcal{A}, \mu)$.

XVIII.6.5.13. If $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$, then $L^1_{\mathbb{F}}(X, \mathcal{B}, \mu|_{\mathcal{B}})$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) is technically not a subset of $L^1_{\mathbb{F}}(X, \mathcal{A}, \mu)$, but it can be naturally identified with the set of equivalence classes of $\mathcal{B}$-measurable and μ-integrable functions, a vector subspace.

XVIII.6.5.14. Now suppose μ is a probability measure. The (not well-defined) map $f \mapsto \mathbf{E}(f|\mathcal{B})$ gives a well-defined function $\mathbf{E}$ from $L^1_{\mathbb{F}}(X, \mathcal{A}, \mu)$ to the subspace $L^1_{\mathbb{F}}(X, \mathcal{B}, \mu|_{\mathcal{B}})$ by XVIII.6.5.7.(iii). This function is linear by XVIII.6.5.7.(i)–(ii), and has other properties from the other parts of XVIII.6.5.7.. From (vi), $\mathbf{E}$ is *idempotent*, i.e. $\mathbf{E}([\mathbf{E}([f])]) = \mathbf{E}([f])$.

XVIII.6.5.15. We can also define $L^\infty_{\mathbb{F}}(X, \mathcal{A}, \mu)$ to be the set of equivalence classes of bounded $\mathcal{A}$-measurable functions. Since μ is finite, a bounded measurable function is integrable, so we may identify $L^\infty_{\mathbb{F}}(X, \mathcal{A}, \mu)$ with a subset of $L^1_{\mathbb{F}}(X, \mathcal{A}, \mu)$, which is clearly a vector subspace. $L^\infty_{\mathbb{F}}(X, \mathcal{A}, \mu)$ also has a multiplication (pointwise product) making it into an $\mathbb{F}$-algebra, and $L^1_{\mathbb{F}}(X, \mathcal{A}, \mu)$ is a module over $L^\infty_{\mathbb{F}}(X, \mathcal{A}, \mu)$ ().

If $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$, then $L^\infty_{\mathbb{F}}(X, \mathcal{B}, \mu|_{\mathcal{B}})$ can be identified with a subset (subalgebra) of $L^\infty_{\mathbb{F}}(X, \mathcal{A}, \mu)$, so $L^1_{\mathbb{F}}(X, \mathcal{A}, \mu)$ can be regarded as a module over $L^\infty_{\mathbb{F}}(X, \mathcal{B}, \mu|_{\mathcal{B}})$.

XVIII.6.5.16. If $f \in L^\infty_{\mathbb{F}}(X, \mathcal{A}, \mu)$, then $|f| \leq M$ for some M which can be regarded as a constant function: thus

$$|\mathbf{E}(f|\mathcal{B})| \leq \mathbf{E}(|f||\mathcal{B}) \leq \mathbf{E}(M|\mathcal{B}) = M$$

by various parts of XVIII.6.5.7.. Thus $\mathbf{E}$ sends $L^\infty_{\mathbb{F}}(X, \mathcal{A}, \mu)$ into $L^\infty_{\mathbb{F}}(X, \mathcal{B}, \mu|_{\mathcal{B}})$. Also, by XVIII.6.5.9., the map $\mathbf{E}$ from $L^1_{\mathbb{F}}(X, \mathcal{A}, \mu)$ to $L^1_{\mathbb{F}}(X, \mathcal{B}, \mu|_{\mathcal{B}})$ is a module-homomorphism when they are regarded as $L^\infty_{\mathbb{F}}(X, \mathcal{B}, \mu|_{\mathcal{B}})$-modules.

See XX.15.8. for more on conditional expectations. There is an important generalization to operator algebras; see e.g. [Bla06, II.6.10].

XVIII.6.6. Change of Variable

The Radon-Nikodým Theorem gives an extremely general change-of-variable theorem for integrals. There are two versions which are easily stated and proved; however, relating them to classical change-of-variable theorems and applying them can be somewhat complicated.

Integration for Push-Forwards

XVIII.6.6.1. We first make a simple observation. Let $(X, \mathcal{A}, \mu)$ be a measure space, $(Y, \mathcal{B})$ a measurable space, and $\phi : X \to Y$ an $(\mathcal{A}, \mathcal{B})$-measurable function. Then () we may push μ forward to a measure $\phi_*\mu$ on $(Y, \mathcal{B})$ defined by $\phi_*\mu(B) = \mu(\phi^{-1}(B))$.

XVIII.6.6.2. Proposition. Let f be either a $\phi_*\mu$-integrable function or a nonnegative $\mathcal{B}$-measurable function on Y. Then $f \circ \phi$ is μ-integrable (if f is $\phi_*\mu$-integrable) or nonnegative and $\mathcal{A}$-measurable (if f is nonnegative), and, for any $E \in \mathcal{B}$,

$$\int_E f \, d(\phi_*\mu) = \int_{\phi^{-1}(E)} (f \circ \phi) \, d\mu \ .$$

Proof: This is a standard bootstrap argument. We may assume $E = Y$ without loss of generality. We have that $f \circ \phi$ is $\mathcal{A}$-measurable. First let $f = \chi_B$ for some $B \in \mathcal{B}$. Then

$$\int_Y \chi_B \, d(\phi_*\mu) = \phi_*\mu(B) = \mu(\phi^{-1}(B)) = \int_X \chi_{\phi^{-1}(B)} \, d\mu = \int_X (\chi_B \circ \phi) \, d\mu \ .$$

The formula thus holds for indicator functions; hence for simple functions by linearity, for nonnegative measurable functions by the Monotone Convergence Theorem, and then for $(\phi_*\mu)$-integrable functions again by linearity. 👍

Converting Integrals

If in the above setting ν is a measure on $(Y, \mathcal{B})$, what we want to do is convert an integral over X with respect to μ into one over Y with respect to ν or (more importantly for applications) vice versa. There are two easy ways to do this using the Radon-Nikodým Theorem. We need modest additional hypotheses, slightly different in the two cases, concerning preservation of null sets under ϕ.

XVIII.6.6.3. Definition. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, and $\phi : X \to Y$ an $(\mathcal{A}, \mathcal{B})$-measurable function.

(i) The function ϕ *preserves null sets backward* if, whenever $B \in \mathcal{B}$ with $\nu(B) = 0$, then $\mu(\phi^{-1}(B)) = 0$.

(ii) The function ϕ *preserves null sets forward* if, whenever $B \in \mathcal{B}$ and $\mu(\phi^{-1}(B)) = 0$, then $\nu(B) = 0$.

XVIII.6.6.4. The function ϕ preserves null sets backward if and only if $\phi_*\mu \ll \nu$, and preserves null sets forward if and only if $\nu \ll \phi_*\mu$. A function ϕ which preserves null sets forward and whose range is in $\mathcal{B}$ must be "almost surjective" in the sense that $\nu(Y \setminus \phi(X)) = 0$. Similarly, if ν is purely nonatomic and singletons are in $\mathcal{B}$, if ϕ preserves null sets backward it is "almost injective" in the sense that the preimage of every point is a null set. (These conditions are far from sufficient.)

There is a closely related property (cf. XIX.5.1.2.):

XVIII.6.6.5. Definition. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, and $\phi : X \to Y$ an $(\mathcal{A}, \mathcal{B})$-measurable function. The function ϕ *has property (N)* if

(i) For every $A \in \mathcal{A}$, $\phi(A) \in \mathcal{B}$.

(ii) Whenever $A \in \mathcal{A}$ with $\mu(A) = 0$, then $\nu(\phi(A)) = 0$.

XVIII.6.6.6. If X and Y are open sets in Euclidean space with Lebesgue measure (this can be relaxed), and ϕ is continuous, then property (i) is automatic from property (ii): any Lebesgue measurable set E is the union of a Borel set and a null set; the null set is a subset of a null Borel set, so its image $\phi(E)$ is the union of an analytic set () and a subset of an analytic null set, and is thus Lebesgue measurable (). (But see ().)

XVIII.6.6.7. Proposition. Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, and $\phi : X \to Y$ an $(\mathcal{A}, \mathcal{B})$-measurable function. If ϕ has property (N) and is almost surjective in the sense of XVIII.6.6.4., then ϕ preserves null sets forward.

Proof: Let $C = Y \setminus \phi(X)$. By hypothesis $C \in \mathcal{B}$ and $\nu(C) = 0$. If $B \in \mathcal{B}$, and $\mu(\phi^{-1}(B)) = 0$, we have $\nu(\phi(\phi^{-1}(B))) = 0$, and $B \subseteq \phi(\phi^{-1}(B)) \cup C$, a ν-null set. 👍

XVIII.6.6.8. THEOREM. [FORWARD CHANGE OF VARIABLE IN INTEGRALS] Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, with ν σ-finite, and $\phi : X \to Y$ an $(\mathcal{A}, \mathcal{B})$-measurable function which preserves null sets backward. Then, if f is a measurable function on Y such that either f is nonnegative or $f \circ \phi$ is μ-integrable, then for every $E \in \mathcal{B}$,

$$\int_{\phi^{-1}(E)} (f \circ \phi)\, d\mu = \int_E \left(f \cdot \frac{d(\phi_* \mu)}{d\nu} \right) d\nu$$

where $\frac{d(\phi_* \mu)}{d\nu}$ is the Radon-Nikodým derivative (in particular, the integrand on the right is ν-integrable if $f \circ \phi$ is μ-integrable).

PROOF: Since $\phi_* \mu \ll \nu$, by XVIII.6.6.2. and the Radon-Nikodým Theorem we have

$$\int_{\phi^{-1}(E)} (f \circ \phi)\, d\mu = \int_E f\, d(\phi_* \mu) = \int_E \left(f \cdot \frac{d(\phi_* \mu)}{d\nu} \right) d\nu \ .$$



XVIII.6.6.9. THEOREM. [BACKWARD CHANGE OF VARIABLE IN INTEGRALS] Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be measure spaces, with μ σ-finite, and $\phi : X \to Y$ an $(\mathcal{A}, \mathcal{B})$-measurable function which preserves null sets forward. Then, if f is a measurable function on Y such that f is either nonnegative or ν-integrable, then for every $E \in \mathcal{B}$,

$$\int_E f\, d\nu = \int_{\phi^{-1}(E)} \left[(f \circ \phi) \cdot \left(\frac{d\nu}{d(\phi_* \mu)} \circ \phi \right) \right] d\mu$$

where $\frac{d\nu}{d(\phi_* \mu)}$ is the Radon-Nikodým derivative (in particular, the integrand on the right is μ-integrable if f is ν-integrable).

PROOF: Since $\nu \ll \phi_* \mu$, by XVIII.6.6.2. and the Radon-Nikodým Theorem we have

$$\int_E f\, d\mu = \int_E \left(f \cdot \frac{d\nu}{d(\phi_* \mu)} \right) d(\phi_* \mu) = \int_{\phi^{-1}(E)} \left[(f \circ \phi) \cdot \left(\frac{d\nu}{d(\phi_* \mu)} \circ \phi \right) \right] d\mu$$

(to apply the Radon-Nikodým Theorem, note that $\phi_* \mu$ is σ-finite).



XVIII.6.6.10. The names "Forward/Backward Change of Variable" are nonstandard and rather arbitrary, but nicely describe the direction the integral is converted (but note that in each the ϕ must preserve null sets in the *opposite* direction). The Backward Change of Variable is generally more useful than the forward, since the integral on the left can be typically taken to be a given one to compute in a problem, and the Backward Theorem is used to convert it to a new integral over a subset of X. The integral on the right appears more complicated, but in practice it sometimes has a nice form which can be evaluated. However, it can be difficult to explicitly compute the Radon-Nikodým derivative, and it will take some work to write it in a more useful form in nice cases.



XVIII.6.7. Exercises

XVIII.6.7.1. What is the difficulty in extending the Radon-Nikodým Theorem to the case where μ is merely decomposable, using the argument in the last part of the proof of XVIII.6.2.1.? (See XVIII.6.1.5..) Give a valid statement in this case; cf. [HS75, 19.27].

XVIII.6.7.2. Let ν be a complex measure on $(X, \mathcal{A})$.

(a) Use the complex Radon-Nikodým Theorem to show that there is a $u \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$ such that

$$\nu(A) = \int_A u \, d|\nu|$$

for all $A \in \mathcal{A}$.

(b) Use Exercise XVII.9.5.1.(b) to show that $|u| = 1$ a.e.

(c) Let $f \in \mathcal{L}^1_{\mathbb{C}}(X, \mathcal{A}, \nu)$, and define a complex measure σ on $(X, \mathcal{A})$ by $\sigma(A) = \int_A f \, d\nu$ (). Show that

$$|\sigma|(A) = \int_A |f| \, d|\nu|$$

for $A \in \mathcal{A}$.

XVIII.6.7.3. Use Theorem XVIII.6.5.6. to give a proof that if $(X, \mathcal{A}, \mu)$ is a σ-finite measure space with completion $(X, \bar{\mathcal{A}}, \bar{\mu})$, and f is a real-valued $\bar{\mathcal{A}}$-measurable function, then there is an $\mathcal{A}$-measurable function g with $g = f$ $\bar{\mu}$-a.e. [The case where μ is finite and f is integrable is immediate. In the general case, decompose X into sets of finite measure and write f as a pointwise limit of integrable functions.]

Chapter XIX

Advanced Analysis on Euclidean Spaces

In this chapter, we collect together results connecting continuity, differentiation, and integration of functions on $\mathbb{R}$ and $\mathbb{R}^n$. The integration used is Lebesgue integration and sometimes integration with respect to more general measures on the Borel sets. In particular, we examine several natural classes of functions and their properties.

A common theme in this chapter is the principle that "what happens on null sets can frequently be ignored" (i.e. that null sets are frequently "negligible"). We will thus often think of two functions as being essentially the same if they agree almost everywhere, and many phenomena only hold almost everywhere. We can even relax the basic notions of limits, continuity, and differentiability to "approximate" ones where "small" sets are ignored. (However, considerable work is sometimes needed to justify ignoring null sets, and it is not always possible.)

XIX.1. Borel Measures on $\mathbb{R}$ and Cumulative Distribution Functions

A *finite Borel measure* on $\mathbb{R}$ is a finite measure defined on $(\mathbb{R}, \mathcal{B})$. We will also study finite Borel measures on intervals in $\mathbb{R}$, and defined in the same way, and σ-finite Borel measures on $\mathbb{R}$ which are finite on bounded sets (*positive Radon measures*). Such measures are important in many contexts, and especially in probability theory.

XIX.1.1. Cumulative Distribution Functions

XIX.1.1.1. We would like to have a way of describing the finite Borel measures on $\mathbb{R}$ (or an interval in $\mathbb{R}$). If μ is a finite Borel measure on $\mathbb{R}$, for each $x \in \mathbb{R}$ we can compute $\mu((-\infty, x])$. It turns out that these numbers are sufficient to recover μ by the Carathéodory extension process (XIX.1.1.4.).

This data is conveniently described by a function F_μ from $\mathbb{R}$ to $\mathbb{R}$, setting $F_\mu(x) = \mu((-\infty, x])$. Then if $a, b \in \mathbb{R}$, $a < b$, we have $\mu((a, b]) = F_\mu(b) - F_\mu(a)$.

XIX.1.1.2. Definition. F_μ is called the *cumulative distribution function* (or *cdf*) of μ.

XIX.1.1.3. Proposition. The function F_μ has the following properties:

(i) F_μ is increasing (nondecreasing).

(ii) F_μ is continuous from the right.

(iii) $\lim_{x \to -\infty} F_\mu(x) = 0$.

(iv) F_μ is bounded, and $\sup_{x \in \mathbb{R}} F_\mu(x) = \lim_{x \to +\infty} F_\mu(x) = \mu(\mathbb{R})$.

Proof: (i) is obvious, and (iii) and (iv) follow immediately from () since $\cup_{n \in \mathbb{N}}(-\infty, n] = \mathbb{R}$ and $\cap_{n \in \mathbb{N}}(-\infty, -n] = \emptyset$. (ii) also follows from () since for any $x \in \mathbb{R}$, $(-\infty, x] = \cap_{n \in \mathbb{N}} \left(-\infty, x + \frac{1}{n}\right]$. 👍

The choice to make cdf's continuous from the right is simply an arbitrary convention. We could have equally well made them continuous from the left by defining $F_\mu(x) = \mu((-\infty, x))$.

Conversely, a function satisfying (i)–(iv) is the cdf of a unique finite Borel measure:

XIX.1.1.4. Theorem. Let $F : \mathbb{R} \to \mathbb{R}$ be a function with the following properties:

(i) F is increasing (nondecreasing).

(ii) F is continuous from the right.

(iii) $\lim_{x \to -\infty} F(x) = 0$.

(iv) F is bounded.

Then there is a unique finite Borel measure μ on $\mathbb{R}$ with $F = F_\mu$, and $\mu(\mathbb{R}) = \sup_{x\in\mathbb{R}} F(x) = \lim_{x\to+\infty} F(x)$.

PROOF: Let $\mathcal{S}$ be the collection of all intervals of the form $(a, b]$, for $a, b \in \mathbb{R}$, $(-\infty, b]$ for $b \in \mathbb{R}$, and $(a, +\infty)$ for $a \in \mathbb{R}$, along with $\mathbb{R}$ and $\emptyset$. Set $M = \sup_{x\in\mathbb{R}} F(x) = \lim_{x\to+\infty} F(x)$. Define

$$\iota((-\infty, b]) = F(b)$$

$$\iota((a, b]) = F(b) - F(a)$$

$$\iota((a, +\infty)) = M - F(a)$$

$$\iota(\mathbb{R}) = M$$

$$\iota(\emptyset) = 0.$$

It is easily verified that $\mathcal{S}$ is a semialgebra () and that ι is a premeasure (XVII.4.5.2.) on $\mathcal{S}$. Thus there is a unique measure μ on the σ-algebra $\sigma(\mathcal{S}) = \mathcal{B}$ extending ι by the Extension Theorem (XVII.4.5.8.), and obviously $F = F_\mu$. 👍

XIX.1.1.5. The finite Borel measure μ for which $F = F_\mu$ is called the *Lebesgue-Stieltjes measure* of (or corresponding to) F. If g is a μ-integrable function, $\int g\, d\mu$ is sometimes written $\int g\, dF$ and called the *Lebesgue-Stieltjes integral* of g with respect to F.

XIX.1.1.6. DEFINITION. Let F be a nondecreasing function from $\mathbb{R}$ to $\mathbb{R}$. Then F is continuous except possibly at a countable number of jump discontinuities at $\{x_n : n \in \mathbb{N}\}$. If $E \subseteq \mathbb{R}$, set

$$F^{\rightarrow}(E) = F(E) \cup \{[F(x_n-), F(x_n+)] : x_n \in E\} .$$

$F^{\rightarrow}(E)$ is the "image" of E, saturated by adding intervals for jump discontinuities of F as in ().

XIX.1.1.7. THEOREM. Let μ be a finite Borel measure on $\mathbb{R}$, with cdf F_μ. If E is any Borel subset of $\mathbb{R}$, then $F_\mu^{\rightarrow}(E)$ is a Borel set, and $\mu(E) = \lambda(F_\mu^{\rightarrow}(E))$.

Note that this result is false, even meaningless, if E is merely a Lebesgue measurable subset of $\mathbb{R}$; $F_\mu^{\rightarrow}(E)$ need not be Lebesgue measurable (), and $\mu(E)$ may not be defined.

XIX.1.2. Density Functions and Absolute Continuity

XIX.1.2.1. For some, but not all, finite Borel measures, there is a corresponding density function. If μ is a finite Borel measure on $\mathbb{R}$, and there is a Borel measurable function ϕ from $\mathbb{R}$ to $\mathbb{R}$ (necessarily nonnegative and integrable) such that $\mu(E) = \int_E \phi\, d\lambda$ for all Borel sets $E \subseteq \mathbb{R}$, then ϕ is called a *density function* for μ. Two density functions for μ must agree almost everywhere (with respect to λ), so by slight abuse of terminology we may refer to *the* density function of μ. Any nonnegative integrable function ϕ on $\mathbb{R}$ is the density function of a unique measure μ_ϕ, defined by $\mu_\phi(E) = \int_E \phi\, d\lambda$ for $E \subseteq \mathbb{R}$ Borel.

The function ϕ is actually the "derivative" of the function F_{μ_ϕ}. This is a precise relationship in nice cases:

XIX.1.2.2. EXAMPLE. Let ϕ be a nonnegative continuous function on $\mathbb{R}$ with $\int_{-\infty}^{\infty} \phi(t)\,dt$ finite. For a Borel set $E \subseteq \mathbb{R}$, set $\mu_\phi(E) = \int_E \phi\,d\lambda$ as above. Then, for $a, b \in \mathbb{R}$, $a < b$, we have

$$F_{\mu_\phi}(b) - F_{\mu_\phi}(a) = \mu_\phi((a,b]) = \int_a^b \phi(t)\,dt$$

and thus by the Fundamental Theorem of Calculus 1, F_{μ_ϕ} is an antiderivative for ϕ, i.e. F_{μ_ϕ} is differentiable and $F'_{\mu_\phi} = \phi$.

This example generalizes to arbitrary density functions ϕ:

XIX.1.2.3. THEOREM. Let $\phi \in \mathcal{L}^1_+(\mathbb{R})$, and F_{μ_ϕ} the cdf of the corresponding measure μ_ϕ with density function ϕ. Then F_{μ_ϕ} is continuous, differentiable λ-a.e., and $F'_{\mu_\phi} = \phi$ λ-a.e.

This theorem will be proved in XIX.5.2.6..

A finite Borel measure need not have a density function. Its cdf need not even be continuous:

XIX.1.2.4. EXAMPLE. Let δ_0 be the Dirac measure (unit point mass) at 0. Then

$$F_{\delta_0}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ \\ 0 & \text{if } x < 0 \end{cases}.$$

XIX.1.2.5. In fact, the cdf of a finite Borel measure μ is a continuous function if and only if μ is a continuous measure: in general, for $x \in \mathbb{R}$,

$$F_\mu(x+) - F_\mu(x-) = F_\mu(x) - F_\mu(x-) = \mu(\{x\})$$

where, as usual, $F_\mu(x+) = \lim_{y \to x+} F_\mu(y)$ and $F_\mu(x-) = \lim_{y \to x-} F_\mu(y)$. (This fact was crucial motivation for the choice of the term "continuous" for a measure with no atoms.)

But even a continuous μ need not have a density function:

XIX.1.2.6. EXAMPLE. Let F be the Cantor function on $[0,1]$ (), extended to be 0 on $(-\infty, 0]$ and 1 on $[1, \infty)$. F satisfies the conditions of XIX.1.1.4., and thus it is F_μ for a unique continuous finite Borel measure μ on $\mathbb{R}$ with $\mu(\mathbb{R}) = \mu([0,1]) = 1$. If μ had a density function ϕ, then ϕ would have to be equal to F' λ-a.e. But $F' = 0$ λ-a.e., so ϕ would have to be 0 λ-a.e. and μ would be 0. Thus μ has no density function.

This μ is not absolutely continuous with respect to λ. For example, if K is the Cantor set, $\lambda(K) = 0$, but $\mu(K) = 1$; this follows from XIX.1.1.7., but can also easily be seen directly by noting that $\mu(I) = 0$ for any interval I in the complement of K, since F is constant on I. This μ is actually the push-forward of the measure of () under the natural map from $\{0,1\}^{\mathbb{N}}$ to $K \subseteq \mathbb{R}$.

XIX.1.2.7. So which finite Borel measures have density functions? It follows immediately from () that for μ to have a density function, it is necessary for μ to be absolutely continuous with respect to λ. This condition also turns out to be sufficient:

XIX.1.2.8. THEOREM. A finite Borel measure has a density function if and only if it is absolutely continuous with respect to λ.

The "if" part of this theorem is a special case of the Radon-Nikodým Theorem (). See XIX.5.2.6. for a proof.

XIX.1.2.9. COROLLARY. A finite Borel measure μ is absolutely continuous with respect to λ if and only if its cdf F_μ is absolutely continuous in the sense of XIX.5.2.1..

In fact, this result is the reason behind the choice of the term "absolutely continuous" for the functions of XIX.5.2.1..

XIX.1.2.10. Even if μ does not have a density function, it has a density in a generalized sense (), which is the "derivative" of F_μ. Thus, for example, we informally think of the "Dirac δ-function" as the "derivative" of the function $\chi_{[0,\infty)}$. We can think of the actual measure μ as being identified with F_μ, in which case the "density" of μ is regarded as the "derivative" of μ, or we can regard μ as being identified with its "density", with F_μ regarded as the "antiderivative" of μ. The second point of view is more common.

XIX.1.3. Finite Borel Measures on Intervals

XIX.1.3.1. If I is an interval in $\mathbb{R}$, then finite Borel measures on I can be thought of as finite Borel measures on $\mathbb{R}$ giving measure 0 to $\mathbb{R} \setminus I$. The cdf of such a measure is constant on $\mathbb{R} \setminus I$.

If I is bounded above, and $b = \sup(I)$, then the cdf F_μ of a finite Borel measure μ on I must satisfy $F_\mu(b) = \mu(I)$ $(= \mu(\mathbb{R}))$. If $b \notin I$, then F_μ is continuous at b; but if $b \in I$, then F_μ can have a discontinuity at b if $\mu(\{b\}) > 0$. In any event, $\mu(I) = \sup_{x \in I} F_\mu(x)$.

If I is bounded below with $a = \inf(I)$, then F_μ is identically 0 on $(-\infty, a)$, but if $a \in I$, then $F_\mu(a) = \mu(\{a\})$, so $\inf_{x \in I} F_\mu(x)$ is not necessarily 0. But if $a \notin I$, then we must have $\inf_{x \in I} F_\mu(x) = 0$.

The Mean Value Theorem for Integrals Revisited

Here is perhaps the most general version of the Mean Value Theorem for Integrals IX.2.7.14.:

XIX.1.3.2. THEOREM. [MEAN VALUE THEOREM FOR INTEGRALS, VERSION 4] Let f be a continuous function on $[a, b]$, and μ a finite Borel measure on $[a, b]$. Then there is a $c \in [a, b]$ for which

$$f(c)\mu([a, b]) = \int_{[a,b]} f \, d\mu \ .$$

PROOF: If $\mu = 0$ there is nothing to prove. If $\mu([a,b]) > 0$, then, if M and m are the maximum and minimum values of f on $[a,b]$, we have

$$m\mu([a,b]) \leq \int_{[a,b]} f\, d\mu \leq M\mu([a,b])$$

and thus by the Intermediate Value Theorem there is a $c \in [a,b]$ with

$$f(c) = \frac{\int_{[a,b]} f\, d\mu}{\mu([a,b])} \ .$$

👍

An apparently more general version is:

XIX.1.3.3. COROLLARY. [MEAN VALUE THEOREM FOR INTEGRALS, VERSION 5] Let f be a continuous function on $[a,b]$, μ a finite Borel measure on $[a,b]$, and g a nonnegative μ-integrable function on $[a,b]$. Then fg is μ-integrable on $[a,b]$, and there is a $c \in [a,b]$ for which

$$f(c) \int_{[a,b]} g\, d\mu = \int_{[a,b]} fg\, d\mu \ .$$

PROOF: If E is a Borel subset of $[a,b]$, set $\nu(E) = \int_E g\, d\mu$. Then ν is again a finite Borel measure on $[a,b]$. We have that fg is μ-integrable on $[a,b]$ and that $\int_{[a,b]} fg\, d\mu = \int_{[a,b]} f\, d\nu$ by XVIII.1.9.2.. Apply XIX.1.3.2. with ν in place of μ. 👍

XIX.1.4. Stieltjes Integrals

XIX.1.5. Unbounded Distribution Functions and Positive Radon Measures

XIX.1.5.1. The requirement that $\lim_{x \to -\infty} F(x) = 0$ for a cdf F is a convenient normalization, but is not really necessary. If μ is a finite Borel measure on $\mathbb{R}$, c is any constant, and $G(x) = F_\mu(x) + c$, then G can be used just as well as F_μ to recover μ, since for any finite interval $(a,b]$, we have

$$\mu((a,b]) = F_\mu(b) - F_\mu(a) = G(b) - G(a) \ .$$

If F_μ is regarded as an "antiderivative" of μ, then the general "antiderivative" of μ is a constant translate of F_μ, and one antiderivative is as good as another for most purposes.

XIX.1.5.2. Taking this observation one step farther, if F is any nondecreasing, right-continuous function from $\mathbb{R}$ to $\mathbb{R}$, not necessarily bounded above or below, then F has a Lebesgue-Stieltjes measure μ defined by the Carathéodory extension process using $\mu((a,b]) = F(b) - F(a)$ for $a, b \in \mathbb{R}$, $a < b$, and F can be regarded as a cdf for μ. This μ is a positive Radon measure (finite on bounded sets) which is always σ-finite, but not finite if F is unbounded. Two nondecreasing right-continuous functions define the same Lebesgue-Stieltjes measure if and only if they differ by a constant. The function $F(x) = x$ is the cdf of Lebesgue measure itself.

XIX.1.5.3. If F is such a cdf, and $a, b \in \mathbb{R}$, then F can be truncated at a and b to a function G by setting

$$G(x) = \begin{cases} F(a-) & \text{if } x < a \\ F(x) & \text{if } a \leq x \leq b \\ F(b) & \text{if } x > b \end{cases} .$$

Then G is a cdf for the measure $\mu|_{[a,b]}$, which is a finite Borel measure. Thus the study of positive Radon measures on $\mathbb{R}$ and their cdf's can be effectively reduced to the theory of finite Borel measures.

XIX.1.5.4. The theory can be further extended by allowing F to take extended real values. Then $\{x : F(x) \in \mathbb{R}\}$ is an interval I, and F may be thought of as a positive Radon measure on I, i.e. finite on closed subintervals in the interior of I.

XIX.1.5.5. Example. Let $F(x) = \tan x$ for $-\pi/2 < x < \pi/2$. Set $F = -\infty$ on $(-\infty, -\pi/2]$ and $+\infty$ on $[\pi/2, \infty)$. Then F is the cdf of a σ-finite Borel measure μ on $(-\pi/2, \pi/2)$ which is finite on any closed subinterval. This μ is absolutely continuous with respect to $\lambda|_{(-\pi/2,\pi/2)}$, and has density function $g(x) = \sec^2 x$.

XIX.1.6. Signed Measures and Nonincreasing Distribution Functions

Using the theory of signed measures, many of the results of this section generalize to cdf's which are not increasing. There are considerable complications in this case, but also important applications. Details are found in ().

XIX.1.7. Borel Measures and Distribution Functions on $\mathbb{R}^n$

XIX.1.8. Integration By Parts

We may combine the above notions with the Fubini-Tonelli Theorem to obtain a very general version of Integration by Parts.

XIX.1.8.1. Let $[a, b]$ be a closed bounded interval. Let μ and ν be finite Borel measures on $[a, b]$ with cdf's F and G respectively. Set $I = (a, b]$. We have

$$(\mu \times \nu)(I^2) = \mu(I)\nu(I) = [F(b) - F(a)][G(b) - G(a) = F(b)G(b) + F(a)G(a) - F(a)G(b) - F(b)G(a) .$$

Set $A = \{(x, y) \in I^2 : x \leq y\}$ and $B = \{(x, y) \in I^2 : y \leq x\}$. Then $A \cup B = I^2$ and $A \cap B = \Delta = \{(x, x) : a < x \leq b\}$, the diagonal. By Tonelli's Theorem we have

$$(\mu \times \nu)(A) = \int_I \mu((a, y])\, d\nu(y) = \int_I (F(y)) - F(a))\, d\nu(y) = \int_I F(y)\, d\nu(y) - F(a)[G(b) - G(a)]$$

$$(\mu \times \nu)(B) = \int_I \nu((a, x])\, d\mu(x) = \int_I (G(x) - G(a))\, d\mu(x) = \int_I G(x)\, d\mu(x) - G(a)[F(b) - F(a)]$$

so we have

$$(\mu \times \nu)(A) + (\mu \times \nu)(B) = (\mu \times \nu)(I^2) - (\mu \times \nu)(\Delta)$$

$$\int_I F\,d\nu - F(a)G(b) + F(a)G(a) + \int_I G\,d\mu - G(a)F(b) + G(a)F(a)$$
$$= F(b)G(b) + F(a)G(a) - F(a)G(b) - F(b)G(a) - (\mu\times\nu)(\Delta)$$
$$\int_I F\,d\nu + \int_I G\,d\mu = F(b)G(b) - F(a)G(a) - (\mu\times\nu)(\Delta)\ .$$

By Tonelli's Theorem we have

$$(\mu\times\nu)(\Delta) = \int_I \nu(\{x\})\,d\mu(x) = \sum_x \mu(x)\nu(x)$$

(sum over the countably many x with $\mu(x)\nu(x) > 0$). So $(\mu\times\nu)(\Delta) = 0$ if and only if μ and ν do not both give a positive measure to the same singleton, i.e. F and G have no common discontinuities. In this case we obtain the Integration by Parts formula, i.e. we have proved:

XIX.1.8.2. Theorem. [Integration by Parts, General Version] Let $[a,b]$ be a closed bounded interval. Let μ and ν be finite Borel measures on $[a,b]$ with cdf's F and G respectively. If F and G have no common discontinuities, then

$$\int_{(a,b]} G\,d\mu = F(b)G(b) - F(a)G(a) - \int_{(a,b]} F\,d\nu\ .$$

XIX.1.8.3. Using the Stieltjes integral notation, the last formula can be put into more familiar form: if we write dF for $d\mu$ and dG for $d\nu$, we have

$$\int_{(a,b]} G\,dF = F(b)G(b) - F(a)G(a) - \int_{(a,b]} F\,dG\ .$$

XIX.1.8.4. If μ and ν are absolutely continuous with respect to Lebesgue measure, with density functions f and g respectively, then F and G are absolutely continuous on $[a,b]$ and $F' = f$ a.e., $G' = g$ a.e., and the formula becomes

$$\int_a^b fG\,d\lambda = F(b)G(b) - F(a)G(a) - \int_a^b gF\,d\lambda\ .$$

See XIX.5.5.2. for another derivation of this case not requiring Tonelli's Theorem.

XIX.1.8.5. Various further generalizations are possible:

The result extends to unbounded intervals and Radon measures.

The result extends to signed measures.

The result works on regions in $\mathbb{R}^n$.

The straightforward details are left to the interested reader. See () for another general version.

XIX.2. Monotone Functions

In this section, we describe the properties of monotone functions.

XIX.2.1. Definitions and Basics

XIX.2.1.1. DEFINITION. Let I be an interval. A function $f : I \to \mathbb{R}$ is *nondecreasing* on I if $f(x) \leq f(y)$ whenever $x, y \in I$ and $x < y$, and *nonincreasing* on I if $f(x) \geq f(y)$ whenever $x, y \in I$ and $x < y$. A function is *monotone* on I if it is either nondecreasing or nonincreasing on I.

XIX.2.1.2. A nondecreasing function is sometimes called *increasing*, but this term is somewhat ambiguous. If $f(x) < f(y)$ whenever $x, y \in I$, $x < y$, we will say that f is *strictly increasing* on I. A strictly increasing function is, of course, nondecreasing. Strictly decreasing functions are defined analogously. A function is *strictly monotone* on I if it is strictly increasing or strictly decreasing on I. "Increasing" could potentially mean either "nondecreasing" or "strictly increasing," and different authors use both meanings. We will avoid using the terms "increasing" and "decreasing" with regard to functions on intervals.

XIX.2.1.3. We will work with monotone functions on a closed bounded interval $[a, b]$ for simplicity. Analogous results hold for monotone functions on intervals which are not closed and/or not bounded, but there are some technical complications (e.g. such functions need not be bounded). We will also concentrate on the case of nondecreasing functions; the nonincreasing case is essentially identical with only a few notational changes needed.

The first result is almost obvious.

XIX.2.1.4. PROPOSITION. Let f be a monotone function on $[a, b]$. Then f is bounded on $[a, b]$.

PROOF: If f is nondecreasing on $[a, b]$, then $f(a) \leq f(x) \leq f(b)$ for all $x \in [a, b]$. The argument for nonincreasing functions is almost identical. 👍

One of the most important facts about monotone functions is that they are "almost continuous", and that the only possible discontinuities are of a very simple form.

XIX.2.1.5. DEFINITION. Let f be a function defined in a neighborhood of $c \in \mathbb{R}$, and $h > 0$. Then f has a *jump discontinuity of height* h at c if $\lim_{x \to c+} f(x)$ and $\lim_{x \to c-} f(x)$ both exist and

$$\left| \left[\lim_{x \to c+} f(x) \right] - \left[\lim_{x \to c-} f(x) \right] \right| = h$$

(One could also keep track of the sign of the jump by removing the absolute values and allowing $h < 0$.)

XIX.2.1.6. If $[\lim_{x \to c+} f(x)]$ exists, we will denote it by $f(c+)$. Similarly, if $[\lim_{x \to c-} f(x)]$ exists, we will denote it by $f(c-)$. We will extend this notation and the notion of a jump discontinuity at an endpoint of the domain of a function f: if a is the left endpoint of the domain, set $f(a-) = f(a)$, and at the right endpoint b set $f(b+) = f(b)$.

XIX.2.1.7. PROPOSITION. Let f be a monotone function on an interval I, and $c \in I$. Then $f(c+)$ and $f(c-)$ both exist, and $f(c)$ satisfies

$f(c-) \le f(c) \le f(c+)$ if f is nondecreasing on I.

$f(c-) \ge f(c) \ge f(c+)$ if f is nonincreasing on I.

If $d \in I$ and $c < d$, and f is nondecreasing [*resp.* nonincreasing], then $f(c+) \le f(d-)$ [*resp.* $f(c+) \ge f(d-)$].

PROOF: We will suppose c is an interior point of I; the case of an endpoint is similar but easier. Suppose f is nondecreasing on I. Set $u = \inf_{x>c} f(x)$. Since $f(c) \le f(x)$ for all $x > c$, $f(c)$ is a lower bound for $\{f(x) : x > c\}$, so u exists and $f(c) \le u$. We claim $u = f(c+)$. Let $\epsilon > 0$. There is a $y > c$ with $f(y) < u + \epsilon$, by definition of u. Set $\delta = y - c$; then $\delta > 0$ and, if $c < x < c + \delta$, then $c < x < y$ and $u \le f(x) \le f(y) < u + \epsilon$, so $|f(x) - u| < \epsilon$. Similarly, if $v = \sup_{x<c} f(x)$, then $v = f(c-)$ and $v \le f(c)$. If $d > c$, choose x with $c < x < d$; then $f(c+) \le f(x) \le f(d-)$. The proof for nonincreasing f is almost identical. 👍

It follows that if f is monotone on I and $c \in I$, then either f is continuous at c (if $f(c-) = f(c+)$), or f has a jump discontinuity at c (if $f(c-) \ne f(c+)$).

XIX.2.1.8. PROPOSITION. Let f be monotone on a closed interval $[a, b]$, with jump discontinuities at $x_1, \ldots, x_n \in [a, b]$. Let h_j be the height of the jump discontinuity of f at x_j. If $M = |f(b) - f(a)|$, then $\sum_{j=1}^{n} h_j \le M$.

PROOF: Suppose f is nondecreasing. We may assume

$$a \le x_1 < x_2 < \cdots < x_n \le b$$

so we have

$$f(a) \le f(x_1-) < f(x_1+) \le f(x_2-) < f(x_2+) \le \cdots \le f(x_n-) < f(x_n+) \le f(b) \ .$$

Thus,

$$h_1 + \cdots + h_n = [f(x_1+) - f(x_1-)] + \cdots + [f(x_n+) - f(x_n-)]$$
$$\le [f(x_1+) - f(a)] + [f(x_2+) - f(x_1+)] + \cdots + [f(b) - f(x_{n-1}+)] = M \ .$$

The case of nonincreasing f is the same with some inequalities reversed. 👍

XIX.2.1.9. COROLLARY. Let f be monotone on $[a, b]$, and $\epsilon > 0$. If $M = |f(b) - f(a)|$, then f has no more than M/ϵ jump discontinuities on $[a, b]$ of height $\ge \epsilon$.

We summarize these results in a theorem.

XIX.2.1.10. THEOREM. Let f be monotone on an interval I. Then f has only countably many discontinuities on I, each of which is a jump discontinuity.

PROOF: We have proved everything except the countability of the set of discontinuities. But if $[a,b] \subseteq I$, and $M = |f(b) - f(a)|$, then for each $n \in \mathbb{N}$, f has only finitely many jump discontinuities of height $\geq \frac{1}{n}$ (no more than Mn) on $[a,b]$, and hence only countably many jump discontinuities in all. Since I is a countable union of closed bounded subintervals, f can only have countably many jump discontinuities on all of I. 👍

XIX.2.1.11. COROLLARY. Let f be a monotone function on an interval I. Then there is a monotone function g on I which is right continuous on I, and which agrees with f except at countably many points.

PROOF: For $x \in I$, set $g(x) = f(x+)$. Wherever f is continuous, $f(x) = f(x+) = g(x)$. 👍

The same result holds if g is to be left continuous: set $g(x) = f(x-)$ for all x.

XIX.2.2. Differentiability of Monotone Functions

A much deeper and more remarkable result is true about monotone functions: they are not only continuous almost everywhere, but even differentiable almost everywhere. This result is due to Lebesgue.

XIX.2.2.12. THEOREM. Let f be monotone on an interval I. Then f is differentiable almost everywhere (with respect to Lebesgue measure).

However, the set where a monotone function is nondifferentiable is not necessarily countable. For example, the Cantor function () is not differentiable at the points of K, which is uncountable (but has measure 0).

XIX.2.2.13. THEOREM. Let f be a nondecreasing function on $[a,b]$. Then f' is integrable on $[a,b]$ and

$$\int_a^b f' \, d\lambda \leq f(b) - f(a) \ .$$

PROOF: By XIX.2.2.12., f' is defined almost everywhere on $[a,b]$. Extend f to $[a,b+1]$ by setting $f(x) = f(b)$ for $b \leq x \leq b+1$. Define

$$g_n(x) = n\left[f\left(x + \frac{1}{n}\right) - f\left(x\right)\right]$$

for $n \in \mathbb{N}$. Then $g_n \to f'$ pointwise a.e., so f' is measurable. The g_n and f' are nonnegative, so by Fatou's Lemma we have

$$\int_a^b f' \, d\lambda \leq \liminf_n \int_a^b g_n \, d\lambda = \liminf_n \ n\int_a^b \left[f\left(x + \frac{1}{n}\right) - f\left(x\right)\right] \, d\lambda(x)$$

$$= \liminf_n \left[n\int_{a+\frac{1}{n}}^{b+\frac{1}{n}} f \, d\lambda - n\int_a^b f \, d\lambda\right] = \liminf_n \left[n\int_b^{b+\frac{1}{n}} f \, d\lambda - n\int_a^{a+\frac{1}{n}} f \, d\lambda\right] \ .$$

Since f is constant on $\left[b, b+\frac{1}{n}\right]$, the first integral in the last expression is $\frac{f(b)}{n}$; and $f(x) \geq f(a)$ for $a \leq x \leq a+\frac{1}{n}$, so the second integral is at least $\frac{f(a)}{n}$. Thus we have

$$\int_a^b f' \, d\lambda \leq f(b) - \limsup_n \; n \int_a^{a+\frac{1}{n}} f \, d\lambda \leq f(b) - f(a) \; .$$



XIX.2.2.14. We do not have equality in XIX.2.2.13. in general. If f is the Cantor function (), then f is nondecreasing on $[0,1]$, but $f' = 0$ a.e., so we have

$$0 = \int_0^1 f' \, d\lambda < f(1) - f(0) = 1 \; .$$

We do have equality for absolutely continuous functions (and only for absolutely continuous functions) ().

XIX.2.3. Fubini's Series Theorem

We have the following interesting and useful result of Fubini. We call this "Fubini's Series Theorem" (not a standard name) since the name "Fubini's Theorem" is usually reserved for his better-known result on iterated integrals ().

XIX.2.3.1. Theorem. [Fubini's Series Theorem] Let (f_k) be a sequence of nondecreasing functions on a closed bounded interval $[a,b]$. Suppose $\sum_{k=1}^{\infty} f_k$ converges pointwise on $[a,b]$ to a real-valued function f. Then

(i) The series $\sum_{k=1}^{\infty} f_k$ converges to f uniformly on $[a,b]$.

(ii) The series $\sum_{k=1}^{\infty} f_k'$ converges to f' pointwise a.e. on $[a,b]$. In particular, $f_k' \to 0$ pointwise a.e. on $[a,b]$. (Note that by XIX.2.2.12., f and each f_k are differentiable a.e. on $[a,b]$.)

Of course, there is an analogous result for nonincreasing functions.

The statement of the theorem can be rephrased in terms of convergence of sequences of functions (i.e. the partial sums of the series), but it is a little awkward to do so since the hypothesis is that the successive differences of the functions in the sequence, not the functions themselves, must be nondecreasing.

For the proof of XIX.2.3.1., we need two lemmas:

XIX.2.3.2. Lemma. To prove the theorem, it suffices to assume that each f_k is nonnegative and $f_k(a) = 0$.

Proof: Let (f_k) be as in the statement of the theorem. For each k, set $g_k(x) = f_k(x) - f_k(a)$. Then $g_k(a) = 0$, and g_k is nonnegative and nondecreasing since f_k is nondecreasing. We have

$$\sum_{k=1}^{n} g_k(x) = \sum_{k=1}^{n} f_k(x) - \sum_{k=1}^{n} f_k(a)$$

for each n. Since the first sum converges to $f(x)$ and the second sum to $f(a)$ as $n \to \infty$, $\sum_{k=1}^{\infty} g_k(x)$ converges to $g(x) := f(x) - f(a)$ for every $x \in [a,b]$. If the conclusion of the theorem holds for the g_k, then $\sum_{k=1}^{\infty} g_k$ converges uniformly to g, so for any $\epsilon > 0$ there is an N such that

$$\left| g(x) - \sum_{k=1}^{n} g_k(x) \right| < \frac{\epsilon}{2}$$

for all $n \geq N$ and all $x \in [a,b]$, and

$$\left| f(a) - \sum_{k=1}^{n} f_k(a) \right| < \frac{\epsilon}{2}$$

for all $n \geq N$. Then, for all $n \geq N$ and $x \in [a,b]$,

$$\left| f(x) - \sum_{k=1}^{n} f_k(x) \right| = \left| (g(x) + f(a)) - \sum_{k=1}^{n} (g_k(x) + f_k(a)) \right| \leq \left| f(x) - \sum_{k=1}^{n} f_k(x) \right| + \left| f(a) - \sum_{k=1}^{n} f_k(a) \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

so $\sum_{k=1}^{\infty} f_k$ converges uniformly to f on $[a,b]$. Part (ii) for f obviously follows from part (ii) for g since $f_k' = g_k'$ for each k and $f' = g'$. 👍

XIX.2.3.3. LEMMA. Under the hypotheses of XIX.2.3.1., we have that $\sum_{k=1}^{\infty} f_k$ converges uniformly to f and that $f_k' \to 0$ pointwise a.e. on $[a,b]$.

PROOF: By XIX.2.3.2., we may assume each f_k is nonnegative and $f_k(a) = 0$. For each n, set

$$s_n(x) = \sum_{k=1}^{n} f_k(x) \ .$$

If $\epsilon > 0$, choose N such that

$$0 \leq f(b) - s_n(b) = f(b) - \sum_{k=1}^{n} f_k(b) < \epsilon$$

for all $n \geq N$. Then, for any $n \geq N$ and $x \in [a,b]$, we have

$$0 \leq f(x) - s_n(x) \leq f(b) - s_n(b) < \epsilon$$

since $f - s_n$ is nonnegative and nondecreasing. Thus $s_n \to f$ uniformly on $[a,b]$.

Set

$$E = \{x \in [a,b] : f \text{ and each } f_k \text{ are differentiable at } x\} \ .$$

By XIX.2.2.12., $[a,b] \setminus E$ is a countable union of null sets, so is a null set. Each s_n is differentiable on E, and since $s_{n+1} - s_n = f_{n+1}$ and $f - s_n$ are nondecreasing, we have $(s_{n+1} - s_n)'(x) \geq 0$ and $(f - s_n)'(x) \geq 0$ for all $x \in E$. So, for every $x \in E$, we have

$$s_1'(x) \leq s_2'(x) \leq \cdots \leq f'(x)$$

and hence $(s_n'(x))$ converges for $x \in E$, which implies that $f_n'(x) \to 0$ for all $x \in E$. 👍

XIX.2.3.4. To complete the proof of XIX.2.3.1., we need to show that $s_n' \to f'$ pointwise a.e. on $[a,b]$. Let E be the set in the proof of XIX.2.3.3.. Since the sequence $(s_n'(x))$ is nondecreasing for each $x \in E$, it suffices to find a subsequence that converges pointwise to f' a.e. on E. For each j, choose an n_j such that $f(b) - s_{n_j}(b) < 2^{-j}$, and set $t_j = f - s_{n_j}$. Then

$$0 \le t_j(x) = f(x) - s_{n_j}(x) \le f(b) - s_{n_j}(b) < 2^{-j}$$

for all $x \in [a,b]$, and thus the series $\sum_{j=1}^{\infty} t_j$ satisfies the hypotheses of XIX.2.3.1.. So by XIX.2.3.3. we have $t_j' \to 0$ pointwise on

$$E' = \{x \in [a,b] : t \text{ and each } t_j \text{ are differentiable at } x\}$$

(where $t = \sum_{j=1}^{\infty} t_j$), i.e. $s_{n_j}' \to f'$ pointwise on $E \cap E'$. Thus $s_n' \to f'$ pointwise on $E \cap E'$, whose complement is a null set.

XIX.2.3.5. Note that we do not necessarily have $E \subset E'$ since there is no assurance that t is differentiable on E, and thus $\sum_{k=1}^{\infty} f_k'$ does not necessarily converge to f' on all of E. See Exercise XIX.2.4.2..

Examples

EXAMPLES. (i) A nonconstant increasing function f can have $f' = 0$ a.e. The Cantor function () is an example. One can even have a strictly increasing function g with $g' = 0$ a.e.: take $g = \sum_{n=1}^{\infty} 2^{-n} f_n$, where the f_n are copies of the Cantor function scaled to each of the subintervals where f and all the previous rescalings of f are constant. (Expand)***

(ii) There is a strictly increasing function f on $[0,1]$ (or any other interval), which is continuously differentiable on all of $[0,1]$, and whose derivative vanishes on a set of positive measure. Let K be a Cantor set of positive measure (), and $g : [0,1] \to [0,1]$ a continuous function with $g^{-1}(\{0\}) = K$ (e.g. $g(x) = \rho(x,K)$; cf. ()). Define

$$f(x) = \int_0^x g(t)\,dt\ .$$

Then $f' = g$ exists and is continuous on all of $[0,1]$ (), and, for any $a, b \in [0,1]$, $a < b$,

$$f(b) - f(a) = \int_a^b g(t)\,dt > 0$$

since g is continuous, nonnegative, and not identically 0 on $[a,b]$ (); recall that K contains no nontrivial intervals). Thus f is strictly increasing. But $f' = g = 0$ on K. K can be taken to have measure $1 - \epsilon$, for any $\epsilon > 0$.

Note that if f is differentiable everywhere on an interval I and $f' = 0$ a.e. on I, then f is constant on I ().

(iii) [POMPEIU] [?] There is a strictly increasing everywhere differentiable function f on a closed bounded interval (actually any interval), with bounded derivative, for which $f' = 0$ on a dense set. Let (r_k) be an enumeration of the rationals in $[0,1]$, and set $g_k(x) = 2^{-k}\sqrt[3]{x - r_k}$,

$$g(x) = \sum_{k=1}^{\infty} g_k(x) = \sum_{k=1}^{\infty} 2^{-k}\sqrt[3]{x - r_k}$$

for $x \in [0,1]$. The series converges uniformly on $[0,1]$, and thus g is a strictly increasing continuous function on $[0,1]$. Let $[a,b]$ be the range of g, and $f : [a,b] \to [0,1]$ the inverse function.

We claim that f is differentiable everywhere on $[a,b]$. Each g_k is differentiable everywhere except at r_k, where its extended derivative is $+\infty$. Thus, by the Mean Value Theorem, we have, for all $x, y \in [0,1]$, $x \neq y$, we have

$$\frac{g(y) - g(x)}{y - x} \geq \frac{g_1(y) - g_1(x)}{y - x} \geq \frac{1}{6}$$

and so all Dini derivatives of g are $\geq \frac{1}{6}$ at every point. Thus to show that f is differentiable everywhere on $[a,b]$, by the Inverse Function Theorem () it suffices to show that g has an extended derivative everywhere on $[0,1]$. It will then also follow that $0 \leq f'(x) \leq 6$ for all $x \in [a,b]$.

In fact, we will show that at every $x \in [0,1]$, the extended derivative of g at x equals $\sum_{k=1}^{\infty} g_k'(x)$. This is clear if $x = r_k \in \mathbb{Q}$ [g is the sum of g_k and another nondecreasing continuous function, so $g'(x) = g_k'(x) = +\infty$]. Suppose $x \notin \mathbb{Q}$ but $\sum_{k=1}^{\infty} g_k'(x) = +\infty$. Set $h_n = \sum_{k=1}^{n} g_k$. Then for each M there is an n such that

$$h_n'(x) = \lim_{y \to x} \frac{h_n(y) - h_n(x)}{y - x} > M$$

so there is a $\delta > 0$ such that for all $y \in [x - \delta, x + \delta]$ we have

$$\frac{g(y) - g(x)}{y - x} \geq \frac{h_n(y) - h_n(x)}{y - x} > M$$

and thus $g'(x) = +\infty = \sum_{k=1}^{\infty} g_k'(x)$.

Now suppose $x \in [0,1]$ and $\sum_{k=1}^{\infty} g_k'(x) < \infty$. Then $x \notin \mathbb{Q}$. We have

$$|s^{1/3} - t^{1/3}| \leq \frac{4}{3} s^{-2/3} |s - t|$$

for all $s, t \in \mathbb{R}$, $s \neq 0$ (Exercise XIX.2.4.1.). Setting $s = x - r_k$ (recall that $x \notin \mathbb{Q}$) and $t = y - r_k$ we obtain

$$\left| \frac{g_k(y) - g_k(x)}{y - x} \right| \leq 4 g_k'(x)$$

for all k and all $y \neq x$. Thus we have, for $y \neq x$ and any n,

$$\left| \frac{g(y) - g(x)}{y - x} - \sum_{k=1}^{\infty} g_k'(x) \right| = \left| \sum_{k=1}^{\infty} \left(\frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right) \right|$$

$$\leq \sum_{k=1}^{\infty} \left| \frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right| = \sum_{k=1}^{n} \left| \frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right| + \sum_{k=n+1}^{\infty} \left| \frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right|$$

for every $y \neq x$ and every n. Fix $\epsilon > 0$ and choose n such that $\sum_{k=n+1}^{\infty} 5 g_k'(x) < \frac{\epsilon}{2}$. There is a $\delta > 0$ such that if $0 < |y - x| < \delta$, we have

$$\left| \frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right| < \frac{\epsilon}{2n}$$

for $1 \leq k \leq n$. For $0 < |y - x| < \delta$, we have

$$\left| \frac{g(y) - g(x)}{y - x} - \sum_{k=1}^{\infty} g_k'(x) \right| \leq \sum_{k=1}^{n} \left| \frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right| + \sum_{k=n+1}^{\infty} \left| \frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right|$$

$$\leq \sum_{k=1}^{n} \left| \frac{g_k(y) - g_k(x)}{y - x} - g_k'(x) \right| + \sum_{k=n+1}^{\infty} 5g_k'(x) < n\frac{\epsilon}{2n} + \frac{\epsilon}{2} = \epsilon \ .$$

Thus $g'(x)$ exists and equals $\sum_{k=1}^{\infty} g_k'(x)$.

Since the extended derivative of g is $+\infty$ at every point of $\mathbb{Q} \cap [0,1]$, we have $f' = 0$ at all points of $g(\mathbb{Q} \cap [0,1])$ (), which is dense in $[a,b]$. We also have $f'(x) = 0$ for many x not in this countable set: the set

$$A = \{x \in [a,b] : f'(x) = 0\}$$

is dense in $[a,b]$, hence A must contain the set

$$B = \{x \in [a,b] : f' \text{ is continuous at } x\}$$

which is a residual set by (). (However, the meager set $[a,b] \setminus A$ must be dense in $[a,b]$ and have positive measure in every subinterval by ().)

XIX.2.4. Exercises

XIX.2.4.1. Show that if $s, t \in \mathbb{R}$, $s \neq 0$, then

$$|s^{1/3} - t^{1/3}| \leq \frac{4}{3} s^{-2/3} |s - t| \ .$$

[Convert the problem into showing that

$$|y - z| \leq \frac{4}{3y^2} |y^3 - z^3|$$

for $y \neq 0$ and use the factorization

$$y^3 - z^3 = (y - z)(y^2 + yz + z^2) \ .]$$

XIX.2.4.2. [vRS82, 4.H] For each k, let

$$f_k(x) = \frac{1}{k} \tan^{-1} kx - \frac{1}{k+1} \tan^{-1}(k+1)x \ .$$

(a) Show that f_k is strictly increasing and differentiable on $\mathbb{R}$.
(b) Show that $\sum_{k=1}^{\infty} f_k(x)$ converges pointwise (in fact, uniformly) to $f(x) = \tan^{-1} x$.
(c) Show that $\sum_{k=1}^{\infty} f_k'(0)$ does not converge to $f'(0)$.

XIX.2.4.3. Let f be a monotone function on $[a,b]$.
(a) Show that f is Borel measurable. [Show that $f^{-1}((c, +\infty))$ is an interval for any $c \in \mathbb{R}$.]
(b) Show that f is Baire class 1 (cf. XVI.2.10.5..)

XIX.3. Convex Functions

In this section, we define and derive the most important properties of convex functions.

XIX.3.1. Definitions

XIX.3.1.1. Definition. Let f be a function from an interval I to $\mathbb{R}$. Then f is *convex* on I if for all $x, y \in I$ and $0 < t < 1$, we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \ .$$

Geometrically, this means that the graph of f between x and y lies below the chord between $(x, f(x))$ and $(y, f(y))$, for any $x, y \in I$ (Figure 1).

The terms *concave upward* and *convex downward* are sometimes used as synonyms for *convex* (although for us *concave upward* will mean "strictly convex," cf. V.8.3.1.). The term *convex* is most common, and is motivated by the following observation:

XIX.3.1.2. Proposition. Let f be a real-valued function on an interval I. Then f is a convex function if and only if its *supergraph*

$$S_f = \{(x, y) : x \in I, y \geq f(x)\}$$

is a convex subset () of $\mathbb{R}^2$.

XIX.3.1.3. Proposition. [Chordal Slope Criterion] Let I be an interval and $f : I \to \mathbb{R}$. The following are equivalent:

(i) f is convex on I.

(ii) For all $x, y, z \in I$, $x < y < z$, we have

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \ .$$

(iii) For all $x, y, z \in I$, $x < y < z$, we have

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y} \ .$$

(iv) For all $x, y, z \in I$, $x < y < z$, we have

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(y)}{z - y} \ .$$

(See Figure 2.)

PROOF: (i) $\Rightarrow$ (ii): Suppose f is convex on I and $x, y, z \in I$, $x < y < z$. Set $t = \frac{z-y}{z-x}$. Then

$$tx + (1-t)z = \frac{z-y}{z-x}x + \frac{y-x}{z-x}z = \frac{xz - xy + yz - xz}{z-x} = \frac{yz-xy}{z-x} = y$$

so

$$f(y) \le tf(x) + (1-t)f(z) = \frac{z-y}{z-x}f(x) + \frac{y-x}{z-x}f(z)$$

and therefore

$$f(y) - f(x) \le \left[\frac{z-y}{z-x} - 1\right] f(x) + \frac{y-x}{z-x}f(z)$$

$$= -\frac{y-x}{z-x}f(x) + \frac{y-x}{z-x}f(z) = \frac{y-x}{z-x}[f(z) - f(x)]$$

So we have

$$\frac{f(y)-f(x)}{y-x} \le \frac{f(z)-f(x)}{z-x} \ .$$

(ii) $\Rightarrow$ (i): Let $x, z \in I$, $x < z$, and $0 < t < 1$, and set $y = tx + (1-t)z$. Then $x < y < z$, and $t = \frac{z-y}{z-x}$. The steps in (i) $\Rightarrow$ (ii) can be reversed to show that

$$f(y) \le \frac{z-y}{z-x}f(x) + \frac{y-x}{z-x}f(z) = tf(x) + (1-t)f(z) \ .$$

(ii) $\Rightarrow$ (iii): Let $x, y, z \in I$, $x < y < z$. Suppose

$$\frac{f(y)-f(x)}{y-x} \le \frac{f(z)-f(x)}{z-x} \ .$$

Since

$$\frac{f(z)-f(x)}{z-x}(z-x) = f(z) - f(x) = \frac{f(y)-f(x)}{y-x}(y-x) + \frac{f(z)-f(y)}{z-y}(z-y)$$

$$\le \frac{f(z)-f(x)}{z-x}(y-x) + \frac{f(z)-f(y)}{z-y}(z-y)$$

we have

$$\frac{f(z)-f(x)}{z-x}(z-x) - \frac{f(z)-f(x)}{z-x}(y-x) \le \frac{f(z)-f(y)}{z-y}(z-y)$$

$$\frac{f(z)-f(x)}{z-x}[(z-x)-(y-x)] = \frac{f(z)-f(x)}{z-x}(z-y) \le \frac{f(z)-f(y)}{z-y}(z-y)$$

and the result follows from dividing by $z-y$.

(iii) $\Rightarrow$ (ii) is almost identical to (ii) $\Rightarrow$ (iii), and is left to the reader.

(iii) $\Rightarrow$ (iv): (iv) is an immediate combination of (ii) and (iii).

(iv) $\Rightarrow$ (ii) and (iii): Let $x, y, z \in I$, $x < y < z$. Then

$$\frac{f(z)-f(x)}{z-x} = \frac{y-x}{z-x}\frac{f(y)-f(x)}{y-x} + \frac{z-y}{z-x}\frac{f(z)-f(y)}{z-y}$$

so $\frac{f(z)-f(x)}{z-x}$ is a weighted average of $\frac{f(y)-f(x)}{y-x}$ and $\frac{f(z)-f(y)}{z-y}$ and thus lies in between. 👍

XIX.3.1.4. This result can be rephrased (cf. [vRS82, §2]). For a function f defined on an interval I, define, for $x, y \in I$,

$$\Phi_1 f(x,y) = \frac{f(y) - f(x)}{x - y} \ .$$

For $x, y, z \in I$, all distinct, define

$$\Phi_2 f(x,y,z) = \frac{\frac{f(z)-f(y)}{z-y} - \frac{f(y)-f(x)}{y-x}}{z - x} \ .$$

Note that the values of $\Phi_1 f$ and $\Phi_2 f$ are unchanged if the variables are permuted, i.e. $\Phi_1 f(x,y) = \Phi_1 f(y,x)$, $\Phi_2(x,y,z) = \Phi_2(x,z,y)$, etc.

XIX.3.1.5. Corollary. Let f be a real-valued function on an interval I. The following are equivalent:

(i) f is convex on I.

(ii) For each $x \in I$, the function $\Phi_1 f(x, \cdot)$ is increasing on $I \setminus \{x\}$.

(iii) $\Phi_2 f(x,y,z) \geq 0$ for all distinct $x, y, z \in I$.

Here are some other consequences of XIX.3.1.3.:

XIX.3.1.6. Corollary. Let f be convex on I, and $w, x, y, z \in I$, $w < x < y < z$. Then

$$\frac{f(x) - f(w)}{x - w} \leq \frac{f(z) - f(y)}{z - y} \ .$$

Proof: Compare both to $\frac{f(y)-f(x)}{y-x}$. 👍

XIX.3.1.7. Corollary. If f is differentiable on I, and f' is nondecreasing on I, then f is convex on I. In particular, if f'' exists and is nonnegative on I, then f is convex on I.

The converse is also true, and the assumption that f is differentiable is almost automatic (XIX.3.1.16.).

Proof: Suppose f' is nondecreasing on I. Let $x, y, z \in I$, $x < y < z$. By the Mean Value Theorem (), there are $c, d \in I$ with $x < c < y < d < z$ such that

$$f'(c) = \frac{f(y) - f(x)}{y - x} \quad \text{and} \quad f'(d) = \frac{f(z) - f(y)}{z - y} \ .$$

Since f' is nondecreasing, $f'(c) \leq f'(d)$. Apply XIX.3.1.3.(iv). 👍

XIX.3.1.8. EXAMPLES. (i) Any linear (affine) function is convex on $(-\infty, \infty)$.
(ii) $f(x) = x^r$ is convex on $[0, \infty)$ if $r = 0$ or $r \geq 1$, convex on $(0, \infty)$ if $r < 0$, and not convex on any interval if $0 < r < 1$. If $n \in \mathbb{N}$ is even, $f(x) = x^n$ is convex on $(-\infty, \infty)$ and $g(x) = x^{-n}$ is convex on $(-\infty, 0)$.
(iii) $f(x) = e^x$ is convex on $(-\infty, \infty)$.
(iv) A convex function need not be differentiable everywhere: $f(x) = |x|$ is convex on $(-\infty, \infty)$.
(v) A convex function need not even be continuous. The following function is convex on $[0, 1]$:

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

XIX.3.1.9. PROPOSITION. (i) If f and g are convex on I, then $f + g$ is convex on I.
(ii) If f is convex on I and $c \geq 0$, then cf is convex on I.
(iii) If I and J are intervals, f is convex on I, $f(I) \subseteq J$, and g is nondecreasing and convex on J, then $g \circ f$ is convex on I.

PROOF: (i) and (ii) are straightforward, and are left to the reader. For (iii), if $x, y \in I$ and $0 < t < 1$, then

$$g(f(tx + (1-t)y)) \leq g(tf(x) + (1-t)f(y)) \leq tg(f(x)) + (1-t)g(f(y)) \ .$$



XIX.3.1.10. Note, however, that a product of convex functions is not convex in general, even if both are nonnegative and one is linear. For example $f(x) = x^2$ and $g(x) = 1 - x$ are convex and nonnegative on $[0, 1]$, but fg is not convex on $[0, 1]$. (A product of two nonnegative, nondecreasing, convex functions is convex (XIX.3.1.24.).) Also note that (iii) fails in general if g is not nondecreasing, e.g. if $g(x) = -x$.

Limits and Suprema of Convex Functions

XIX.3.1.11. PROPOSITION. Let I be an interval, and (f_n) a sequence of convex functions on I converging pointwise on I to a function f. Then f is convex on I.

PROOF: If $x, y \in I$ and $0 < t < 1$, then $f_n(tx + (1-t)y) \to f(tx + (1-t)y)$ and $tf_n(x) + (1-t)f_n(y) \to tf(x) + (1-t)f(y)$. It follows immediately that f is convex. 

In fact, such a convergent sequence of convex functions almost converges uniformly (XIX.3.3.1.).

XIX.3.1.12. PROPOSITION. Let $\{f_i\}$ be a set of convex functions on an interval I. If $f = \sup_i f_i$ is finite on I, then f is convex on I. In particular, if $f_1, \ldots, f_n$ are convex on I, then $f_1 \vee \cdots \vee f_n = \max(f_1, \ldots, f_n)$ is convex on I.

PROOF: This follows immediately from XIX.3.1.2. and (), since $S_f = \cap_i S_{f_i}$.

Continuity of Convex Functions

The discontinuity in XIX.3.1.8.(v) can only happen at an endpoint. In fact, we have:

XIX.3.1.13. PROPOSITION. Let f be convex on I, and $[a,b]$ a closed bounded interval contained in the interior of I. Then f is Lipschitz on $[a,b]$.

PROOF: Let $c,d \in I$, $c < a < b < d$. If $x,y \in [a,b]$, then by XIX.3.1.6. we have

$$\frac{f(a)-f(c)}{a-c} \leq \frac{f(y)-f(x)}{y-x} \leq \frac{f(d)-f(b)}{d-b}$$

and so $|f(y)-f(x)| \leq M|y-x|$, where

$$M = \max\left(\left|\frac{f(a)-f(c)}{a-c}\right|, \left|\frac{f(d)-f(b)}{d-b}\right|\right) .$$



XIX.3.1.14. PROPOSITION. Let f be convex on I. Then f is continuous on the interior of I. If I has a left endpoint a [*resp.* right endpoint b], then $\lim_{x\to a+} f(x)$ exists [*resp.* $\lim_{x\to b-} f(x)$] exists, and $\lim_{x\to a+} f(x) \leq f(a)$ [*resp.* $\lim_{x\to b-} f(x) \leq f(b)$]. In particular, f is lower semicontinuous at endpoints of I.

PROOF: Continuity of f on the interior of I follows immediately from XIX.3.1.13., since a Lipschitz function is (uniformly) continuous.

Suppose I has a left endpoint a. Let $c \in I$, $c > a$. Then, if $a < x < c$, we have by XIX.3.1.3.(iii),

$$g(x) := \frac{f(c)-f(x)}{c-x} \geq \frac{f(c)-f(a)}{c-a}$$

and $g(x)$ decreases as x decreases; thus

$$L := \lim_{x\to a+} \frac{f(c)-f(x)}{c-x} = \inf_{a<x<c} \frac{f(c)-f(x)}{c-x}$$

exists, and satisfies

$$L \geq \frac{f(c)-f(a)}{c-a} .$$

Since $\lim_{x\to a+}(c-x) = c-a$ exists and is nonzero, $m := \lim_{x\to a+} f(x)$ also exists and satisfies

$$L = \frac{f(c)-m}{c-a} \geq \frac{f(c)-f(a)}{c-a}$$

and so $m = \lim_{x\to a+} f(x) \leq f(a)$. The argument for a right endpoint b of I is virtually identical.

Differentiability of Convex Functions

An almost identical argument shows that the left and right derivatives of a convex function exist at every point:

XIX.3.1.15. THEOREM. Let f be a convex function on an interval I. Then the left and right derivatives $D_L f$ and $D_R f$ of f exist at every point of I (except a left or right endpoint respectively), and are finite except possibly at endpoints of I. If $x, y \in I$, $x < y$, then

$$D_L f(x) \le D_R f(x) \le D_L f(y) \le D_R f(y) \ .$$

$D_L f$ is continuous from the left wherever it is finite (i.e. except possibly at a right endpoint of I), and $D_R f$ is continuous from the right wherever it is finite (i.e. except possibly at a left endpoint of I).

In fact, if I has a right endpoint b, and f is continuous at b, and $D_L f(b) = +\infty$, then $\lim_{x \to b-} D_L f(x) = +\infty$ (cf. **??**), i.e. $D_L f$ is "continuous" at b. A similar statement holds for $D_R f$ at a left endpoint of I.

PROOF: Let $x \in I$, not a right endpoint. Then there is an $\epsilon > 0$ such that $x + \epsilon \in I$. Then

$$g(h) := \frac{f(x+h) - f(x)}{h}$$

exists for $0 < h < \epsilon$. By XIX.3.1.3.(ii), $g(h)$ decreases as h decreases, so

$$D_R f(x) = \lim_{h \to 0+} \frac{f(x+h) - f(x)}{h} = \inf_{0<h<\epsilon} \frac{f(x+h) - f(x)}{h}$$

exists as an extended real number and is not $+\infty$. If x is not a left endpoint, let $a \in I$, $a < x$; then by XIX.3.1.3.(iv) we have

$$\frac{f(x+h) - f(x)}{h} \ge \frac{f(x) - f(a)}{x - a}$$

for $0 < h < \epsilon$, so $D_R f(x)$ is finite (i.e. not $-\infty$). A similar argument shows that if x is not a left endpoint, then $D_L f(x)$ exists and is not $-\infty$; and if x is not a right endpoint, $D_L f(x)$ is finite (not $+\infty$). If $x, y \in I$, $x < y$, then

$$\frac{f(x+h) - f(x)}{h} \le \frac{f(y) - f(y-h)}{h}$$

for $0 < h < (y - x)/2$ by XIX.3.1.6., and hence $D_R f(x) \le D_L f(y)$.

To prove left-continuity of $D_L f$, suppose x is not a left endpoint of I and $D_L f(x)$ is finite. It follows that f is continuous at x. Let $\epsilon > 0$. Then there is a $y \in I$, $y < x$, such that

$$D_L f(x) - \epsilon < \frac{f(y) - f(x)}{y - x} \le D_L f(x) \ .$$

If (x_n) is a sequence in (y, x) converging to x, we have $f(x_n) \to f(x)$, so for sufficiently large n we have

$$D_L f(x) - \epsilon < \frac{f(y) - f(x_n)}{y - x_n} \le D_L f(x_n) \le D_L f(x) \ .$$

Right-continuity of $D_R f$ is almost identical. 👍

XIX.3.1.16. COROLLARY. Let f be convex on I. Then f is differentiable on I except on a countable set (which may be finite or empty), f' is continuous on the set on which it is defined, and f' is nondecreasing on I.

PROOF: It is immediate from XIX.3.1.15. that $D_L f$ is nondecreasing on I, and hence is continuous except at countably many points (). If $x \in I$, then $D_L f(x) \le D_R f(x) \le D_L f(y)$ for all $y > x$, and hence $D_R f(x) = D_L f(x)$ if $D_L f$ is continuous from the right at x (and only at such points). Thus f is differentiable at any point where $D_L f$ is continuous, hence at all but countably many points. 👍

We then get the following "geometrically obvious" consequence:

XIX.3.1.17. COROLLARY. Let f be convex and continuous on an interval I. Then one of the following holds:

(i) $D_L f(x) \ge 0$ for all $x \in I$.

(ii) $D_L f(x) \le 0$ for all $x \in I$.

(iii) There are unique $c, d \in I$, $c \le d$, with $D_L f(x) < 0$ for all $x \in I$, $x < c$, $D_L f(x) > 0$ for all $x \in I$, $x > d$, and, if $c < d$, $D_L f(x) = 0$ for $c < x \le d$.

Thus either f is nondecreasing on I, f is nonincreasing on I, or there are unique $c, d \in I$, $c \le d$, such that f is strictly decreasing on $I \cap (-\infty, c]$, strictly increasing on $I \cap [d, +\infty)$, and, if $c < d$, constant on $[c, d]$.

In fact, if $D_L f$ takes both positive and negative values on I, then $c = \sup\{x \in I : D_L f(x) < 0\}$ and $d = \inf\{x \in I : D_L f(x) > 0\}$. Note that we often have $c = d$; this will happen unless f is constant on a subinterval.

XIX.3.1.18. COROLLARY. Let f be convex on an interval I. Then for each c in the interior of I there is a linear function L on I with $L(c) = f(c)$ and $L(x) \le f(x)$ for all $x \in I$.

PROOF: We can take for L the linear function with $L(c) = f(c)$ and with slope m, for any m satisfying

$$D_L f(c) \le m \le D_R f(c) \ .$$

👍

XIX.3.1.19. Conversely, if f is a continuous function on an interval I with the property that for each c in the interior of I, there is a linear function L_c on I with $L_c(c) = f(c)$ and $L_c(x) \le f(x)$ for all $x \in I$, then $f = \sup_c(L_c)$, and each L_c is convex on I, so f is convex on I by XIX.3.1.12..

Let I be an interval, and $g : I \to \mathbb{R}$ a nondecreasing function. Then g is Riemann integrable over every closed bounded subset of I (). Fix $c \in I$, and for $x \in I$ define

$$f(x) = \int_c^x g(t)\, dt \ .$$

XIX.3.1.20. PROPOSITION. f is convex on I.

PROOF: Let $x, y, z \in I$, $x < y < z$. Then

$$\frac{f(y) - f(x)}{y - x} = \frac{1}{y - x} \int_x^y g(t)\, dt$$

is the average value of g on $[x, y]$, and similarly $\frac{f(z)-f(y)}{z-y}$ is the average value of g on $[y, z]$. Since g is nondecreasing,

$$\frac{f(y) - f(x)}{y - x} \leq g(y) \leq \frac{f(z) - f(y)}{z - y} \ .$$

Apply XIX.3.1.3.. 👍

XIX.3.1.21. We will show that the converse is true, i.e. that every convex function on an open interval is the primitive of a nondecreasing function. This follows immediately from XIX.3.1.13. along with () and (), using the theory of Lebesgue integration and absolute continuity. However, it is not necessary to use Lebesgue integration since bounded nondecreasing functions are Riemann integrable on closed bounded intervals:

XIX.3.1.22. THEOREM. Let f be convex on an open interval I. Then f' is Riemann integrable over every closed bounded subinterval of I. If $a \in I$ is fixed, then, for every $x \in I$,

$$f(x) = f(a) + \int_a^x f'(t)\, dt \ .$$

PROOF: This is a special case of (). We have that $D_L f$ is increasing on I, hence Riemann integrable over every closed bounded subinterval of I, and equal to f' except at countably many points of I. 👍

XIX.3.1.23. COROLLARY. Let f be a function on an open interval I. Then f is convex on I if and only if f is absolutely continuous on every closed bounded subinterval of I and f' is increasing on a subset A of I with $\lambda(I \setminus A) = 0$.

XIX.3.1.24. COROLLARY. Let f and g be nonnegative, nondecreasing, and convex on an open interval I. Then fg is convex on I.

PROOF: Let $[a, b] \subseteq I$. Then f and g are bounded and Lipschitz on $[a, b]$ (XIX.3.1.13.), and hence fg is also Lipschitz and thus absolutely continuous on $[a, b]$. Let A be the set of points where f and g are both differentiable; then $I \setminus A$ is countable. By the product rule, if $x \in A$, then fg is differentiable at x and

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x) \ .$$

Since f, g, f', and g' are all nonnegative and nondecreasing on A, $(fg)'$ is nondecreasing on A. 👍

*** Is there an elementary proof of this? ***

Here is a generalization of the last sentence of XIX.3.1.7. (cf. [**?**, p. 39]; [HS75, 17.37(c)]). Recall () that

$$\overline{D}_S^{(2)}(f)(x) = \limsup_{h\to 0} \frac{f(x+h)+f(x-h)-2f(x)}{h^2}$$

$$\underline{D}_S^{(2)}(f)(x) = \liminf_{h\to 0} \frac{f(x+h)+f(x-h)-2f(x)}{h^2}$$

XIX.3.1.25. PROPOSITION. Let f be a function on an open interval I. Then the following are equivalent:

(i) f is convex on I.

(ii) f is continuous on I and $\overline{D}_S^{(2)}(f)(x) \geq 0$ for all $x \in I$.

(iii) f is continuous on I and $\underline{D}_S^{(2)}(f)(x) \geq 0$ for all $x \in I$.

PROOF: (i) $\Rightarrow$ (iii): If f is convex on I and $x \in I$, then there is an $\epsilon > 0$ such that $(x-\epsilon, x+\epsilon) \subseteq I$. Then, if $|h| < \epsilon$,

$$f(x) \leq \frac{1}{2}f(x-h) + \frac{1}{2}f(x+h)$$

and hence $f(x+h)+f(x-h)-2f(x) \geq 0$. Thus $\underline{D}_S^{(2)}(f)(x) \geq 0$. And f is continuous on I by XIX.3.1.14..
(iii) $\Rightarrow$ (ii) is trivial.
(ii) $\Rightarrow$ (i): Suppose f is continuous and $\overline{D}_S^{(2)}(f) \geq 0$ on I, and let $[a,b] \subseteq I$, and $\epsilon > 0$. Set

$$\phi(x) = f(x) - \left[f(a) + \frac{f(b)-f(a)}{b-a}(x-a)\right] + \epsilon(x-a)(x-b)$$

for $a \leq x \leq b$. Then ϕ is continuous on $[a,b]$, $\phi(a) = \phi(b) = 0$, and $\overline{D}_S^{(2)}(\phi) \geq 2\epsilon$ on $[a,b]$. If $\max_{[a,b]} \phi > 0$, then ϕ has a local maximum at some $x_0 \in (a,b)$, and we would have $\overline{D}_S^{(2)}(\phi)(x_0) \leq 0$, a contradiction. Thus $\phi(x) \leq 0$ on $[a,b]$,

$$f(x) \leq f(a) + \frac{f(b)-f(a)}{b-a}(x-a) - \epsilon(x-a)(x-b)$$

for all $x \in [a,b]$. Since $\epsilon > 0$ is arbitrary, we have

$$f(x) \leq f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$$

for all $x \in [a,b]$. Since $[a,b]$ is arbitrary, f is convex on I. 👍

XIX.3.2. Midpoint-Convex Functions

There is a weaker condition which almost implies convexity:

XIX.3.2.1. Definition. A real-valued function f on an interval I is *midpoint-convex* if, for every $x, y \in I$, we have

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x)+f(y)}{2} \ .$$

A convex function is midpoint-convex. The converse is false (assuming the Axiom of Choice – Exercise XIX.3.7.2.), but is "almost true" (i.e. true under very mild additional hypotheses), as the results of this subsection show. These results are quite useful since midpoint-convexity is often much easier to check than convexity.

This exposition is adapted from [**?**].

XIX.3.2.2. Proposition. Let f be a midpoint-convex function on an interval I. Then, for every $x, y \in I$ and dyadic rational $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \ .$$

Proof: We prove this for t of the form $\frac{p}{2^n}$ by induction on n. The case $n = 1$ holds by definition. Suppose it is true for n, and let $t = \frac{p}{2^{n+1}}$. Since the condition is symmetric in t and $1-t$, we may assume $p < 2^n$, so $q := 2^{n+1} - p > 2^n$. Set $r = q - 2^n$, so $p + r = 2^n$. Then

$$z := \frac{p}{2^{n+1}}x + \frac{q}{2^{n+1}}y = \frac{1}{2}\left[\frac{p}{2^n}x + \frac{r}{2^n}y + y\right]$$

so we have

$$f(z) \leq \frac{1}{2}\left[f\left(\frac{p}{2^n}x + \frac{r}{2^n}y\right) + f(y)\right] \leq \frac{1}{2}\left[\frac{p}{2^n}f(x) + \frac{r}{2^n}f(y) + f(y)\right] = \frac{p}{2^{n+1}}f(x) + \frac{q}{2^{n+1}}f(y) \ .$$



XIX.3.2.3. Corollary. If f is continuous and midpoint-convex on an interval I, then f is convex on I.

Proof: For fixed $x, y \in I$, the inequality

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

for arbitrary $t \in [0,1]$ follows from the inequality for dyadic rational t by continuity and density of the dyadic rationals.



We can say much more. A midpoint-convex function which is not convex must be not only discontinuous, but rather bizarre:

XIX.3.2.4. THEOREM. Let f be a midpoint-convex function on an open interval I. If f is not convex on I, then f is not bounded above in any subinterval. If $c \in I$, then there is a sequence (x_n) in I with $x_n \to c$ and $\lim_{n\to\infty} f(x_n) = +\infty$. In particular, f is discontinuous everywhere on I.

PROOF: By XIX.3.2.3., f is discontinuous at some $x_0 \in I$. By translation and adding a constant, we may assume $x_0 = 0$ and that $f(0) = 0$ for ease of notation. There is a sequence (y_n) in I with $y_n \to 0$ and $f(y_n) \to m \neq 0$. If $f(y_n) < 0$, we have that $f(-y_n) \geq -f(y_n) > 0$ since $0 = f(0) \leq \frac{1}{2}(f(y_n) + f(-y_n))$. Thus, by replacing (y_n) by a subsequence of $(-y_n)$ if necessary, we may assume $m > 0$. If $m = +\infty$, we are done. Otherwise, we have

$$2f(y_n) \leq f(0) + f(2y_n) = f(2y_n)$$

and so $\liminf_{n\to\infty} f(2y_n) \geq 2m$. Continuing inductively, for each k we have

$$f(2^k y_n) \geq 2^k f(y_n)$$

for all n large enough that $2^k y_n \in I$, and hence $\liminf_{n\to\infty} f(2^k y_n) \geq 2^k m$. Since for fixed k we have $2^k y_n \to 0$, we have that f is unbounded above in every neighborhood of 0, and there is a sequence $(z_n) \in I$ with $z_n \to 0$ and $f(z_n) \to +\infty$.

Now let a be the largest number (or $+\infty$) for which $(-a, a) \subseteq I$, and let $c \in (-a, a)$. We have that $x_n := c + 2z_n \to c$, so $x_n \in I$ for sufficiently large n. Also,

$$f(z_n) = f\left(\frac{c + 2z_n - c}{2}\right) \leq \frac{f(c + 2z_n) + f(-c)}{2} = \frac{f(x_n) + f(-c)}{2}$$

and, since $f(z_n) \to +\infty$, it follows that $f(x_n) \to +\infty$. Thus f is unbounded above in every subinterval of $(-a, a)$ and discontinuous everywhere on $(-a, a)$.

We end up with an open subinterval J of I, extending to at least one end of I, such that f is unbounded above on every subinterval of J and hence discontinuous everywhere on J. If $J \neq I$, repeat the argument with an $x_0 \in J$ near the endpoint of J not exhausting I to get another such interval J_2 centered at x_0. If $J \cup J_2 \neq I$, then $J \subseteq J_2$, and J_2 is approximately twice the length of J, with the same properties. Repeating the process, in finitely many steps we cover any bounded subinterval of I, which suffices to prove the result. 👍

XIX.3.2.5. A midpoint-convex function which is not convex can be bounded below: in fact, if f is midpoint-convex, so is $\max(f, 0)$.

XIX.3.2.6. COROLLARY. A midpoint-convex function on an interval I which is bounded on at least one subinterval of I is convex.

This result considerably extends XIX.3.2.3..

We can extend XIX.3.2.4. even further. A midpoint-convex function which is not convex cannot even be bounded on a set of positive Lebesgue measure. The next result is due to Ostrowski [**?**,]

XIX.3.2.7. THEOREM. Let f be midpoint-convex on an interval I. If there is a Lebesgue-measurable subset E of I with $\lambda(E) > 0$ on which f is bounded, then f is convex on I.

PROOF: Assume f is not convex on I. By replacing I by $I \cap (-n, n)$ and E by $E \cap (-n, n)$ for some sufficiently large n, we may assume $I = (a, b)$ is bounded. There is an open set $U \subseteq (a, b)$ with $E \subseteq U$ and $\lambda(U) < \frac{4}{3}\lambda(E)$ (). U is a disjoint union of finitely or countably many open intervals J_n, and we have

$$\lambda(E) = \sum_n \lambda(E \cap J_n) \leq \sum_n \lambda(J_n) < \frac{4}{3} \sum_n \lambda(E \cap J_n)$$

so there is an n such that

$$0 < \lambda(E \cap J_n) \leq \lambda(J_n) < \frac{4}{3}\lambda(E \cap J_n)$$

and replacing I by J_n and E by $E \cap J_n$ we may assume $\lambda(I) < \frac{4}{3}\lambda(E)$.

Suppose $f(x) \leq M$ for all $x \in E$. The set

$$S = \{x \in I : f(x) > M\}$$

is dense in I by XIX.3.2.4., so by contracting I slightly we may assume the midpoint of I is in S. By translating, we may assume $I = (-c, c)$ and that $0 \in S$, to simplify notation. Set

$$\bar{S} = \{-x : x \in S\} .$$

For any $x \in I$, we have

$$M < f(0) \leq \frac{f(x) + f(-x)}{2}$$

so either x or $-x$ is in S, i.e. $S \cup \bar{S} = I$. The set S is not obviously Lebesgue measurable, but in any case we have $\lambda^*(\bar{S}) = \lambda^*(S)$, where λ^* denotes Lebesgue outer measure (). We have $E \cap S = \emptyset$, so $E \subseteq \bar{S}$ and $S \subseteq (I \setminus E)$. Then

$$\lambda(E) \leq \lambda^*(\bar{S}) = \lambda^*(S) \leq \lambda(I \setminus E) = \lambda(I) - \lambda(E)$$

$$2\lambda(E) \leq \lambda(I) < \frac{4}{3}\lambda(E)$$

which is a contradiction (to the assumption that f is not convex on I). 👍

We then obtain the following corollary, due to Sierpiński:

XIX.3.2.8. COROLLARY. Let f be a midpoint-convex function on an interval I. If f is a Lebesgue measurable function, then f is convex on I.

PROOF: We have that $E_n = f^{-1}([-n, n])$ is a Lebesgue measurable subset of I for all n, and $\cup_n E_n = I$, so $\sup_n \lambda(E_n) = \lambda(I) > 0$, and f is bounded on each E_n. 👍

Summarizing, and combining with XIX.3.1.13., we have:

XIX.3.2.9. THEOREM. Let f be a midpoint-convex function on an interval I, and $[a,b]$ a closed bounded interval in the interior of I. The following are equivalent:

(i) f is convex on $[a,b]$.

(ii) f is Lipschitz on $[a,b]$.

(iii) f is continuous on $[a,b]$.

(iv) f is bounded on $[a,b]$.

(v) f is Lebesgue measurable on $[a,b]$.

XIX.3.3. Convergence of Sequences of Convex Functions

Pointwise convergence of a sequence of convex functions implies several much stronger types of convergence. We describe some of the rather remarkable consequences of pointwise convergence in this subsection.

XIX.3.3.1. THEOREM. Let I be an interval, and (f_n) a sequence of convex functions on I converging pointwise on I to a function f. Then f is convex on I, and the convergence is uniform on any closed bounded interval contained in I on which f is continuous (e.g. any closed bounded interval contained in the interior of I).

PROOF: It has been noted in XIX.3.1.11. that f is convex.

If $[a,b]$ is a closed bounded interval in I, suppose the convergence is not uniform on $[a,b]$. Then there is an $\eta > 0$ such that, for infinitely many n, there is a $c_n \in [a,b]$ with $|f_n(c_n) - f(c_n)| \geq \eta$. Passing to a subsequence, we may assume we have such a c_n for all n.

The sequence (c_n) cannot have any eventually constant subsequences, which would contradict the pointwise convergence $f_n \to f$. Thus (c_n) has a strictly monotone subsequence. Passing to this subsequence, we may assume (c_n) is strictly monotone. In the rest of the argument, we will assume (c_n) is strictly increasing; the strictly decreasing case is essentially identical, interchanging the roles of a and b.

We have $c_n \to c$ for some c, $a < c \leq b$. There are two cases to consider:

Case 1. $f_n(c_n) \geq f(c_n) + \eta$ for infinitely many n. Again passing to a subsequence, in this case we may assume $f_n(c_n) \geq f(c_n) + \eta$ for all n. Fix $\epsilon > 0$. Choose n_0 so that, for all $n \geq n_0$, we have $c_n > c - \epsilon$, $|f_n(c) - f(c)| < \frac{\eta}{2}$, and $|f_n(a) - f(a)| < \frac{\eta}{2}$. Then, for $n \geq n_0$, we have

$$\frac{f(c) - f(a) - \eta}{c - a} < \frac{f_n(c) - f_n(a)}{c - a} \leq \frac{f_n(c) - f_n(c_n)}{c - c_n} < -\frac{\eta/2}{\epsilon} \ .$$

This is a contradiction if $f(c) - f(a) - \eta \geq 0$, or if $f(c) - f(a) - \eta < 0$ and ϵ is sufficiently small, specifically if

$$\epsilon < \frac{\eta(c-a)}{2[\eta + f(a) - f(c)]} \ .$$

Case 2. $f_n(c_n) \le f(c_n) - \eta$ for infinitely many n; we may again assume this holds for all n. Fix z with $a < z < c$ and let $\lambda > 0$ with

$$-\lambda < \frac{f(z) - f(a) - \eta}{z - a} .$$

Fix d with $a < d < c$ such that $d > c - \frac{\eta}{4\lambda}$ and $|f(x) - f(c)| < \frac{\eta}{4}$ for all $x \in [d, c]$ (this is possible since f is continuous at c). Choose n_0 such that, for all $n \ge n_0$, $c_n > d$, $|f_n(a) - f(a)| < \frac{\eta}{2}$, $|f_n(z) - f(z)| < \frac{\eta}{2}$, and $|f_n(d) - f(d)| < \frac{\eta}{4}$. Then, for $n \ge n_0$, we have $f_n(d) > f(c) - \frac{\eta}{2}$ and $f_n(c_n) < f(c_n) - \eta < f(c) - \frac{3\eta}{4}$. Thus $f_n(c_n) - f_n(d) < -\frac{\eta}{4}$. We also have

$$\frac{f_n(c_n) - f_n(d)}{c_n - d} \ge \frac{f_n(z) - f_n(a)}{z - a} \ge \frac{f(z) - f(a) - \eta}{z - a} > -\lambda .$$

Then, since $0 < c_n - d < \frac{\eta}{4\lambda}$, we get

$$\frac{4\lambda}{\eta}(f_n(c_n) - f_n(d)) > -\lambda$$

$$f_n(c_n) - f_n(d) > -\frac{\eta}{4}$$

which is a contradiction.

(If $c < b$ in Case 2, a contradiction can be obtained somewhat more simply using an argument similar to the one in Case 1.) 👍

The first statement of this result which I have found is in [Con74]; however, the proof there is incorrect. A stronger assertion, which is incorrect (see the following example), is made in [**?**, IX, Book IV, Chapter I, §4, p. 52]; the proof outlined there only gives a slightly weaker conclusion than XIX.3.3.1. (cf. XIX.3.7.6.(a)-(b)), but can be adapted to give the full result (XIX.3.7.6.(c)-(d)).

XIX.3.3.2. Example. If f is not continuous on I, we do not have uniform convergence on all of I in general, even if I is closed and bounded. If $f_n(x) = x^n$, then f_n is convex on $[0, 1]$ and converges pointwise on $[0, 1]$ to the function of XIX.3.1.8.(v). But the convergence is not uniform on $[0, 1]$.

If I is not closed and bounded, simple examples show that we do not have uniform convergence on all of I in general, even if f is continuous on I: if I is unbounded, consider $f_n(x) = \frac{x}{n}$, and if $I = (0, 1]$, consider $f_n(x) = \frac{1}{x + \frac{1}{n}}$.

Here is a closely related result:

XIX.3.3.3. Theorem. Let (f_n) be a sequence of convex functions on an interval I, converging pointwise on I to a (convex) function f. Then

(i) If x is not a left endpoint of I, then

$$\liminf_{n\to\infty} D_L f_n(x) \ge D_L f(x) .$$

(ii) If x is not a right endpoint of I, then

$$\limsup_{n\to\infty} D_R f_n(x) \le D_R f(x) .$$

In particular, if x is an interior point of I and f is differentiable at x, then

$$\lim_{n\to\infty} D_L f_n(x) = \lim_{n\to\infty} D_R f_n(x) = f'(x)$$

(i.e. the limits both exist and equal $f'(x)$).

PROOF: We prove only (i); the proof of (ii) is entirely symmetric and is left to the reader. So suppose x is not a left endpoint of I, and fix $m < D_L f(x)$ (note that $D_L f(x) \neq -\infty$). Then there is a $y \in I$, $y < x$, such that

$$\frac{f(y) - f(x)}{y - x} > m$$

and thus there is an $\epsilon > 0$ such that

$$m' := \frac{f(y) - f(x) + 2\epsilon}{y - x} > m$$

(note that $y - x < 0$). There is an N such that $|f_n(x) - f(x)| < \epsilon$ and $|f_n(y) - f(y)| < \epsilon$ for all $n \geq N$. Then, for $n \geq N$, we have $f_n(y) < f(y) + \epsilon$ and $f_n(x) > f(x) - \epsilon$, and thus, for any z, $y \leq z < x$, we have, using convexity of f_n,

$$\frac{f_n(z) - f_n(x)}{z - x} \geq \frac{f_n(y) - f_n(x)}{y - x} > \frac{[f(y) + \epsilon] - [f(x) - \epsilon]}{y - x} = m'$$

so $D_L f_n(x) \geq m' > m$ for $n \geq N$, and hence $\liminf_{n\to\infty} D_L f_n(x) \geq m' > m$.

For the final statement, since $D_L f_n(x) \leq D_R f_n(x)$ for all n, we have

$$f'(x) \leq \liminf_{n\to\infty} D_L f_n(x) \leq \limsup_{n\to\infty} D_L f_n(x) \leq \limsup_{n\to\infty} D_R f_n(x) \leq f'(x)$$

$$f'(x) \leq \liminf_{n\to\infty} D_L f_n(x) \leq \liminf_{n\to\infty} D_R f_n(x) \leq \limsup_{n\to\infty} D_R f_n(x) \leq f'(x)$$

so all the inequalities are equalities. 👍

XIX.3.3.4. COROLLARY. Let (f_n) be a sequence of convex functions on an interval I converging pointwise on I to a (convex) function f. Then $f_n' \to f'$ pointwise almost everywhere.

This corollary can also be proved by combining the next theorem with ().

XIX.3.3.5. THEOREM. [Con74] Let (f_n) be a sequence of convex functions on a closed bounded interval $[a, b]$, converging pointwise on $[a, b]$ to a continuous (convex) function f. Then $f_n' \to f'$ in the L^1-norm (), i.e.

$$\lim_{n\to\infty} \|f_n' - f'\|_1 = \lim_{n\to\infty} \int_a^b |f_n'(x) - f'(x)|\, dx = 0 \ .$$

XIX.3.3.6. COROLLARY. Let (f_n) be a sequence of convex functions on a closed bounded interval $[a,b]$, converging pointwise on $[a,b]$ to a continuous (convex) function f. Then $f_n \to f$ in the total variation norm ().

To see this, we may assume that each f_n is continuous on $[a,b]$ by redefining it at the endpoints. Then f and each f_n is absolutely continuous on I, and by () the total variation norm of an absolutely continuous function g on $[a,b]$ is

$$\|g\|_{TV} = |g(a)| + \int_a^b |g'(x)|\, d\lambda(x) \ .$$

To prove XIX.3.3.5., we need a lemma:

XIX.3.3.7. LEMMA. Let f and g be convex functions on an interval I, and $P(x) = f(a) + \frac{f(b)-f(a)}{b-a}(x-a)$. Then

$$\|f' - P'\|_1 = 2\|f - P\|_\infty$$

$$\|g' - P'\|_1 \le \|f' - P'\|_1 + 4\|f - g\|_\infty$$

where $\|\phi\|_\infty = \sup\{|\phi(x)| : a \le x \le b\}$ is the uniform norm.

PROOF: For the equality, note that $f - P$ is convex and is 0 at a and b, and there is a $c \in (a,b)$ such that f is nonincreasing on $[a,c]$ and nondecreasing on $[c,b]$, so the total variation $\|f' - P'\|_1$ is $2|m|$, where $m = f(c)$ is the minimum of $f - P$ on $[a,b]$; and we also have $|m| = \|f - P\|_\infty$.

For the inequality, $g - P$ is convex on $[a,b]$, so either $g - P$ is monotone on $[a,b]$, in which case

$$\|g' - P'\|_1 = |[g(b) - P(b)] - [g(a) - P(a)]| = |[g(b) - f(b)] - [g(a) - f(a)]| \le 2\|f - g\|_\infty$$

or there is a $d \in (a,b)$ with $g - P$ nonincreasing on $[a,d]$ and nondecreasing on $[d,b]$, in which case

$$\|g' - P'\|_1 = ([g(a) - P(a)] - [g(d) - P(d)]) + ([g(b) - P(b)] - [g(d) - P(d)])$$

$$= ([g(a) - f(a)] - [g(d) - P(d)]) + ([g(b) - f(b)] - [g(d) - P(d)]) = 2[P(d) - g(d)] + [g(a) - f(a)] + [g(b) - f(b)]$$

$$\le 2|g(d) - P(d)| + 2\|g - f\|_\infty \le 2(|g(d) - f(d)| + |f(d) - P(d)|) + \|g - f\|_\infty$$

$$\le 2\|f - P\|_\infty + 4\|g - f\|_\infty = \|f' - P'\|_1 + 4\|f - g\|_\infty \ .$$



We now give the proof of XIX.3.3.5..

PROOF: Let $\epsilon > 0$. Using (), fix a partition $\mathcal{P} = \{a = t_0 < t_1 < \cdots < t_m = b\}$ of $[a,b]$ for which

$$\|f' - (\mathcal{P}f)'\|_1 < \frac{\epsilon}{4}$$

where $\mathcal{P}f$ is the polygonal function whose graph has vertices $(t_k, f(t_k))$ for $0 \le k \le m$. Since $f_n \to f$ uniformly, choose N such that

$$\|f_n - f\|_\infty < \frac{\epsilon}{8m}$$

for all $n \geq N$. For such an n, by Lemma XIX.3.3.7. we have, for each k,

$$\int_{t_{k-1}}^{t_k} |f_n' - (\mathcal{P}f)'| \leq \left[\int_{t_{k-1}}^{t_k} |f' - (\mathcal{P}f)'|\right] + 4\|f_n - f\|_\infty < \left[\int_{t_{k-1}}^{t_k} |f' - (\mathcal{P}f)'|\right] + \frac{\epsilon}{2m} .$$

Thus we have

$$\begin{aligned}
\int_a^b |f_n' - f'| &\leq \int_a^b |f_n' - (\mathcal{P}f)'| + \int_a^b |(\mathcal{P}f)' - f'| \\
&\leq \left[\sum_{k=1}^m \int_{t_{k-1}}^{t_k} |f_n' - (\mathcal{P}f)'|\right] + \frac{\epsilon}{4} \\
&< \sum_{k=1}^m \left[\left(\int_{t_{k-1}}^{t_k} |f' - (\mathcal{P}f)'|\right) + \frac{\epsilon}{2m}\right] + \frac{\epsilon}{4} \\
&= \int_a^b |f' - (\mathcal{P}f)'| + \frac{\epsilon}{2} + \frac{\epsilon}{4} < \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon .
\end{aligned}$$



If the limit function is differentiable on I, we can say something stronger:

XIX.3.3.8. PROPOSITION. Let (f_n) be a sequence of convex functions on a closed bounded interval $[a, b]$, converging pointwise on $[a, b]$ to a (convex) function f. If f is differentiable on $[a, b]$, then $(D_L f_n)$ and $(D_R f_n)$ converge uniformly to f' on $[a, b]$ (i.e. (f_n') "converges uniformly" to f' on $[a, b]$).

Indeed, under the hypotheses, f' is continuous on $[a, b]$, so the proposition follows from the next more general result, which is strongly reminiscent of Dini's Theorem ():

XIX.3.3.9. PROPOSITION. Let (g_n) be a sequence of nondecreasing functions on a closed bounded interval $[a, b]$ converging pointwise on $[a, b]$ to a (necessarily nondecreasing) continuous function g. Then $g_n \to g$ uniformly on $[a, b]$.

PROOF: Let $\epsilon > 0$. By continuity of g, choose a partition $\{a = t_0 < t_1 < \cdots < t_m = b\}$ of $[a, b]$ such that $g(t_k) - g(t_{k-1}) < \frac{\epsilon}{2}$ for $1 \leq k \leq m$. Choose N such that $|g_n(t_k) - g(t_k)| < \frac{\epsilon}{2}$ for all $n \geq N$ and $0 \leq k \leq m$. Then, if $t_{k-1} \leq x \leq t_k$, we have, for $n \geq N$,

$$g(t_k) - \epsilon < g(t_{k-1}) - \frac{\epsilon}{2} \leq g_n(t_{k-1}) \leq g_n(x) \leq g_n(t_k) \leq g(t_k) + \frac{\epsilon}{2}$$

and we also have

$$g(t_k) - \frac{\epsilon}{2} < g(t_{k-1}) \leq g(x) \leq g(t_k)$$

and thus it follows that $|g_n(x) - g(x)| < \epsilon$. This is true for every k, i.e. for every $x \in [a, b]$.



If f is not differentiable everywhere, we do not have $f_n' \to f'$ uniformly in general, even on closed bounded intervals contained in the interior of I:

XIX.3.3.10. EXAMPLE. Let $f(x) = |x|$, and for each n let $f_n(x) = |x|^{1+1/n}$. Then f and each f_n are convex, f_n is differentiable everywhere, with f_n' continuous, and $f_n \to f$ pointwise on $\mathbb{R}$ (and uniformly on any bounded subinterval). But f_n' does not converge uniformly to f' on any (deleted) interval around 0.

XIX.3.4. Jensen's Inequality

A simple induction on n, beginning with $n = 2$, shows:

XIX.3.4.1. PROPOSITION. If f is convex on an interval I, $x_1, \ldots, x_n \in I$, and $t_1, \ldots, t_n \geq 0$, $t_1 + \cdots + t_n = 1$, then $t_1x_1 + \cdots + t_nx_n \in I$ and

$$f(t_1x_1 + \cdots + t_nx_n) \leq t_1f(x_1) + \cdots + t_nf(x_n) \ .$$

The inequality is strict if the x_k are distinct and the t_k nonzero, unless f is linear on an interval containing the x_k.

This can be rephrased:

XIX.3.4.2. COROLLARY. Let f be convex on an interval I, $x_1, \ldots, x_n \in I$, and $s_1, \ldots, s_n \geq 0$. Then

$$\frac{s_1x_1 + \cdots + s_nx_n}{s_1 + \cdots + s_n} \in I$$

and

$$f\left(\frac{s_1x_1 + \cdots + s_nx_n}{s_1 + \cdots + s_n}\right) \leq \frac{s_1f(x_1) + \cdots + s_nf(x_n)}{s_1 + \cdots + s_n} \ .$$

The inequality is strict if the x_k are distinct and the s_k nonzero, unless f is linear on an interval containing the x_k.

For the proof, just set $t_k = \frac{s_k}{s_1+\cdots+s_n}$.

A particularly important sample application is:

XIX.3.4.3. COROLLARY. Let $x_1, \ldots, x_n$ be positive numbers, and $\lambda_1, \ldots, \lambda_n$ positive numbers with $\lambda_1 + \cdots + \lambda_n = 1$. Then

$$x_1^{\lambda_1}x_2^{\lambda_2}\cdots x_n^{\lambda_n} \leq \lambda_1x_1 + \cdots + \lambda_nx_n$$

with strict inequality unless all the x_k are equal.

PROOF: Note that $f(x) = e^x$ is convex on $\mathbb{R}$. Set $y_k = \log x_k$, so $x_k = e^{y_k}$. We then have

$$(e^{y_1})^{\lambda_1}\cdots(e^{y_n})^{\lambda_n} = e^{\lambda_1y_1+\cdots+\lambda_ny_n} \leq \lambda_1e^{y_1} + \cdots + \lambda_ne^{y_n}$$

which is the desired inequality. 👍

The special case where all $\lambda_k = \frac{1}{n}$ is

XIX.3.4.4. COROLLARY. Let $x_1, \ldots, x_n$ be positive numbers. Then the geometric mean of $\{x_1, \ldots, x_n\}$ is less than or equal to the arithmetic mean, i.e.

$$[x_1 x_2 \cdots x_n]^{1/n} \leq \frac{1}{n}[x_1 + \cdots + x_n]$$

with strict inequality unless all the x_k are equal.

Jensen's Inequality is a generalization of XIX.3.4.2.. If $(X, \mathcal{A}, \mu)$ is a finite measure space and f is a μ-integrable function, recall () that the *average value* of f is

$$av(f) = \frac{1}{\mu(X)} \int_X f \, d\mu \ .$$

XIX.3.4.5. THEOREM. [JENSEN'S INEQUALITY] Let $(X, \mathcal{A}, \mu)$ be a finite measure space, f a μ-integrable function taking values in an interval I, and ϕ a convex function on I. Let $c = av(f)$. Then $c \in I$, and if either ϕ is nonnegative on I or $\phi \circ f$ is μ-integrable, then

$$\phi(c) \leq \frac{1}{\mu(X)} \int_X \phi \circ f \, d\mu \ .$$

The inequality is strict unless f is constant a.e. or ϕ is linear on a subinterval containing the essential range of f.

The condition that ϕ is nonnegative or that $\phi \circ f$ is integrable is simply needed to define the right side of the inequality. (Integrability of $\phi \circ f$ is not automatic; cf. Exercise XIX.3.7.3..) Actually, in general it turns out that $(\phi \circ f)_-$ is always integrable (Exercise XIX.3.7.4.), so $\phi \circ f$ is integrable in the extended sense of () and the right side is well defined as $+\infty$ if $\phi \circ f$ is not integrable, even if ϕ is not nonnegative.

PROOF: If ϕ is nonnegative but $\phi \circ f$ is not integrable, the inequality is trivial, so it only needs proof in the case where $\phi \circ f$ is integrable. Let us assume this.

To see that $c \in I$, it suffices to show that if $a < f(x) < b$ for all $x \in X$, then $a < c < b$. But in this case the function $g(x) = f(x) - a$ is strictly positive on X, so $\int_X g \, d\mu > 0$ (), and similarly $h(x) = b - f(x)$ has $\int_X h \, d\mu > 0$. In fact c is in the interior of I unless $f(x)$ is equal almost everywhere to one of the endpoints of I, in which case the inequality is a trivial equality.

Let m be any number between $D_L\phi(c)$ and $D_R\phi(c)$ (), and let $L(t) = \phi(c) + m(t - c)$ for $x \in I$. Then L is a linear function, and $L(t) \leq \phi(t)$ for all $t \in I$. Thus we have

$$\phi(c)\mu(X) = \int_X [\phi(c) + m(f(x) - c)] \, d\mu(x) = \int_X L(f(x)) \, d\mu(x) \leq \int_X \phi(f(x)) \, d\mu(x) \ .$$

We have equality if and only if $L \circ f = \phi \circ f$ μ-a.e., which is equivalent to the condition in the last line of the statement.

XIX.3.4.6. The inequality XIX.3.4.2. can be regarded as a special case of Jensen's Inequality, where X is an n-point space and $\{s_1, \ldots, s_n\}$ are the masses of the points.

Here are some of the other most commonly stated special cases of Jensen's Inequality:

XIX.3.4.7. Corollary. Let f be integrable on $[a,b]$ with values in an interval I, and let ϕ be convex on I. Let

$$c = \frac{1}{b-a}\int_a^b f\,d\lambda\ .$$

Then $c \in I$, and if ϕ is nonnegative or $\phi \circ f$ is integrable on $[a,b]$, then

$$\phi(c) \le \frac{1}{b-a}\int_a^b \phi\circ f\,d\lambda\ .$$

A slightly more general version allows a weight function:

XIX.3.4.8. Corollary. Let f be Lebesgue measurable on $[a,b]$ with values in an interval I, and let ϕ be convex on I. Let ω be a nonnegative Lebesgue integrable function on $[a,b]$, with $f\omega$ integrable on $[a,b]$. Set

$$c = \frac{1}{\int_a^b \omega\,d\lambda}\int_a^b f\omega\,d\lambda\ .$$

Then $c \in I$, and if ϕ is nonnegative or $(\phi\circ f)\omega$ is integrable on $[a,b]$, then

$$\phi(c) \le \frac{1}{\int_a^b \omega\,d\lambda}\int_a^b (\phi\circ f)\omega\,d\lambda\ .$$

If f is continuous, then everything can be stated and proved in terms of Riemann integration:

XIX.3.4.9. Corollary. Let f be continuous on $[a,b]$ with values in an interval I, and let ϕ be convex and continuous on I. Let

$$c = \frac{1}{b-a}\int_a^b f(x)\,dx.$$

Then $c \in I$, and

$$\phi(c) \le \frac{1}{b-a}\int_a^b \phi(f(x))\,dx\ .$$

An examination of the proof of XIX.3.4.5. shows that only Riemann integrals need be considered in this case. Without the continuity assumption on ϕ, $\phi\circ f$ can fail to be Riemann-integrable on $[a,b]$ (Exercise XIX.3.7.5.).

XIX.3.5. Log-Convex Functions

XIX.3.5.1. Definition. If I is an interval, a function $g : I \to (0,\infty)$ is *log-convex* on I if $\log g$ is convex on I.

XIX.3.5.2. Alternatively, a positive function g on an interval I is log-convex on I if, for all $a, b \in I$, $0 \leq \lambda \leq 1$,

$$g(\lambda a + (1-\lambda)b) \leq g(a)^{\lambda} g(b)^{1-\lambda} \ .$$

The right-hand side is a sort of weighted geometric mean of $g(a)$ and $g(b)$. If g is continuous, then by XIX.3.2.3. g is log-convex if and only if, for every $a, b \in I$,

$$g\left(\frac{a+b}{2}\right) \leq \sqrt{g(a)g(b)} \ .$$

XIX.3.5.3. Proposition. A log-convex function is convex.

Proof: This follows immediately from XIX.3.5.2. and XIX.3.4.3.. 👍

XIX.3.5.4. The converse is false. $g(x) = x^2$ is convex on $(0, \infty)$, but not log-convex on any subinterval. In fact, an increasing log-convex function on an interval (a, ∞) must grow at least exponentially as $x \to +\infty$.

XIX.3.5.5. Since a sum of convex functions is convex, it is obvious that a product of log-convex functions is log-convex. It is also clear that a positive scalar multiple of a log-convex function is log-convex. It is less obvious, but true, that a sum of log-convex functions is log-convex (XIX.3.5.7.).

XIX.3.5.6. Lemma. Let $a_1, a_2, b_1, b_2, c_1, c_2$ be positive real numbers. If $a_1c_1 - b_1^2 \geq 0$ and $a_2c_2 - b_2^2 \geq 0$, then

$$(a_1 + a_2)(c_1 + c_2) - (b_1 + b_2)^2 \geq 0 \ .$$

Proof: If a, b, c are positive real numbers, then the polynomial $f(x) = ax^2 + 2bx + c$ takes only nonnegative real values (for x real) if and only if $ac - b^2 \geq 0$. By assumption, the polynomials $f_1(x) = a_1x^2 + 2b_1x + c_1$ and $f_2(x) = a_2x^2 + 2b_2x + c_2$ take only nonnegative real values, and hence

$$f_1(x) + f_2(x) = (a_1 + a_2)x^2 + 2(b_1 + b_2)x + (c_1 + c_2)$$

also takes only nonnegative real values. 👍

XIX.3.5.7. Proposition. A sum of log-convex functions on an interval I is log-convex on I.

Proof: Since log-convex functions are continuous, by XIX.3.5.2. it suffices to show that if f and g are log-convex on I and $x, y \in I$, then

$$\left[f\left(\frac{x+y}{2}\right) + g\left(\frac{x+y}{2}\right)\right]^2 \leq (f(x) + g(x))(f(y) + g(y)) \ .$$

Apply XIX.3.5.6. with $a_1 = f(x)$, $a_2 = g(x)$, $b_1 = f\left(\frac{x+y}{2}\right)$, $b_2 = g\left(\frac{x+y}{2}\right)$, $c_1 = f(y)$, $c_2 = g(y)$. 👍

XIX.3.5.8. The conclusion of XIX.3.5.7. extends to arbitrary finite sums. Since a positive pointwise limit of log-convex functions is obviously log-convex, it also extends to a positive pointwise limit of finite sums of log-convex functions, e.g. a pointwise-convergent infinite series. By approximating the integrals by Riemann sums, we also obtain:

XIX.3.5.9. PROPOSITION. Let f be a function on $I \times [a,b]$ for some interval I and $a,b \in \mathbb{R}$, $a < b$. If $f(x,t)$ is log-convex on I for each fixed t, and $f(x,t)$ is continuous on $[a,b]$ for each fixed x, then

$$g(x) = \int_a^b f(x,t)\,dt$$

is log-convex on I.

PROOF: Let $n \in \mathbb{N}$ and $h = \frac{b-a}{n}$. Then $g_n(x) = h(f(x,a)+f(x,a+h)+\cdots+f(x,a+(n-1)h)$ is log-convex, and $g_n \to g$ pointwise. Since f is strictly positive, it is obvious that g is also strictly positive. 👍

This result extends to improper integrals:

XIX.3.5.10. COROLLARY. Let f be a function on $I \times (a,b)$ for some interval I and $a,b \in [-\infty,\infty]$, $a < b$. If $f(x,t)$ is log-convex on I for each fixed t, $f(x,t)$ is continuous on (a,b) for each fixed x, and the improper Riemann integral $\int_a^b f(x,t)\,dt$ converges for all $x \in I$, then

$$g(x) = \int_a^b f(x,t)\,dt$$

is log-convex on I.

PROOF: Let (a_n) be a decreasing sequence with $a_n \to a$, (b_n) an increasing sequence with $b_n \to b$, $a_n < b_n$ for all n. Set

$$g_n(x) = \int_{a_n}^{b_n} f(x,t)\,dt\ .$$

Then g_n is log-convex on I for all n by XIX.3.5.9., and $g_n \to g$ pointwise on I. It is also clear that g is strictly positive. 👍

XIX.3.6. Higher-Dimensional Versions of Convexity

XIX.3.7. Exercises

XIX.3.7.1. Let $f : \mathbb{R} \to \mathbb{R}$ be an additive function, i.e. $f(x+y) = f(x) + f(y)$ for all $x,y \in \mathbb{R}$. Show that if there is any nontrivial closed bounded interval $[a,b]$ on which f is bounded or Lebesgue measurable, then f is linear on $\mathbb{R}$, i.e. there is an $m \in \mathbb{R}$ with $f(x) = mx$ for all $x \in \mathbb{R}$. [Show that f and $-f$ are midpoint-convex, and use XIX.3.2.7..]

XIX.3.7.2. [?] Let $\{x_i : i \in I\}$ be a Hamel basis () for $\mathbb{R}$ over $\mathbb{Q}$, i.e. for each $x \in \mathbb{R}$ there is a unique representation

$$x = \sum_{i \in I} c_i(x) x_i$$

where $c_i(x) \in \mathbb{Q}$ for each i and only finitely many $c_i(x)$ are nonzero. (The Axiom of Choice is needed here.)

(a) Show that each function c_i is $\mathbb{Q}$-linear, i.e. $c_i(x+y) = c_i(x) + c_i(y)$ and $c_i(ax) = ac_i(x)$ for all $x, y \in \mathbb{R}$, $a \in \mathbb{Q}$.

(b) Show that each c_i is a midpoint-convex function on $\mathbb{R}$ (and that $-c_i$ is also midpoint-convex).

(c) Show that c_i is discontinuous everywhere. [If $j \neq i$, then $\mathbb{Q}x_j$ and $x_i + \mathbb{Q}x_j$ are both dense in $\mathbb{R}$.]

(d) Conclude that c_i is not convex on any interval. It then follows from XIX.3.2.4. that c_i is unbounded on every interval (this can be seen directly by noting that if $j \neq i$, then $nx_i + \mathbb{Q}x_j$ is dense in $\mathbb{R}$ for each $n \in \mathbb{Z}$). Conclude also that the restriction of c_i to any interval is not Lebesgue measurable.

XIX.3.7.3. Let $X = (0,1)$ with Lebesgue measure. Let $f(x) = x$, $I = (0,1)$, and $\phi(x) = e^{1/x}$. Show that X, f, I, ϕ satisfy the hypotheses of Jensen's Inequality, but that $\phi \circ f$ is not integrable on X.

XIX.3.7.4. Let $(X, \mathcal{A}, \mu)$ be a finite measure space, f a μ-integrable function on X taking values in an integral I, and ϕ a convex function on I. Show that $(\phi \circ f)_-$ is μ-integrable. [There is a linear function L such that $\phi \geq L$ on I.]

XIX.3.7.5. Let K be a Cantor set of positive measure (), and f a continuous function from $[0,1]$ to $[0,1]$ which takes the value 1 precisely on K, e.g. $f(x) = 1 - dist(x, K)$. Let ϕ be the convex function on $[0,1]$ defined in XIX.3.1.8.(v). Then $\phi \circ f$ is not Riemann-integrable on $[0,1]$.

XIX.3.7.6. Let $\mathcal{F}$ be a collection of convex functions on an interval $[a,b]$, and $[c,d]$ a closed bounded subinterval of (a,b).

(a) If the functions in $\mathcal{F}$ are uniformly bounded above at a and b, and there is an $x_0 \in (a,b)$ such that the functions in $\mathcal{F}$ are uniformly bounded below at x_0, show that $\mathcal{F}$ is uniformly bounded on $[a,b]$.

(b) Under the hypotheses of (a), show that there are real numbers m and M such that

$$m \leq D_L f(x) \leq D_R f(x) \leq M$$

for all $f \in \mathcal{F}$ and all $x \in [c,d]$. Conclude that $\mathcal{F}$ is uniformly equicontinuous () on $[c,d]$. In particular, $\mathcal{F}$ is equicontinuous on (a,b).

(c) If (f_n) is a sequence of convex functions on $[a,b]$ converging pointwise on $[a,b]$ to a (convex) function f, use (a), (b), and the Arzela-Ascoli Theorem () to conclude that $f_n \to f$ uniformly on $[c,d]$ (cf. XIX.3.3.1.).

(d) If (f_n) is a sequence of continuous convex functions on $[a,b]$ converging pointwise on $[a,b]$ to a continuous (convex) function f, show that the set

$$\mathcal{F} = \{f_n : n \in \mathbb{N}\}$$

is uniformly bounded and equicontinuous on $[a,b]$.

(e) If (f_n) is a sequence of convex functions on $[a,b]$ converging pointwise on $[a,b]$ to a function f, use (d) and the Arzela-Ascoli Theorem () to give an alternate proof of XIX.3.3.1..

XIX.3.7.7. Let f be convex on an interval I with right endpoint b, and suppose f is continuous at b. Show that if $\sup_{x<b} D_L f(x) = m < \infty$, then $D_L f(b) = m$.

XIX.3.7.8. Let f be convex on $[a, +\infty)$ for some a. Show that

$$\lim_{x\to+\infty} \frac{f(x)}{x}$$

exists (possibly as $+\infty$, but not $-\infty$) and equals

$$\lim_{x\to+\infty} D_L f(x) = \lim_{x\to+\infty} D_R f(x) = \sup_{x>a} D_L f(x) = \sup_{x\geq a} D_R f(x) \ .$$

Prove a similar statement for $-\infty$.

XIX.3.7.9. Let f be a function on an interval I taking only positive values, with f'' defined everywhere on I.

(a) Show that f is log-convex on I if and only if

$$f(x)f''(x) \geq [f'(x)]^2$$

for all $x \in I$. [Set $g(x) = \log f(x)$, and compute g''.]

(b) What if f'' is only defined on the complement of a countable set? What if f'' is only defined almost everywhere?

XIX.3.7.10. Let f and g be convex functions on an interval of the form $(a, +\infty)$ for some a. Show that if

$$\lim_{n\to\infty} [f(n) - g(n)] = 0$$

(limit for $n \in \mathbb{N}$), then

$$\lim_{x\to+\infty} [f(x) - g(x)] = 0$$

(limit for $x \in \mathbb{R}$).

XIX.4. Functions of Bounded Variation

Functions of bounded variation are ones which vary in a controlled way. Geometrically, a function has bounded variation on an interval if its graph has finite arc length. Functions of bounded variation are well-behaved, with perhaps unexpectedly nice properties which are technically useful.

XIX.4.1. The Total Variation of a Function

XIX.4.1.1. In the definition, we will use the positive and negative part notation: if x is a real number, write $x_+ = \max(x, 0) = \frac{|x|+x}{2}$ and $x_- = \max(-x, 0) = \frac{|x|-x}{2}$, so that $x = x_+ - x_-$ and $|x| = x_+ + x_-$. Thus if y and z are real numbers, we will write $(y-z)_+$ for $\max(y-z, 0)$ and $(y-z)_- = \max(z-y, 0)$. Thus $(y-z)_+ = y - z$ if $y \geq z$ and 0 otherwise, and $(y-z)_- = |y-z|$ if $y \leq z$ and 0 otherwise. We always have $(y-z)_+ - (y-z)_- = y - z$ and $(y-z)_+ + (y-z)_- = |y-z|$.

XIX.4.1.2. Proposition. If $x, y \in \mathbb{R}$, then $(x+y)_+ \leq x_+ + y_+$ and $(x+y)_- \leq x_- + y_-$.

This can be proved brute-force on a case-by-case basis, or using the triangle inequality (Exercise ()).

XIX.4.1.3. Definition. Let f be a real-valued function on an interval $[a, b]$. If

$$\mathcal{P} = \{a = x_0, x_1, \ldots, x_n = b\}$$

is a partition of $[a, b]$, set

$$V^+(f, \mathcal{P}) = \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})]_+$$

$$V^-(f, \mathcal{P}) = \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})]_-$$

$$V(f, \mathcal{P}) = V^+(f, \mathcal{P}) + V^-(f, \mathcal{P}) = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \ .$$

$V^+(f, \mathcal{P})$, $V^-(f, \mathcal{P})$, and $V(f, \mathcal{P})$ are respectively the *positive variation*, *negative variation*, and *total variation* of f over the partition $\mathcal{P}$.

These numbers of course depend on the interval $[a, b]$, which is implicitly coded in the partition $\mathcal{P}$.

XIX.4.1.4. Proposition. If $\mathcal{P}$ is any partition of $[a, b]$, then $V^+(f, \mathcal{P})$ and $V^-(f, \mathcal{P})$ are nonnegative real numbers, and

$$V^+(f, \mathcal{P}) - V^-(f, \mathcal{P}) = f(b) - f(a) \ .$$

Proof: The sums are obviously nonnegative and finite. We have

$$V^+(f, \mathcal{P}) - V^-(f, \mathcal{P}) = \sum_{k-1}^{n} ([f(x_k) - f(x_{k-1})]_+ - [f(x_k) - f(x_{k-1})]_-)$$

$$= \sum_{k=1}^{n} [f(x_k) - f(x_{k-1})] = f(b) - f(a)$$

by telescoping. 👍

XIX.4.1.5. PROPOSITION. If $\mathcal{P}$ is a partition of $[a,b]$ and $\mathcal{Q}$ is a refinement of $\mathcal{P}$, then $V^+(f,\mathcal{P}) \le V^+(f,\mathcal{Q})$, $V^-(f,\mathcal{P}) \le V^-(f,\mathcal{Q})$, and $V(f,\mathcal{P}) \le V(f,\mathcal{Q})$.

PROOF: It suffices to prove the result when $\mathcal{Q}$ has just one additional point y. Fix k with $x_{k-1} < y < x_k$ (where $\mathcal{P} = \{x_0, \dots, x_n\}$). For V^+, we need only show that

$$[f(x_k) - f(x_{k-1})]_+ \le [f(x_k) - f(y)]_+ + [f(y) - f(x_{k-1})]_+$$

which follows immediately from XIX.4.1.2.. The proof for V^- is almost identical (or can be immediately obtained by noting that $V^-(f,\mathcal{P}) = V^+(-f,\mathcal{P})$ for any f and $\mathcal{P}$), and the inequalities add for V. 👍

XIX.4.1.6. DEFINITION. Let f be a real-valued function on an interval $[a,b]$. Define

$$V^+_{[a,b]}(f) = \sup_{\mathcal{P}} V^+(f,\mathcal{P})$$

$$V^-_{[a,b]}(f) = \sup_{\mathcal{P}} V^-(f,\mathcal{P})$$

$$V_{[a,b]}(f) = \sup_{\mathcal{P}} V(f,\mathcal{P})$$

where in each case the supremum is over all partitions $\mathcal{P}$ of $[a,b]$. $V^+_{[a,b]}(f)$, $V^-_{[a,b]}(f)$, $V_{[a,b]}(f)$ are called the *total positive variation*, *total negative variation*, and *total variation* of f over $[a,b]$. These are nonnegative extended real numbers.

The function f has *bounded variation* on $[a,b]$ if $V_{[a,b]}(f) < +\infty$.

XIX.4.1.7. EXAMPLES. (i) Any monotone function on an interval $[a,b]$ has bounded variation. If f is nondecreasing on $[a,b]$, then $V^+_{[a,b]}(f) = V_{[a,b]}(f) = f(b) - f(a)$ and $V^-_{[a,b]}(f) = 0$; if f is nonincreasing on $[a,b]$, then $V^-_{[a,b]}(f) = V_{[a,b]}(f) = f(a) - f(b)$ and $V^+_{[a,b]}(f) = 0$.

(ii) A Lipschitz function on $[a,b]$ has bounded variation. If f is Lipschitz with Lipschitz constant M, i.e. $|f(x) - f(y)| \le M|x - y|$ for all $x, y \in [a,b]$, then if $\mathcal{P} = \{a = x_0, x_1, \dots, x_n = b\}$ is any partition of $[a,b]$, we have

$$V(f,\mathcal{P}) = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \le \sum_{k=1}^{n} M|x_k - x_{k-1}| = M(b-a)$$

so $V_{[a,b]}(f) \le M(b-a)$. In particular, if f is differentiable on $[a,b]$ and f' is bounded, then f has bounded variation on $[a,b]$.

(iii) A continuous function need not have bounded variation. Continuous but everywhere nondifferentiable functions cannot have bounded variation (XIX.4.2.4.(iv)). Even a function which is differentiable everywhere need not have bounded variation. Let

$$f(x) = x^2 \sin\left(\frac{1}{x^2}\right)$$

for $x \neq 0$, $f(0) = 0$ (Figure V.4). Then f is differentiable everywhere, but not of bounded variation on $[0,1]$. For each n, let

$$x_n = \frac{1}{\sqrt{\frac{\pi}{2} + \pi(n-1)}} \ .$$

Then

$$f(x_n) = \pm \frac{1}{\frac{\pi}{2} + \pi(n-1)}$$

where the signs alternate, so

$$|f(x_{n+1}) - f(x_n)| = \frac{1}{\frac{\pi}{2} + \pi n} + \frac{1}{\frac{\pi}{2} + \pi(n-1)} \geq \frac{1}{\pi n}$$

so if $\mathcal{P}$ runs over partitions containing large initial finite segments of $\{x_n : n \in \mathbb{N}\}$, $V(f, \mathcal{P})$ is arbitrarily large since the harmonic series diverges (IV.2.2.11.). Thus $V_{[0,1]}(f) = +\infty$. The problem is, roughly speaking (even precisely speaking; cf. XIX.4.2.5.), that f' is not Lebesgue integrable on $[0,1]$.

It turns out that if f is differentiable everywhere on $[a,b]$ and f' is Lebesgue integrable on $[a,b]$, then f has bounded variation on $[a,b]$, but this requires considerable work to prove (). This cannot, however, be relaxed to just requiring that f be continuous, differentiable almost everywhere, and with integrable derivative ().

Example (i) shows that a function of bounded variation need not be continuous. But it must be "almost continuous": it can have only countably many discontinuities, each of which is a jump discontinuity (XIX.4.2.4.).

XIX.4.1.8. PROPOSITION. Let f be a real-valued function on an interval $[a,b]$, and let Γ_f be its graph

$$\Gamma_f = \{(x, f(x)) : a \leq x \leq b\} \ .$$

Then f has bounded variation on $[a,b]$ if and only if the arc length () $\ell(\Gamma_f)$ of Γ_f is finite. In fact,

$$V_{[a,b]}(f) \leq \ell(\Gamma_f) \leq (b-a) + V_{[a,b]}(f) \ .$$

PROOF: We have that

$$\ell(\Gamma_f) = \sup_{\mathcal{P}} \left\{ \sum_{k=1}^{n} \|(x_k, f(x_k)) - (x_{k-1}, f(x_{k-1}))\| \right\}$$

$$= \sup_{\mathcal{P}} \left\{ \sum_{k=1}^{n} \sqrt{[x_k - x_{k-1}]^2 + [f(x_k) - f(x_{k-1})]^2} \right\}$$

$$\leq \sup_{\mathcal{P}} \left\{ \sum_{k=1}^{n} \left([x_k - x_{k-1}] + |f(x_k) - f(x_{k-1})| \right) \right\}$$
$$= \sup_{\mathcal{P}} \left\{ (b-a) + \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \right\}$$
$$= \sup_{\mathcal{P}} \{(b-a) + V(f, \mathcal{P})\} = (b-a) + V_{[a,b]}(f)$$

where the suprema are over all partitions $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$ of $[a, b]$. The first inequality is evident. 👍

The first inequality is always strict if f has bounded variation, but the second inequality can be an equality ().

XIX.4.1.9. Since a function f of bounded variation need not be continuous, the graph of f need not be an arc (a homeomorphic image of $[0, 1]$). We normally only consider arc length for arcs. But since a function of bounded variation has only jump discontinuities, and only finitely many jumps of length more than ϵ for any $\epsilon > 0$, the graph of f can be expanded to an arc by adding vertical line segments across the jumps. The arc length of this arc is still no more than $(b-a) + V_{[a,b]}(f)$.

We now explore properties of the variations of a function.

XIX.4.1.10. Proposition. Let f be a real-valued function on an interval $[a, b]$. Then

$$V^+_{[a,b]}(f) + V^-_{[a,b]}(f) = V_{[a,b]}(f)$$

and, if f has bounded variation on $[a, b]$, then

$$V^+_{[a,b]}(f) - V^-_{[a,b]}(f) = f(b) - f(a) \ .$$

If f does not have bounded variation, then $V^+_{[a,b]}(f) = V^-_{[a,b]}(f) = +\infty$.

Proof: If $\mathcal{P}$ is any partition of $[a, b]$, we have

$$V(f, \mathcal{P}) = V^+(f, \mathcal{P}) + V^-(f, \mathcal{P}) \leq V^+_{[a,b]}(f) + V^-_{[a,b]}(f)$$

and thus $V_{[a,b]}(f) \leq V^+_{[a,b]}(f) + V^-_{[a,b]}(f)$. For the opposite inequality, there is nothing to prove if $V_{[a,b]}(f) = +\infty$, so assume $V_{[a,b]}(f)$ is finite. Since $V^+_{[a,b]}(f) \leq V_{[a,b]}(f)$ and $V^-_{[a,b]}(f) \leq V_{[a,b]}(f)$, they are both also finite. Let $\epsilon > 0$. Choose a partition $\mathcal{P}$ fine enough that $V^+(f, \mathcal{P}) \geq V^+_{[a,b]}(f) - \epsilon$ and $V^-(f, \mathcal{P}) \geq V^-_{[a,b]}(f) - \epsilon$. Then

$$V^+_{[a,b]}(f) + V^-_{[a,b]}(f) - 2\epsilon \leq V^+(f, \mathcal{P}) + V^-(f, \mathcal{P}) = V(f, \mathcal{P}) \leq V_{[a,b]}(f)$$

and since $\epsilon > 0$ is arbitrary, $V^+_{[a,b]}(f) + V^-_{[a,b]}(f) \leq V_{[a,b]}(f)$.

For the last equation, suppose f has bounded variation. Let $\epsilon > 0$, and fix a partition $\mathcal{P}$ fine enough that $|V^+_{[a,b]}(f) - V^+(f, \mathcal{P})| < \frac{\epsilon}{2}$ and $|V^-_{[a,b]}(f) - V^-(f, \mathcal{P})| < \frac{\epsilon}{2}$. Since by XIX.4.1.4. we have $V^+(f, \mathcal{P}) - V^-(f, \mathcal{P}) = f(b) - f(a)$, we have

$$|[V^+_{[a,b]}(f) - V^-_{[a,b]}(f)] - [f(b) - f(a)]|$$

$$= |[V^+(f,\mathcal{P}) - V^-(f,\mathcal{P})] - [f(b) - f(a)] + [V^+_{[a,b]}(f) - V^+(f,\mathcal{P})] + [V^-(f,\mathcal{P}) - V^-_{[a,b]}(f)]|$$

$$\leq |[V^+(f,\mathcal{P}) - V^-(f,\mathcal{P})] - [f(b) - f(a)]| + |V^+_{[a,b]}(f) - V^+(f,\mathcal{P})| + |V^-(f,\mathcal{P}) - V^-_{[a,b]}(f)| < 0 + \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

and, since $\epsilon > 0$ is arbitrary, $V^+_{[a,b]}(f) - V^-_{[a,b]}(f) = f(b) - f(a)$.

Finally, if $V_{[a,b]}(f) = +\infty$, then at least one of $V^+_{[a,b]}(f)$ and $V^-_{[a,b]}(f)$ is infinite. Suppose $V^+_{[a,b]}(f) = +\infty$ and $V^-_{[a,b]}(f) < +\infty$. If $M \in \mathbb{R}$, fix a partition $\mathcal{P}$ such that $V^+(f,\mathcal{P}) \geq M + V^-_{[a,b]}(f)$. Then

$$f(b) - f(a) = V^+(f,\mathcal{P}) - V^-(f,\mathcal{P}) \geq V^+(f,\mathcal{P}) - V^-_{[a,b]}(f) \geq M \ .$$

This is a contradiction since M is arbitrary. The other case is similar. 👍

The next simple result is used repeatedly.

XIX.4.1.11. PROPOSITION. Let f be a real-valued function on $[a,b]$, and $c \in (a,b)$. Then $V^+_{[a,b]}(f) = V^+_{[a,c]}(f) + V^+_{[c,b]}(f)$, $V^-_{[a,b]}(f) = V^-_{[a,c]}(f) + V^-_{[c,b]}(f)$, and $V_{[a,b]}(f) = V_{[a,c]}(f) + V_{[c,b]}(f)$.

PROOF: Since $V^+(f,\mathcal{P})$ increases as $\mathcal{P}$ becomes finer, in computing $V^+_{[a,b]}(f)$ we may restrict to partitions of $[a,b]$ containing c. Any such $\mathcal{P}$ splits into a partition $\mathcal{P}_1$ of $[a,c]$ and a partition $\mathcal{P}_2$ of $[c,b]$. Conversely, any partitions $\mathcal{P}_1$ of $[a,c]$ and $\mathcal{P}_2$ of $[c,b]$ combine into a partition $\mathcal{P}$ of $[a,b]$ containing c. For any such $\mathcal{P}$, we have $V^+(f,\mathcal{P}) = V^+(f,\mathcal{P}_1) + V^+(f,\mathcal{P}_2)$. Taking suprema over $\mathcal{P}_1$ and $\mathcal{P}_2$ corresponds to taking supremum over $\mathcal{P}$. The argument for V^- and V is identical. 👍

XIX.4.1.12. COROLLARY. If f has bounded variation on $[a,b]$, and $[c,d]$ is any (closed) subinterval of $[a,b]$, then f has bounded variation on $[c,d]$, and $V^+_{[c,d]}(f) \leq V^+_{[a,b]}(f)$, $V^-_{[c,d]}(f) \leq V^-_{[a,b]}(f)$, $V_{[c,d]}(f) \leq V_{[a,b]}(f)$.

XIX.4.1.13. PROPOSITION. Let f and g be real-valued functions on $[a,b]$, and $\alpha \in \mathbb{R}$. Then

(i) $V^+_{[a,b]}(f+g) \leq V^+_{[a,b]}(f) + V^+_{[a,b]}(g)$, $V^-_{[a,b]}(f+g) \leq V^-_{[a,b]}(f) + V^-_{[a,b]}(g)$, $V_{[a,b]}(f+g) \leq V_{[a,b]}(f) + V_{[a,b]}(g)$.

(ii) $V_{[a,b]}(\alpha f) = |\alpha| V_{[a,b]}(f)$. If $\alpha \geq 0$, then $V^+_{[a,b]}(\alpha f) = \alpha V^+_{[a,b]}(f)$ and $V^-_{[a,b]}(\alpha f) = \alpha V^-_{[a,b]}(f)$; if $\alpha < 0$, then $V^+_{[a,b]}(\alpha f) = |\alpha| V^-_{[a,b]}(f)$ and $V^-_{[a,b]}(\alpha f) = |\alpha| V^+_{[a,b]}(f)$.

Part (i) follows immediately from XIX.4.1.2., and the simple proof of (ii) is left as an exercise.

XIX.4.1.14. This result shows that if $BV[a,b]$ is the set of functions of bounded variation on $[a,b]$, then $BV[a,b]$ is a vector space of functions and $\|f\|_{BV} = |f(a)| + V_{[a,b]}(f)$ is a norm on $BV[a,b]$. In fact, $BV[a,b]$ is a Banach space under this norm ().

XIX.4.2. The Total Variation Functions

We can use the results of the previous section to define positive, negative, and total variation functions for any function of bounded variation, which yield strong structure results about such functions.

XIX.4.2.1. DEFINITION. Let f be a function of bounded variation on $[a,b]$. For $x \in [a,b]$, define

$$T^+f(x) = V^+_{[a,x]}(f)$$

$$T^-f(x) = V^-_{[a,x]}(f)$$

$$Tf(x) = V_{[a,x]}(f) = T^+f(x) + T^-f(x)\ .$$

T^+f, T^-f, and Tf are called the *positive variation function*, the *negative variation function*, and the *total variation function* of f on $[a,b]$.

These functions depend (up to constants) on the choice of left endpoint of the interval $[a,b]$; when it is necessary to specify the starting point a, we will write $T^+_a f$, $T^-_a f$, $T_a f$.

XIX.4.2.2. The functions T^+f, T^-f, Tf are nonnegative and nondecreasing by XIX.4.1.12.. We define all to be 0 at a. By XIX.4.1.4., $T^+f(x) - T^-f(x) = f(x) - f(a)$ for all $x \in [a,b]$. Thus we obtain the following important consequence:

XIX.4.2.3. THEOREM. Every function of bounded variation on an interval $[a,b]$ can be written as a difference of two nondecreasing functions on $[a,b]$.

In fact, if f has bounded variation, then $f = (f(a) + T^+f) - T^-f$.

Because of the known properties of nondecreasing functions, we obtain:

XIX.4.2.4. COROLLARY. Let f be a function of bounded variation on an interval $[a,b]$. Then

(i) f is Borel measurable ().

(ii) f has left and right limits at every point of $[a,b]$ (XIX.2.1.7.).

(iii) f has only countably many discontinuities on $[a,b]$, each of which is a jump discontinuity (XIX.2.1.10.).

(iv) f is differentiable a.e. on $[a,b]$ (XIX.2.2.12.).

XIX.4.2.5. PROPOSITION. Let f be a function of bounded variation on $[a,b]$. Then f' (which is defined a.e. on $[a,b]$) is integrable on $[a,b]$, and

$$\int_a^b |f'|\,d\lambda \le V_{[a,b]}(f)\ .$$

PROOF: Extend f to $[a,b+1]$ by setting $f(x)=f(b)$ for $x\in[b,b+1]$. For each n define

$$f_n(x)=n\left[f\left(x+\frac{1}{n}\right)-f\left(x\right)\right]\ .$$

Then each f_n is (Borel) measurable, and $f_n(x)\to f'(x)$ for any $x\in(a,b)$ for which f is differentiable at x. Thus $f_n\to f'$ pointwise a.e., so f' is measurable. Also, since T^+f, T^-f, Tf are nondecreasing, they are differentiable a.e., with nonnegative derivatives, and since $f=(f(a)+T^+f)-T^-f$, we have $f'=(T^+f)'-(T^-f)'$ a.e., and thus

$$|f'|\le (T^+f)'+(T^-f)'=(Tf)' \text{ a.e.}$$

By XIX.2.2.13.,

$$\int_a^b |f'|\,d\lambda \le \int_a^b (Tf)'\,d\lambda \le Tf(b)-Tf(a)=V_{[a,b]}(f)\ .$$



XIX.4.2.6. PROPOSITION. Let f be a function of bounded variation on $[a,b]$. Then the variation functions T^+f, T^-f, Tf are continuous wherever f is continuous.

PROOF: Let $f\in BV([a,b])$, and $x\in(a,b]$. We show that if f is left continuous at x, then T^+f is also left continuous at x. The other parts of the proof are essentially identical.

Let $\epsilon>0$, and choose $\delta>0$ so that if $x-\delta<y<x$, $|f(y)-f(x)|<\frac{\epsilon}{2}$. Choose a partition

$$\mathcal{P}=\{a=x_0,x_1,\dots,x_n,x_{n+1}=x\}$$

of $[a,x]$ with $n+1$ subintervals such that

$$T^+f(x)=V^+_{[a,x]}(f)<V^+(f,\mathcal{P})+\frac{\epsilon}{2}$$

$$V^+(f,\mathcal{P})>T^+f(x)-\frac{\epsilon}{2}\ .$$

By adding an additional point to $\mathcal{P}$ if necessary (which can only increase $V^+(f,\mathcal{P})$), we may assume the next-to-last point x_n satisfies $x_n>x-\delta$. Let $\mathcal{Q}$ be the partition $\{a=x_0,x_1,\dots,x_n\}$ of $[a,x_n]$. Then we have

$$V^+(f,\mathcal{P})=\left(\sum_{k=1}^{n}[f(x_k)-f(x_{k-1})]_+\right)+[f(x)-f(x_n)]_+=V^+(f,\mathcal{Q})+[f(x)-f(x_n)]_+<V^+(f,\mathcal{Q})+\frac{\epsilon}{2}\ .$$

If $x_n \le y \le x$, then

$$T^+f(x) - \epsilon = V^+_{[a,x]}(f) - \epsilon < V^+(f,\mathcal{P}) - \frac{\epsilon}{2} < V^+(f,\mathcal{Q}) \le V^+_{[a,x_n]}(f) = T^+f(x_n) \le T^+f(y) \le T^+f(x)$$

so $|T^+f(y) - T^+f(x)| < \epsilon$. Since $\epsilon > 0$ is arbitrary, T^+f is left continuous at x. 👍

XIX.4.2.7. Corollary. If f is a continuous function of bounded variation on $[a,b]$, then T^+f, T^-f, Tf are also continuous on $[a,b]$. Thus f can be written as a difference of two continuous nondecreasing functions on $[a,b]$.

XIX.4.3. Banach's Multiplicity Theorem

For continuous real-valued functions on a closed bounded interval, there is a close relationship between the notion of bounded variation and the Multiplicity Function (XVI.2.8.1.): a function is informally of bounded variation if its graph does not "wiggle up and down" too much, hence the function does not take "too many values too many times." This theorem is due to Banach ().

XIX.4.3.1. Theorem. Let f be a continuous function from a closed bounded interval $[a,b]$ to $\mathbb{R}$. Then N_f is finite a.e. and (Lebesgue) integrable on $[c,d] = f([a,b])$ if and only if f has bounded variation on $[a,b]$, and in any case

$$\int_c^d N_f \, d\lambda = V_{[a,b]}(f) \ .$$

Proof: Recall (XVI.2.8.4.) that N_f is nonnegative and (Borel) measurable, so $\int_c^d N_f \, d\lambda$ makes sense. For each n let ϕ_n be the function defined in the proof of XVI.2.8.4.. Then we have

$$\int_c^d \phi_n \, d\lambda = \sum_{k=1}^{2^n} \lambda(f(J_k)) = \sum_{k=1}^{2^n} \lambda(f(\bar{J}_k)) \le \sum_{k=1}^{2^n} V_{[x_{k-1},x_k]}(f) = V_{[a,b]}(f)$$

(where the x_k are the numbers from the proof of XVI.2.8.4., i.e. the endpoints of the intervals J_k). We have $\int_c^d \phi_n \, d\lambda \to \int_c^d N_f \, d\lambda$ by the Monotone Convergence Theorem. So if f has bounded variation on $[a,b]$, we have that N_f is integrable on $[c,d]$ and in particular finite a.e. on $[c,d]$, and

$$\int_c^d N_f \, d\lambda \le V_{[a,b]}(f) \ .$$

Conversely, suppose N_f is finite a.e. and integrable on $[c,d]$. For $a \le x \le b$, let f_x be the restriction of f to $[a,x]$, and set $S(x) = \int_c^d N_{f_x} \, d\lambda$. We have $N_{f_x} \le N_f$, and for $a \le x_1 \le x_2 \le b$ we have

$$0 = S(a) \le S(x_1) \le S(x_2) \le S(b) = \int_c^d N_f \, d\lambda$$

and, furthermore, since f takes every value between $f(x_1)$ and $f(x_2)$ on $[x_1, x_2]$ by the Intermediate Value Theorem (), we have $N_{f_{x_2}}(y) \geq N_{f_{x_1}}(y) + 1$ for all y between $f(x_1)$ and $f(x_2)$ and thus we have

$$S(x_2) - S(x_1) = \int_c^d (N_{f_{x_2}} - N_{f_{x_1}})\, d\lambda \geq \left| \int_{f(x_1)}^{f(x_2)} 1\, d\lambda \right| = |f(x_2) - f(x_1)| \ .$$

So if $a \leq x_0 < x_1 < \cdots < x_n \leq b$, we have

$$\int_c^d N_f\, d\lambda = S(b) - S(a) \geq \sum_{k=1}^{n} [S(x_k) - S(x_{k-1})] \geq \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|$$

and taking the supremum over all such finite subsets of $[a, b]$ we obtain

$$\int_c^d N_f\, d\lambda \geq V_{[a,b]}(f)$$

and thus f has bounded variation, and the opposite inequality holds by the first part of the proof, giving equality.

If f does not have bounded variation, equality still holds since both sides are $+\infty$. 👍

XIX.4.4. Extensions of Bounded Variation

Noncompact Intervals

We can extend the notion of bounded variation to noncompact intervals:

XIX.4.4.1. DEFINITION. Let I be an interval, and f a real-valued function on I. Define

$$V_I^+(f) = \sup V_{[a,b]}^+(f)$$

$$V_I^-(f) = \sup V_{[a,b]}^-(f)$$

$$V_I(f) = \sup V_{[a,b]}(f)$$

where the suprema are taken over all closed bounded subintervals $[a, b]$ of I.
The function f has *bounded variation on* I if $V_I(f) < +\infty$.

Most properties of bounded variation for closed bounded intervals carry over directly. We list some without proof; proofs are similar to the earlier ones and are left to the reader.

XIX.4.4.2. PROPOSITION. (i) $V_I(f) = V_I^+(f) + V_I^-(f)$.
(ii) If f has bounded variation on I, then f is bounded on I and

$$\sup_{x \in I} f(x) - \inf_{x \in I} f(x) \leq V_I(f) \ .$$

(iii) If J is a subinterval of I, then $V_J^+(f) \le V_I^+(f)$, $V_J^-(f) \le V_I^-(f)$, $V_J(f) \le V_I(f)$. If f has bounded variation on I, then f has bounded variation on J.

(iv) If f is bounded variation on I, and a and b are the left and right endpoints of I (we do not necessarily have $a, b \in I$; we can have $a = -\infty$ and/or $b = +\infty$), then $\lim_{x \to a+} f(x)$ and $\lim_{x \to b-} f(x)$ exist.

(v) If f has bounded variation on I, then f is Borel measurable, has left and right limits at every point, has only countably many discontinuities each of which is a jump discontinuity, and is differentiable a.e. on I.

(vi) If f has bounded variation on I, then f' is Lebesgue integrable on I and

$$\int_I |f'| \, d\lambda \le V_I(f) \ .$$

We can also define variation functions for a function of bounded variation:

XIX.4.4.3. DEFINITION. Let I be an interval, and a the left endpoint of I (we do not necessarily have $a \in I$; a can be $-\infty$). Define

$$T^+ f(x) = \begin{cases} V^+_{[a,x]}(f) & \text{if} \quad a \in I \\ V^+_{(a,x]}(f) & \text{if} \quad a \notin I \end{cases}$$

$$T^- f(x) = \begin{cases} V^-_{[a,x]}(f) & \text{if} \quad a \in I \\ V^-_{(a,x]}(f) & \text{if} \quad a \notin I \end{cases}$$

$$Tf(x) = \begin{cases} V_{[a,x]}(f) & \text{if} \quad a \in I \\ V_{(a,x]}(f) & \text{if} \quad a \notin I \end{cases}$$

for $x \in I$.

$T^+ f$, $T^- f$, Tf are nonnegative and nondecreasing; $Tf = T^+ f + T^- f$, $T^+ f - T^- f = f - C$, where $C = f(a)$ if $a \in I$ and $C = \lim_{x \to a+} f(x)$ if $a \notin I$.

Complex-Valued Functions

XIX.4.4.4. We can also define bounded variation for complex-valued functions. If f is a complex-valued function on $[a, b]$ and $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$, then $V^+(f, \mathcal{P})$ and $V^-(f, \mathcal{P})$ do not make sense, but we can still define

$$V(f, \mathcal{P}) = \sum_{k=1}^{n} |f(x_k) - f(x_{k-1})|$$

$$V_{[a,b]}(f) = \sup_{\mathcal{P}} V(f, \mathcal{P})$$

as before. Then f has bounded variation on $[a, b]$ if $V_{[a,b]}(f) < +\infty$.

XIX.4.4.5. If g and h are the real and imaginary parts of f, i.e. $f = g + ih$ with g, h real-valued, then we have

$$V_{[a,b]}(g), V_{[a,b]}(h) \le V_{[a,b]}(f) \le V_{[a,b]}(g) + V_{[a,b]}(h)$$

so f has bounded variation if and only if both g and h have bounded variation. (There is, however, no simple formula for $V_{[a,b]}(f)$ in terms of $V_{[a,b]}(g)$ and $V_{[a,b]}(h)$.) We still have:

XIX.4.4.6. PROPOSITION. Let f be a complex-valued function on $[a,b]$, and $c \in (a,b)$. Then

$$V_{[a,b]}(f) = V_{[a,c]}(f) + V_{[c,b]}(f) \ .$$

XIX.4.4.7. PROPOSITION. If f is a complex-valued function of bounded variation on $[a,b]$, then f is differentiable a.e. on $[a,b]$, f' is integrable on $[a,b]$, and

$$\int_a^b |f'| \, d\lambda \le V_{[a,b]}(f) \ .$$

XIX.4.4.8. We can also define the total variation function Tf as in the real case, which is nonnegative and nondecreasing, and continuous wherever f is continuous. If g and h are the real and imaginary parts of f, then

$$Tg, Th \le Tf \le Tg + Th$$

but there is no simple formula for Tf in terms of Tg and Th.

XIX.4.5. Exercises

XIX.4.5.1. Let f be a function of bounded variation on $[a,b]$. Show directly that f has one-sided limits at every point of $[a,b]$, as follows:

(a) Let $c \in (a,b]$. Suppose

$$\liminf_{x \to c-} f(x) < \limsup_{x \to c-} f(x) \ .$$

Choose $\alpha, \beta \in \mathbb{R}$ such that

$$\liminf_{x \to c-} f(x) < \alpha < \beta < \limsup_{x \to c-} f(x) \ .$$

Choose an increasing sequence (x_n) in (a,c) converging to c such that $f(x_n) < \alpha$ for n odd and $f(x_n) > \beta$ for n even.

(b) If $\mathcal{P} = \{a = x_0, x_1, \ldots, x_{n+1}, b\}$, show that $V(f, \mathcal{P}) > n(\beta - \alpha)$. Conclude that $V_{[a,b]}(f)$ must be $+\infty$.

(c) Give a similar argument for limits from the right.

(d) Conclude from V.8.6.7.(a) that f has only countably many discontinuities, each of which is a jump discontinuity.

XIX.4.5.2. Let f be the function of IX.2.9.10.(i). Show that f does not have bounded variation on $[0,1]$ (or on $[0,b]$ for any $b>0$). Note that f is differentiable everywhere, and f' is continuous except at 0.

XIX.4.5.3. Slightly modify the proof of XIX.4.2.6. to show that if $f \in BV([a,b])$, $x \in (a,b]$, and f is left lower semicontinuous at x, then T^+f is left continuous at x. State and prove similar statements for T^-f and for right semicontinuity.

XIX.4.5.4. Let I be an interval in $\mathbb{R}$, and f a one-to-one continuous function from I to $\mathbb{R}$. Then

(a) f is a homeomorphism onto its range J, which is also an interval (XIII.19.1.28., or give a more direct argument).

(b) f has bounded variation on any closed bounded subinterval of I (V.3.10.15.).

(c) f is differentiable a.e. on I (XIX.2.2.12.).

(d) If $\mathbb{R}$ is replaced by $\mathbb{R}^n$ for $n>1$, then (a) remains true (XIII.19.1.28.), but (b) does not make sense and (c) is false in general: let g be a continuous function from $\mathbb{R}$ to $\mathbb{R}$ which is not differentiable anywhere, and define $\boldsymbol{f}:\mathbb{R}^2\to\mathbb{R}^2$ by $\boldsymbol{f}(x,y)=(x,y+g(x))$. Then $\boldsymbol{f}$ is a homeomorphism from $\mathbb{R}^2$ onto $\mathbb{R}^2$ which is not differentiable anywhere.

XIX.5. Absolute Continuity and Lebesgue Density

A function which is differentiable everywhere on an interval $[a, b]$ need not have a (Lebesgue) integrable derivative, hence may not be the integral of its derivative (IX.2.9.10.(i)); the problem is that such a function need not have bounded variation. But even bounded variation is not sufficient to insure that a continuous function can be recovered from its derivative by integration if it is only differentiable almost everywhere. The Cantor function is the prototype example of a continuous monotone function (hence of bounded variation) which is differentiable almost everywhere, yet is not the integral of its derivative. In this section we discuss conditions which do suffice to recover a function by integration.

Applications of the results to indicator functions demonstrate a remarkable density property of Lebesgue measurable sets: such sets cannot be too uniformly distributed in the real line and must be either thick or sparse almost everywhere.

XIX.5.1. Luzin's Property (N)

XIX.5.1.1. The principal property of the Cantor function f which causes difficulties in the context of this section is that there are null sets N for which $f(N)$ is not a null set. In fact, the Cantor set K is a null set and $f(K)$ is the entire interval $[0, 1]$.

From the point of view of Borel measures, the measure on $[0, 1]$ whose cdf is the Cantor function is a probability measure which is continuous, but concentrated on K, and hence is not absolutely continuous with respect to Lebesgue measure.

We thus make the following definition. This condition was first defined and studied by N. Luzin, who appreciated its importance and relevance to the problem of reconstructing a function by integration.

XIX.5.1.2. Definition. Let I be an interval, and $f : I \to \mathbb{R}$. Then f has *Property (N)* (or *Luzin's Property (N)*) on I if, whenever N is a subset of I with $\lambda(N) = 0$, we have $\lambda(f(N)) = 0$ (i.e. f sends λ-null sets to λ-null sets).

The next result is an immediate rephrasing of the definition.

XIX.5.1.3. Proposition. Let μ be a continuous finite Borel measure on $[a, b]$, with cdf F. Then $\mu \ll \lambda$ if and only if F has property (N).

The corresponding statement for signed measures is false ($\mu \ll \lambda$ implies Property (N) is true, but the converse fails in general), and it is quite tricky to get the right condition. See ().

XIX.5.1.4. Proposition. [**?**] Let I be an interval, and $f : I \to \mathbb{R}$ a continuous function. Then f has property (N) if and only if, whenever A is a Lebesgue measurable subset of I, $f(A)$ is Lebesgue measurable.

Proof: Suppose f has property (N). If $A \subseteq I$ is Lebesgue measurable, then $A = E \cup N$, where E is an F_σ and N a null set (XVII.4.4.5.(iv)). Then $f(E)$ is an F_σ since f is continuous (XIII.11.7.11.), and $f(N)$ is a null set since f has Property (N); hence $f(A) = f(E) \cup f(N)$ is Lebesgue measurable. Conversely, if N is a null set and $f(N)$ is not a null set (i.e. $f(N)$ has positive Lebesgue outer measure), then $f(N)$ contains a

nonmeasurable set B (XVII.4.10.7.). If $A = f^{-1}(B) \cap N$, then $A \subseteq N$ and hence A is Lebesgue measurable, and $f(A) = B$ is nonmeasurable. 👍

XIX.5.1.5. Property (N) is complementary to bounded variation, even for continuous functions: the Cantor function is continuous and monotone, hence has bounded variation, but not Property (N), and there are continuous functions with Property (N) which do not have bounded variation, and are even not differentiable a.e. (XIX.5.8.2.).

There is an interesting variation on Property (N), due to Banach:

XIX.5.1.6. Definition. Let I be an interval, and $f : I \to \mathbb{R}$. Then f has *Property (S)* (or *Banach's Property (S)*) on I if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever A is a Lebesgue measurable subset of I with $\lambda(A) < \delta$, we have $\lambda^*(f(A)) < \epsilon$.

XIX.5.1.7. A function with Property (S) obviously has Property (N); hence, if f is continuous with Property (S), it sends measurable sets to measurable sets, and λ^* can be replaced by λ in the definition. There are continuous functions with Property (N) which do not have Property (S) (XIX.5.8.3.).

In fact, we have:

XIX.5.1.8. Theorem. [**?**] Let I be an interval, and $f : I \to \mathbb{R}$ a continuous function. Then f has property (S) if and only if it has Property (N) and N_f () is finite a.e.

Combining this with XIX.4.3.1., we obtain:

XIX.5.1.9. Corollary. Let I be an interval, and $f : I \to \mathbb{R}$ a continuous function of bounded variation. Then f has property (S) if and only if it has Property (N).

XIX.5.1.10. There are continuous functions with Property (S) which do not have bounded variation, and are even not differentiable a.e. (XIX.5.8.2.).

XIX.5.1.11. Banach said a function f satisfies Property (T_1) if N_f is finite a.e., and Property (T_2) if N_f is countable a.e. (i.e. f takes almost all values only countably many times). We will avoid this terminology since it could be confused with separation axioms of topology.

The next goal is to show that differentiable functions have Property (N). The next lemma is crucial.

XIX.5.1.12. LEMMA. Let A be a Lebesgue measurable subset of $\mathbb{R}$, and $f : A \to \mathbb{R}$ a function. Suppose there is a constant M such that, for every $x \in A$, there is a $\delta > 0$ such that $|f(y) - f(x)| \leq M(y - x)$ for all $y \in A \cap (x, x + \delta)$. Then $\lambda^*(f(A)) \leq M\lambda(A)$.

PROOF: We first make a reduction. For each n, let

$$B_n = \{x \in A : |f(y) - f(x)| \leq M(y - x) \text{ for all } y \in A \cap (x, x + 1/n)\} .$$

If $x \in A \cap \bar{B}_n$, there is a monotone sequence (x_k) in B_n with $x_k \to x$, so for sufficiently large k we have $|f(x) - f(x_k)| \leq M|x - x_k|$. Thus $f(x) = \lim_{k\to\infty} f(x_k)$ and, if $y \in A \cap (x, x + 1/n)$, then for sufficiently large k we have $y \in (x_k, x_k + 1/n)$, so $|f(y) - f(x_k)| \leq M(y - x_k)$. Taking $k \to \infty$, we conclude $x \in B_n$. Thus $B_n = A \cap \bar{B}_n$, and in particular B_n is measurable. We have that $A = \cup_n B_n$, so $\lambda(A) = \sup_n \lambda(B_n)$ and $\lambda^*(f(A)) = \sup_n \lambda^*(f(B_n))$. Thus it suffices to show $\lambda^*(f(B_n)) \leq M\lambda(B_n)$ for each n. Fix n; write B_n as a disjoint union of measurable sets A_{nm} of diameter $< 1/n$. Then $\lambda(B_n) = \sum_m \lambda(A_{nm})$ and $\lambda^*(f(B_n)) \leq \sum_m \lambda^*(f(A_{nm}))$. Thus it suffices to show that $\lambda^*(f(A_{nm})) \leq M\lambda(A_{nm})$ for all n, m. But A_{nm} has the property that if $x, y \in A_{nm}$, $x < y$, then $|f(y) - f(x)| \leq M(y - x)$. Replacing A by A_{nm}, we may assume A has the property that $|f(y) - f(x)| \leq M|y - x|$ for all $x, y \in A$. We may also assume $\lambda(A) < \infty$.

Let $\epsilon > 0$. Fix a sequence (I_n) of open intervals covering A with $\sum_n \ell(I_n) < \lambda(A) + \epsilon$. If $x, y \in A \cap I_n$, then $|f(y) - f(x)| \leq M|y - x|$, so $f(A \cap I_n)$ is contained in an interval J_n of length $\leq M\ell(I_n)$. We then have

$$\lambda^*(f(A)) \leq \sum_{n=1}^{\infty} \lambda^*(f(A \cap I_n)) \leq \sum_{n=1}^{\infty} \ell(J_n) \leq M \sum_{n=1}^{\infty} \ell(I_n) \leq M(\lambda(A) + \epsilon) .$$

This is true for all $\epsilon > 0$, and the result follows. 👍

XIX.5.1.13. COROLLARY. Let A be a Lebesgue measurable subset of $\mathbb{R}$, and $f : A \to \mathbb{R}$ a function. Suppose there is a constant M such that, for every $x \in A$, f is differentiable at x and $|f'(x)| \leq M$. Then $\lambda^*(f(A)) \leq M\lambda(A)$.

PROOF: Let $\epsilon > 0$. For each $x \in A$, since

$$\left| \lim_{y\to x} \frac{f(y) - f(x)}{y - x} \right| \leq M$$

there is a $\delta > 0$ such that

$$\left| \frac{f(y) - f(x)}{y - x} \right| \leq M + \epsilon$$

whenever $0 < |y - x| < \delta$. Thus $\lambda^*(f(A)) \leq (M + \epsilon)\lambda(A)$. Since $\epsilon > 0$ is arbitrary, the result follows. 👍

As a special case, we obtain the general one-dimensional version of Sard's Theorem (), due to LUZIN:

XIX.5.1.14. COROLLARY. Let I be an interval, and $f : I \to \mathbb{R}$ a function. Set

$$A = \{x \in I : f'(x) \text{ exists and equals } 0\} .$$

Then $f(A)$ is a null set.

PROOF: A is Lebesgue measurable by (). We may write $A = \cup_n A_n$, where $\lambda(A_n) < \infty$. Then $\lambda^*(f(A_n)) = 0$ for all n, and hence $\lambda^*(f(A)) \leq \sum_n \lambda^*(f(A_n)) = 0$. 👍

We can now obtain a sharpening of XIX.5.1.13.:

XIX.5.1.15. THEOREM. Let f be a function on $[a, b]$, and $A \subseteq [a, b]$ a Lebesgue measurable set. If f is differentiable at every point of A, then $f(A)$ is Lebesgue measurable and

$$\lambda(f(A)) \leq \int_A |f'| \, d\lambda .$$

PROOF: A is a union of a null set N and a countable collection (K_n) of compact sets (). Then $f(N)$ is a null set by XIX.5.1.14., and $f(K_n)$ is compact for each n since f is continuous on A. Thus $f(A) = f(N) \cup (\cup_n f(K_n))$ is Lebesgue measurable.
Let $\epsilon > 0$. For each $n \in \mathbb{N}$ let

$$A_n = \{x \in A : (n-1)\epsilon \leq |f'(x)| < n\epsilon\} .$$

Then A is the disjoint union of the A_n. For each n, A_n is measurable since $|f'|$ is measurable, and

$$\lambda^*(f(A_n)) \leq n\epsilon\lambda(A_n) \leq \int_{A_n} [|f'| + \epsilon] \, d\lambda = \int_{A_n} |f'| \, d\lambda + \epsilon\lambda(A_n)$$

by XIX.5.1.13.. Thus

$$\lambda(f(A)) \leq \sum_{n=1}^{\infty} \lambda^*(f(A_n)) \leq \sum_{n=1}^{\infty} \left[\int_{A_n} |f'| \, d\lambda \right] + \epsilon \sum_{n=1}^{\infty} \lambda(A_n) = \int_A |f'| \, d\lambda + \epsilon\lambda(A) .$$

Since $\epsilon > 0$ is arbitrary, the result follows. 👍

We then obtain:

XIX.5.1.16. THEOREM. Let B be a Lebesgue measurable set, and $f : B \to \mathbb{R}$ a function which is differentiable on B except possibly at countably many points of B. If A is a subset of B with $\lambda(A) = 0$, then $\lambda(f(A)) = 0$.

PROOF: Let C be a countable subset of B such that f is differentiable on $B \setminus C$. For each n let

$$D_n = \{x \in B \setminus C : |f'(x)| \leq n\} .$$

Each D_n is measurable. If A is a measurable subset of B with $\lambda(A) = 0$, then $f(A)$ is the union of $f(A \cap C)$ and $f(A \cap D_n)$ for all n. $f(A \cap C)$ is countable, hence has measure 0; and by XIX.5.1.13. we have $\lambda^*(f(A \cap D_n)) \leq n\lambda(A \cap D_n) = 0$. Thus

$$\lambda^*(f(A)) \leq \lambda^*(A \cap C) + \sum_{n=1}^{\infty} \lambda(A \cap D_n) = 0 .$$



XIX.5.1.17. COROLLARY. Let I be an interval, and $f : I \to \mathbb{R}$ a function. If f is differentiable on I except possibly at countably many points of I, then f has Property (N) on I.

Property (N) is relevant in extending consequences of the Mean Value Theorem:

XIX.5.1.18. THEOREM. Let I be an interval, and $f : I \to \mathbb{R}$ a continuous function with Property (N) on I. If f is differentiable a.e. and $f' \geq 0$ a.e. on I, then f is nondecreasing on I (and hence $f' \geq 0$ everywhere it is defined). Similarly, if $f' \leq 0$ a.e. on I, then f is nonincreasing on I. If $f' = 0$ a.e. on I, then f is constant on I.

PROOF: If $f' > 0$ a.e., the proof is identical to the proof of V.8.4.1.(i), since if f is differentiable except on a null set A, then $f(A)$ is also a null set since f has Property (N) and thus contains no interval (cf. V.8.4.4.). In the general case, we proceed as follows. Set

$$A = \{x \in I : f \text{ is not differentiable at } x\}$$

$$B = \{x \in I : f'(x) < 0\}$$

$$C = \{x \in I : f'(x) = 0\}$$

Then A and B are null sets by assumption, so $f(A)$ and $f(B)$ are null sets since f has Property (N). And $f(C)$ is a null set by XIX.5.1.14.. Thus $f(A \cup B \cup C)$ is a null set, and contains no interval.

Suppose there are $a, b \in I$ with $a < b$ and $f(a) > f(b)$. Then there is an $r \notin f(A \cup B \cup C)$ with $f(a) > r > f(b)$. There is an $x \in [a, b]$ with $f(x) = r$ by the Intermediate Value Theorem. Set

$$c = \sup\{x \in [a, b] : f(x) = r\} .$$

Then $f(c) = r$ by continuity, and $c < b$. Then $c \notin A \cup B \cup C$ since $r = f(c) \notin f(A \cup B \cup C)$, so f is differentiable at c and $f'(c) > 0$. We have $c < b$ since $f(b) < r$, and if $c < x \leq b$ we have $f(x) < f(c) = r$, so $\frac{f(x)-f(c)}{x-c} < 0$. Thus

$$f'(c) = \lim_{x \to c+} \frac{f(x) - f(c)}{x - c} \leq 0$$

contradicting that $f'(c) > 0$. 👍

Combining this result with XIX.5.1.17., we obtain:

XIX.5.1.19. Corollary. Let I be an interval, and $f : I \to \mathbb{R}$ a continuous function, which is differentiable on I except possibly at countably many points. If $f' \geq 0$ a.e. on I, then f is nondecreasing on I (and hence $f' \geq 0$ everywhere it is defined). Similarly, if $f' \leq 0$ a.e. on I, then f is nonincreasing on I. If $f' = 0$ a.e. on I, then f is constant on I.

Here is an interesting consequence of XIX.5.1.14.:

XIX.5.1.20. Corollary. Let $[a, b]$ be a closed bounded interval, and $f : [a, b] \to \mathbb{R}$ a continuous function. If f is differentiable on $[a, b]$ except possibly at countably many points, then N_f is finite a.e. on $[a, b]$, so f has Property (S) on $[a, b]$.

Proof: Let C be a countable set such that f is differentiable on $[a, b] \setminus C$. Let

$$A = \{x \in [a, b] : f'(x) = 0\} .$$

By XIX.5.1.14., $f(A)$ has measure 0. Since C is countable, $f(A \cup C)$ has measure 0.

Suppose $N_f(y) = \infty$. Then $f^{-1}(\{y\})$ is an infinite subset of $[a, b]$, hence has an accumulation point x. If $f'(x)$ exists, it must be zero; so $x \in A \cup C$, and hence $y \in f(A \cup C)$. So if $B = \{y : N_f(y) = \infty\}$, then $B \subseteq f(A \cup C)$, $\lambda(B) = 0$. The last statement follows from XIX.5.1.8.. 👍

XIX.5.2. Absolutely Continuous Functions

Our basic definition is a refinement of uniform continuity:

XIX.5.2.1. Definition. Let $[a, b]$ be a closed bounded interval, and $f : [a, b] \to \mathbb{R}$ (or $\mathbb{C}$) a function. Then f is *absolutely continuous* on $[a, b]$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are finitely many nonoverlapping intervals in $[a, b]$ (nonoverlapping means distinct intervals may have endpoints but not interior points in common) with total length $< \delta$, i.e.

$$\sum_{k=1}^{n} (b_k - a_k) < \delta$$

we have

$$\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \epsilon \ .$$

Note that the definition of uniform continuity is the same, but with only one subinterval allowed.

XIX.5.2.2. Although it will not be necessary, we may assume the intervals $[a_1, b_1], \ldots, [a_n, b_n]$ are in increasing order. The nonoverlapping condition then means

$$a \leq a_1 < b_1 \leq a_2 < b_2 \leq \cdots < b_n \leq b \ .$$

There is also no harm, and some technical advantage, in allowing some intervals to be degenerate, i.e. $a_k = b_k$, since such degenerate intervals make no contribution to the sums $\sum_{k=1}^{n}(b_k - a_k)$ and $\sum_{k=1}^{n} |f(b_k) - f(a_k)|$. We may also allow infinite collections of subintervals:

XIX.5.2.3. Proposition. Let $[a, b]$ be a closed bounded interval, and $f : [a, b] \to \mathbb{R}$ a function. Then f is absolutely continuous on $[a, b]$ if and only if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $\{[a_k, b_k] : k \in \mathbb{N}\}$ is a sequence of pairwise nonoverlapping subintervals (not necessarily nondegenerate) of $[a, b]$ with

$$\sum_{k=1}^{\infty}(b_k - a_k) < \delta$$

we have

$$\sum_{k=1}^{\infty} |f(b_k) - f(a_k)| < \epsilon \ .$$

Proof: If f satisfies this condition, it is clearly absolutely continuous (take all but finitely many of the intervals to be degenerate). Conversely, suppose f is absolutely continuous. Let $\epsilon > 0$, and choose $\delta > 0$ corresponding to $\frac{\epsilon}{2}$ in the definition of absolute continuity. If $\{[a_k, b_k] : k \in \mathbb{N}\}$ is a sequence of pairwise nonoverlapping subintervals (not necessarily nondegenerate) of $[a, b]$ with

$$\sum_{k=1}^{\infty}(b_k - a_k) < \delta$$

we have, for each n,

$$\sum_{k=1}^{n}(b_k - a_k) < \delta$$

and therefore

$$\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \frac{\epsilon}{2} \ .$$

Thus

$$\sum_{k=1}^{\infty} |f(b_k) - f(a_k)| \leq \frac{\epsilon}{2} < \epsilon \ .$$

XIX.5.2.4. EXAMPLES. (i) If f is Lipschitz on $[a,b]$, i.e. there is a constant $M>0$ such that

$$|f(y)-f(x)| \le M(y-x)$$

whenever $x,y \in [a,b]$, $x<y$, let $\epsilon>0$, and set $\delta = \frac{\epsilon}{M}$. If $[a_1,b_1],\dots,[a_n,b_n]$ are finitely many nonoverlapping intervals in $[a,b]$ with $\sum_{k=1}^n (b_k-a_k)<\delta$, then

$$\sum_{k=1}^n |f(b_k)-f(a_k)| \le \sum_{k=1}^n M(b_k-a_k) = M\sum_{k=1}^n (b_k-a_k) < M\delta = \epsilon$$

so f is absolutely continuous on $[a,b]$. (Conversely, we will see that every absolutely continuous function is "almost Lipschitz.")

(ii) Let f be the Cantor function (). The standard construction of the Cantor set K is as $K=\cap_n K_n$, where K_n is a disjoint union of 2^n closed intervals of length 3^{-n}. If $[a_1,b_1],\dots,[a_m,b_m]$ $(m=2^n)$ are the intervals in K_n, we have that

$$\sum_{k=1}^m (b_k-a_k) = \left(\frac{2}{3}\right)^n$$

but

$$\sum_{k=1}^m |f(b_k)-f(a_k)| = 1$$

since f is constant on the subintervals not included in K_n. Since $\left(\frac{2}{3}\right)^n \to 0$, f is not absolutely continuous on $[0,1]$. Thus a continuous function, even one of bounded variation and in particular differentiable a.e., is not necessarily absolutely continuous. There are even functions which are differentiable everywhere on an interval $[a,b]$ which are not absolutely continuous; cf. ().

Our main result, which will be proved in stages, is the following theorem giving various characterizations of absolute continuity, which shows that absolutely continuous functions are precisely the functions which can be recovered from their derivatives by integration.

XIX.5.2.5. THEOREM. Let $[a,b]$ be a closed bounded interval, and $F:[a,b]\to\mathbb{R}$ (or $\mathbb{C}$) a function. The following are equivalent:

(i) F is absolutely continuous on $[a,b]$.

(ii) F is continuous, has bounded variation, and property (N) on $[a,b]$.

(iii) F is continuous, has bounded variation, and $V_{[a,b]}(F) = \int_a^b |F'|\,d\lambda$.

(iv) There is a Lebesgue integrable function $f:[a,b]\to\mathbb{R}$ such that F is the indefinite integral of f on $[a,b]$, i.e.

$$F(x) = F(a) + \int_a^x f\,d\lambda$$

for all $x\in[a,b]$.

(v) F is differentiable a.e. on $[a,b]$, F' is integrable on $[a,b]$, and F is the indefinite integral of F' on $[a,b]$, i.e.

$$F(x) = F(a) + \int_a^x F'\, d\lambda$$

for all $x \in [a,b]$.

XIX.5.2.6. Corollary. Let μ be a finite continuous Borel measure on $[a,b]$, with CDF F. Then $\mu \ll \lambda$ if and only if F is absolutely continuous on $[a,b]$. If $\mu \ll \lambda$, then μ has a density function F', i.e. $\mu(B) = \int_B F'\, d\lambda$ for any Borel set $B \subseteq [a,b]$.

Proof: Since μ is continuous, F is a continuous function. Since it is nondecreasing, it has bounded variation. So F satisfies (ii) if and only if it has Property (N). Apply XIX.5.1.3.. If $\mu \ll \lambda$, then F is absolutely continuous, hence the last statement follows from (v). 👍

This corollary motivates the term *absolutely continuous*. The last statement is the Radon-Nikodým Theorem for Lebesgue measure on $(\mathbb{R}, \mathcal{B})$. An extension to signed and complex measures will be given in XIX.5.3.7..

We obtain a series of results leading to the proof of Theorem XIX.5.2.5.. Our exposition roughly follows [vRS82].

XIX.5.2.7. Proposition. A (finite) linear combination of absolutely continuous functions is absolutely continuous.

Proof: Since a scalar multiple of an absolutely continuous function is obviously absolutely continuous, it suffices to show that if f and g are absolutely continuous on $[a,b]$, then so is $f+g$. Let $\epsilon > 0$. Choose δ_1 so that whenever $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta_1$, then $\sum_{k=1}^n |f(b_k) - f(a_k)| < \frac{\epsilon}{2}$, and δ_2 so that if $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta_2$, then $\sum_{k=1}^n |g(b_k) - g(a_k)| < \frac{\epsilon}{2}$. Set $\delta = \min(\delta_1, \delta_2)$. If $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, then

$$\sum_{k=1}^n |f(b_k) + g(b_k) - f(a_k) - g(a_k)| \le \sum_{k=1}^n [|f(b_k) - f(a_k)| + |g(b_k) - g(a_k)|] < \epsilon\ .$$

👍

XIX.5.2.8. Lemma. An absolutely continuous function on $[a,b]$ has bounded variation on $[a,b]$.

Proof: Let f be absolutely continuous on $[a,b]$. Choose $\delta > 0$ such that if $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, then $\sum_{k=1}^n |f(b_k) - f(a_k)| < 1$. Let $N > \frac{b-a}{\delta}$, and partition $[a,b]$ into N subintervals $[x_{k-1}, x_k]$ $(1 \le k \le N)$ of equal length $\frac{b-a}{N} < \delta$. Fix k. If $\mathcal{P}$ is any

partition of $[x_{k-1}, x_k]$, then by choice of δ we have $V(f, \mathcal{P}) < 1$. Thus $V_{[x_{k-1},x_k]}(f) \le 1$. This is true for all k, so

$$V_{[a,b]}(f) = \sum_{k=1}^{N} V_{[x_{k-1},x_k]}(f) \le N \ .$$

XIX.5.2.9. LEMMA. An absolutely continuous function on $[a, b]$ has property (N).

PROOF: Let f be absolutely continuous on $[a, b]$, $A \subseteq [a, b]$ with $\lambda(A) = 0$, and $\epsilon > 0$. Choose $\delta > 0$ such that if $\{[a_k, b_k] : k \in \mathbb{N}\}$ is a sequence of nonoverlapping subintervals of $[a, b]$ with $\sum_{k=1}^{\infty}(b_k - a_k) < \delta$, then $\sum_{k=1}^{\infty} |f(b_k) - f(a_k)| < \epsilon$ (XIX.5.2.3.). Let U be an open set containing A with $\lambda(U) < \delta$ (). Then U is the disjoint union of a sequence (I_k) of open intervals. Set $\tilde{I}_k = \bar{I}_k \cap [a, b]$ (discard any $\tilde{I}_k$ that are empty). Since each $\tilde{I}_k$ is a closed bounded subinterval of $[a, b]$, f attains a maximum and a minimum on $\tilde{I}_k$ at points c_k and d_k respectively. Let $J_k = [a_k, b_k]$ be the closed interval with endpoints c_k and d_k. Then the J_k are nonoverlapping closed subintervals of $[a, b]$, and

$$\sum_{k=1}^{\infty}(b_k - a_k) \le \sum_{k=1}^{\infty} \ell(\tilde{I}_k) \le \sum_{k=1}^{\infty} \ell(I_k) = \lambda(U) < \delta \ .$$

We have $f(\tilde{I}_k) = f(J_k) = [f(d_k), f(c_k)]$, so $\lambda(f(\tilde{I}_k)) = |f(b_k) - f(a_k)|$ for each k. If $B = \cup_{k=1}^{\infty} \tilde{I}_k$, then $A \subseteq B$, so $f(A) \subseteq f(B) = \cup_{k=1}^{\infty} f(\tilde{I}_k)$ and

$$\lambda^*(f(A)) \le \lambda(f(B)) \le \sum_{k=1}^{\infty} \lambda(f(\tilde{I}_k)) = \sum_{k=1}^{\infty} |f(b_k) - f(a_k)| < \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $\lambda^*(f(A)) = 0$.

XIX.5.2.10. LEMMA. Let f be an integrable function on $[a, b]$, and $F(x) = \int_a^x f \, d\lambda$ for $x \in [a, b]$. Then F has bounded variation on $[a, b]$ and $V_{[a,b]}(F) \le \int_a^b |f| \, d\lambda$.

PROOF: If $a \le c < d \le b$, we have $F(d) - F(c) = \int_c^d f \, d\lambda$. Let $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$ be a partition of $[a, b]$. Then

$$V(F, \mathcal{P}) = \sum_{k=1}^{n} |F(x_k) - F(x_{k-1})| = \sum_{k=1}^{n} \left| \int_{x_{k-1}}^{x_k} f \, d\lambda \right| \le \sum_{k=1}^{n} \left[\int_{x_{k-1}}^{x_k} |f| \, d\lambda \right] = \int_a^b |f| \, d\lambda \ .$$

Take the supremum over all $\mathcal{P}$.

XIX.5.2.11. LEMMA. Let f be an integrable function on $[a,b]$, and $F(x) = \int_a^x f\,d\lambda$ for $x \in [a,b]$. Then F is absolutely continuous on $[a,b]$.

PROOF: If $a \le c < d \le b$, we have $F(d) - F(c) = \int_c^d f\,d\lambda$. Let $\epsilon > 0$. Let g be a bounded measurable function (e.g. a simple function) on $[a,b]$ with

$$\int_a^b |g - f|\,d\lambda < \frac{\epsilon}{2}\ .$$

Suppose $|g| \le M$ on $[a,b]$. Set $\delta = \frac{\epsilon}{2M}$. If $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, set $B = \cup_{k=1}^n [a_k,b_k]$. Then $\lambda(B) < \delta$. We have

$$\sum_{k=1}^n |F(b_k) - F(a_k)| = \sum_{k=1}^n \left| \int_{a_k}^{b_k} f\,d\lambda \right| \le \sum_{k=1}^n \left[\int_{a_k}^{b_k} |f|\,d\lambda \right]$$

$$= \int_B |f|\,d\lambda \le \int_B |g|\,d\lambda + \int_B |f-g|\,d\lambda < M\lambda(B) + \frac{\epsilon}{2} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\ .$$



XIX.5.2.12. LEMMA. Let F be a continuous function on $[a,b]$ with property (N), such that F is differentiable a.e. on $[a,b]$ and F' is integrable on $[a,b]$ (satisfied if F has bounded variation on $[a,b]$). Then F is absolutely continuous on $[a,b]$, and $V_{[a,b]}(F) = \int_a^b |F'|\,d\lambda$.

PROOF: Set

$$G(x) = \int_a^x |F'|\,d\lambda$$

for $x \in [a,b]$. Then G is absolutely continuous on $[a,b]$ by XIX.5.2.11.. Let $N \subseteq [a,b]$ be a null set such that F is differentiable everywhere on $[a,b] \setminus N$. Then, since F is continuous and has property (N), for $a \le c < d \le b$,

$$|F(d) - F(c)| \le \lambda(F([c,d])) \le \lambda^*(F([c,d] \setminus N)) + \lambda(F(N)) = \lambda^*(F([c,d] \setminus N)) \le \int_{[c,d]\setminus N} |F'|\,d\lambda = \int_c^d |F'|\,d\lambda$$

by XIX.5.1.15..

Let $\epsilon > 0$, and choose $\delta > 0$ so that whenever $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, we have

$$\sum_{k=1}^n |G(b_k) - G(a_k)| = \sum_{k=1}^n \left[\int_{a_k}^{b_k} |F'|\,d\lambda \right] < \epsilon\ .$$

Then

$$\sum_{k=1}^n |F(b_k) - F(a_k)| \le \sum_{k=1}^n \left[\int_{a_k}^{b_k} |F'|\,d\lambda \right] < \epsilon\ .$$

Thus F is absolutely continuous.
If $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$ is a partition of $[a, b]$, we have

$$V(F, \mathcal{P}) = \sum_{k=1}^{n} |F(x_k) - F(x_{k-1})| \leq \sum_{k=1}^{n} \left[\int_{x_{k-1}}^{x_k} |F'| \, d\lambda \right] = \int_a^b |F'| \, d\lambda \ .$$

Taking the supremum over all $\mathcal{P}$, we obtain $V_{[a,b]}(F) \leq \int_a^b |F'| \, d\lambda$. The opposite inequality follows from XIX.4.2.5.. 👍

XIX.5.2.13. PROPOSITION. Let f be an integrable function on $[a, b]$. Suppose $\int_a^x f \, d\lambda = 0$ for all $x \in [a, b]$. Then $f = 0$ a.e. on $[a, b]$.

PROOF: If $[c, d] \subseteq [a, b]$, then

$$\int_c^d f \, d\lambda = \int_a^d f \, d\lambda - \int_a^c f \, d\lambda = 0 - 0 = 0 \ .$$

Thus if I is any subinterval of $[a, b]$, $\int_I f \, d\lambda = 0$. If U is a relatively open set in $[a, b]$, then U is a disjoint union of relatively open intervals I_k, so $\int_U f \, d\lambda = \sum_k \int_{I_k} f \, d\lambda = 0$. If $B = \cap_n U_n$ is a G_δ in $[a, b]$, then

$$0 = \int_{U_n} f \, d\lambda \to \int_B f \, d\lambda$$

by the DCT, so $\int_B f \, d\lambda = 0$. If A is any Lebesgue measurable subset of $[a, b]$, there is a G_δ $B \subseteq [a, b]$ with $A \subseteq B$ and $\lambda(B \setminus A) = 0$, so

$$\int_A f \, d\lambda = \int_B f \, d\lambda = 0 \ .$$

The result now follows from (). 👍

XIX.5.2.14. COROLLARY. Let f and g be integrable functions on $[a, b]$. If

$$\int_a^x f \, d\lambda = \int_a^x g \, d\lambda$$

for all $x \in [a, b]$, then $f = g$ a.e. on $[a, b]$.

PROOF: Apply XIX.5.2.13. to $f - g$. 👍

The Fundamental Theorem of Calculus 1

We can now prove the following version of the Fundamental Theorem of Calculus, which is a strong version of XIX.5.2.5.[(iv) ⇒ (v)]:

XIX.5.2.15. THEOREM. [FUNDAMENTAL THEOREM OF CALCULUS, LEBESGUE VERSION 1A] Let $[a,b]$ be a closed bounded interval, and $f:[a,b]\to\mathbb{R}$ a Lebesgue integrable function. Let

$$F(x)=\int_a^x f\,d\lambda$$

for all $x\in[a,b]$. Then F is differentiable a.e. and $F'=f$ a.e. on $[a,b]$.

See XIX.5.2.24. for a variation.

PROOF: Suppose first that f is nonnegative. Then F is nondecreasing. By XIX.5.2.11., F is absolutely continuous, and thus has bounded variation and Property (N) by XIX.5.2.8. and XIX.5.2.9.. By XIX.5.2.12. we have that

$$F(d)-F(c)=V_{[c,d]}(F)=\int_c^d |F'|\,d\lambda=\int_c^d F'\,d\lambda$$

for any $[c,d]\subseteq[a,b]$ since $F'\geq 0$ whenever it exists, and in particular

$$F(x)=\int_a^x F'\,d\lambda=\int_a^x f\,d\lambda$$

for any $x\in[a,b]$. Thus $F'=f$ a.e. by XIX.5.2.14..

For a general real f, write $f=f_+-f_-$, and set $F_+(x)=\int_a^x f_+\,d\lambda$ and $F_-(x)=\int_a^x f_-\,d\lambda$; then $F=F_+-F_-$, so $F'=F'_+-F'_-$ (defined a.e.) We have $F'_+=f_+$ a.e. and $F'_-=f_-$ a.e., so $F'=f$ a.e. For complex f, consider real and imaginary parts. 👍

The Proof of the Main Theorem

XIX.5.2.16. We can now give the proof of Theorem XIX.5.2.5.. (i) ⇒ (ii) by XIX.5.2.8. and XIX.5.2.9., and (ii) ⇒ (i) and (ii) ⇒ (iii) by XIX.5.2.12.. (iv) ⇒ (i) by XIX.5.2.11.. (iv) ⇒ (v) by XIX.5.2.15. (strictly speaking we do not need this implication, although we need XIX.5.2.15. for (i) ⇒ (v) below), and (v) ⇒ (iv) is trivial.

It remains to show (iii) ⇒ (i) ⇒ (v). For (i) ⇒ (v), suppose F is absolutely continuous on $[a,b]$. Then F has bounded variation on $[a,b]$ and F' is integrable on $[a,b]$. Set

$$G(x)=\int_a^x F'\,d\lambda$$

for $x\in[a,b]$. Then G is absolutely continuous on $[a,b]$, and $G'=F'$ a.e. by XIX.5.2.15.. So $F-G$ is absolutely continuous on $[a,b]$ by XIX.5.2.7., so has property (N) by XIX.5.2.9.. We have $(F-G)'=0$ a.e. on $[a,b]$, so $F-G$ is constant on $[a,b]$ by XIX.5.1.18.. Since $G(a)=0$, we have $F-G=F(a)$, so

$$F(x)=F(a)+G(x)=F(a)+\int_a^x F'\,d\lambda$$

for all $x\in[a,b]$.

For (iii) $\Rightarrow$ (i), suppose F is continuous, bounded variation, and $V_{[a,b]}(F) = \int_a^b |F'|\,d\lambda$. For any $x \in [a,b]$, we have $\int_a^x |F'|\,d\lambda \le V_{[a,x]}(F)$ and $\int_x^b |F'|\,d\lambda \le V_{[x,b]}(F)$, and

$$V_{[a,b]}(F) = \int_a^b |F'|\,d\lambda = \int_a^x |F'|\,d\lambda + \int_x^b |F'|\,d\lambda \le V_{[a,x]}(F) + V_{[x,b]}(F) = V_{[a,b]}(F)$$

and it follows that $V_{[a,x]}(F) = \int_a^x |F'|\,d\lambda$ (and $V_{[x,b]}(F) = \int_x^b |F'|\,d\lambda$), i.e.

$$(TF)(x) = \int_a^x |F'|\,d\lambda$$

for all $x \in [a,b]$, where TF is the total variation function of F. Thus TF is absolutely continuous by XIX.5.2.11.. Let $\epsilon > 0$, and choose $\delta > 0$ so that if $\{[a_k, b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, then $\sum_{k=1}^n |TF(b_k) - TF(a_k)| < \epsilon$. Then, since

$$|F(b_k) - F(a_k)| \le V_{[a_k,b_k]}(F) = TF(b_k) - TF(a_k)$$

for all k, we have $\sum_{k=1}^n |F(b_k) - F(a_k)| < \epsilon$. Thus F is absolutely continuous.

This completes the proof of Theorem XIX.5.2.5.. 

An interesting corollary of the proof, combined with previous results, is:

XIX.5.2.17. Corollary. Let F be a continuous function on $[a,b]$. Then F is absolutely continuous if and only if it has bounded variation and its total variation function TF is absolutely continuous. If TF is absolutely continuous, then $(TF)' = |F'|$ a.e. on $[a,b]$.

Proof: The last part of the proof of (iii) $\Rightarrow$ (i) above shows that if TF is absolutely continuous, so is F. Conversely, if F is absolutely continuous, then F has bounded variation and Property (N), so $V_{[a,b]}(F) = \int_a^b |F'|\,d\lambda$ by XIX.5.2.12., and by the first half of the proof of (iii) $\Rightarrow$ (i) the function TF is absolutely continuous, and

$$(TF)(x) = \int_a^x |F'|\,d\lambda$$

for all $x \in [a,b]$. Thus $(TF)' = |F'|$ a.e. by XIX.5.2.15..

Lipschitz Functions

XIX.5.2.18. Theorem. Let $[a,b]$ be a closed bounded interval, and $F : [a,b] \to \mathbb{R}$ a function. Then F is Lipschitz on $[a,b]$ if and only if F is absolutely continuous, hence differentiable a.e. on $[a,b]$, and F' is bounded on $[a,b]$. Set

$$M = \sup\left\{\left|\frac{F(y) - F(x)}{y - x}\right| : x, y \in [a,b], x \ne y\right\}$$

$$M' = \sup\{|D^+F(x)|, |D_+F(x)|, |D^-F(x)|, |D_-F(x)| : x \in [a,b]\}$$

$$M'' = \sup\{|F'(x)| : x \in [a,b], F'(x) \text{ exists}\}\ .$$

If F is Lipschitz, then $M = M' = M''$ is finite and is the smallest Lipschitz constant for F on $[a, b]$.

PROOF: We have already seen (XIX.5.2.4.) that a Lipschitz function is absolutely continuous. Conversely, if F is absolutely continuous on $[a, b]$ and M'' is finite, then for $x, y \in [a, b]$, $x < y$, we have

$$|F(y) - F(x)| = \left| \int_x^y F' \, d\lambda \right| \leq \int_x^y |F'| \, d\lambda \leq M''(y - x)$$

so F is Lipschitz on $[a, b]$ and M'' is a Lipschitz constant for F. If F is Lipschitz, i.e. M is finite, then M is the smallest Lipschitz constant for F on $[a, b]$, so we have $M \leq M''$. But $M'' \leq M' \leq M$ since every Dini derivative (in particular, any derivative) of F is a limit of difference quotients, so $M = M' = M''$ (and in particular M'' is finite). 👍

Everywhere Differentiable Functions

If a function F is differentiable everywhere on $[a, b]$, we have a more precise result, which applies more generally if F fails to be differentiable at only countably many points:

XIX.5.2.19. THEOREM. Let F be continuous on $[a, b]$ and differentiable except possibly on a countable subset of $[a, b]$. Then the following are equivalent:

(i) F is absolutely continuous on $[a, b]$.

(ii) F has bounded variation on $[a, b]$.

(iii) F' is Lebesgue integrable on $[a, b]$.

PROOF: (i) $\Rightarrow$ (ii) by XIX.5.2.8., and (ii) $\Rightarrow$ (iii) by (). If F satisfies the hypotheses, then F has property (N) by XIX.5.1.17., so if F' is Lebesgue integrable on $[a, b]$, then F is absolutely continuous by XIX.5.2.12., giving (iii) $\Rightarrow$ (i). 👍

XIX.5.2.20. In XIX.5.2.19., "differentiable except possibly on a countable subset" cannot be weakened to "differentiable a.e." The Cantor function is a counterexample. The conditions can fail, even if F is differentiable everywhere on $[a, b]$ (IX.2.9.10.(i)). ✋

The Fundamental Theorem of Calculus Again

The equivalence of (i) and (v) of XIX.5.2.5. can be rephrased as the other version of the Fundamental Theorem of Calculus:

XIX.5.2.21. THEOREM. [FUNDAMENTAL THEOREM OF CALCULUS, LEBESGUE VERSION 2A] Let F be a function which is absolutely continuous on $[a, b]$, with derivative f (defined a.e.). Then f is Lebesgue integrable on $[a, b]$, and for all $x \in [a, b]$ we have

$$F(x) = F(a) + \int_a^x f \, d\lambda$$

and so, in particular,

$$\int_a^b f \, d\lambda = F(b) - F(a) \ .$$

Conversely, if F can be so represented, it is absolutely continuous.

Combining XIX.5.2.15. with XIX.5.2.19., we get another version of Version 2 for everywhere differentiable functions:

XIX.5.2.22. THEOREM. [FUNDAMENTAL THEOREM OF CALCULUS, LEBESGUE VERSION 2B] Let F be a function which is continuous on $[a, b]$ and differentiable except possibly on a countable subset of $[a, b]$, with derivative f (defined a.e.). If f is Lebesgue integrable on $[a, b]$, then for all $x \in [a, b]$ we have

$$F(x) = F(a) + \int_a^x f \, d\lambda$$

and so, in particular,

$$\int_a^b f \, d\lambda = F(b) - F(a) \ .$$

The same caution as in XIX.5.2.20. applies here.

XIX.5.2.23. Thus care must be taken in the statement that Lebesgue integration and differentiation are inverse processes: the derivative of an integral is the function started with, but only almost everywhere in general by XIX.5.2.15. (a better result cannot be expected in general since functions which agree almost everywhere have the same Lebesgue integral), but a function is the integral of its derivative if and only if it is absolutely continuous (XIX.5.2.21.).

Theorems XIX.5.2.21. and XIX.5.2.22. are true generalizations of the Fundamental Theorem of Calculus 2. But Theorem XIX.5.2.15. is not a true generalization of the Fundamental Theorem of Calculus 1. In fact, XIX.5.2.15. does not assert that $F'(x) = f(x)$ for any *specific* $x \in [a, b]$, only for "almost all" x's. But there is a true generalization of the FTC1, obtained by just replacing Riemann integrals with Lebesgue integrals in the statement and proof:

XIX.5.2.24. THEOREM. [FUNDAMENTAL THEOREM OF CALCULUS, LEBESGUE VERSION 1B] Let f be Lebesgue integrable on $[a,b]$, and

$$F(x) = \int_a^x f\,d\lambda$$

for $x \in [a,b]$. If f is continuous at $x_0 \in [a,b]$, then F is differentiable at x_0 and $F'(x_0) = f(x_0)$.

PROOF: We have, for $x \in [a,b]$, $x \neq x_0$,

$$F(x) - F(x_0) = \int_a^x f\,d\lambda - \int_a^{x_0} f\,d\lambda = \int_{x_0}^x f\,d\lambda$$

(interpreted as $-\int_x^{x_0} f\,d\lambda$ if $x < x_0$), so

$$\frac{F(x) - F(x_0)}{x - x_0} = \frac{1}{x - x_0}\int_{x_0}^x f\,d\lambda\ .$$

We also have

$$f(x_0) = \frac{1}{x - x_0}\int_{x_0}^x f(x_0)\,d\lambda$$

(since $f(x_0)$ is just a constant), and so

$$\frac{F(x) - F(x_0)}{x - x_0} - f(x_0) = \frac{1}{x - x_0}\int_{x_0}^x [f - f(x_0)]\,d\lambda\ .$$

Let $\epsilon > 0$. Since f is continuous at x_0, there is a $\delta > 0$ such that, whenever $x \in [a,b]$ and $|x - x_0| < \delta$, $|f(x) - f(x_0)| < \epsilon$. If $x \in [a,b]$ and $|x - x_0| < \delta$, then $|t - x_0| < \delta$ for all t between x_0 and x, and hence

$$\left|\int_{x_0}^x [f - f(x_0)]\,d\lambda\right| < \epsilon|x - x_0|$$

by (); so we have

$$\left|\frac{F(x) - F(x_0)}{x - x_0} - f(x_0)\right| = \frac{1}{|x - x_0|}\left|\int_{x_0}^x [f - f(x_0)]\,d\lambda\right| < \epsilon$$

and therefore

$$F'(x_0) = \lim_{x \to x_0}\frac{F(x) - F(x_0)}{x - x_0} = f(x_0)\ .$$



Characterizing Functions as Derivatives

Theorem XIX.5.2.5. has another interesting consequence: integrable functions can be characterized without any explicit reference to integration. The next result generalizes part of IX.2.9.4. (see XIX.5.2.31. for the other part)

XIX.5.2.25. Corollary. Let f be a Lebesgue measurable real-valued function on an interval $[a,b]$. Then f is Lebesgue integrable on $[a,b]$ if and only if there is an absolutely continuous function F on $[a,b]$ with $F' = f$ a.e. on $[a,b]$.

If the word "absolutely" is removed from the statement, the result is false (). In fact, it is remarkable that *every* Lebesgue measurable function is a.e. the derivative of a continuous function. This result is due to Luzin; our proof is adapted from [Sak64].

XIX.5.2.26. Theorem. Let f be a Lebesgue measurable real-valued function on an interval $[a,b]$. Then there is a continuous function F on $[a,b]$ which is differentiable a.e. and $F' = f$ a.e. In fact, if G is any continuous function on $[a,b]$, for any $\epsilon > 0$ there is a continuous function F on $[a,b]$ which is differentiable a.e., $F' = f$ a.e., and $|F(x) - G(x)| < \epsilon$ for all $x \in [a,b]$.

The proof is a little complicated technically; we begin with an observation and two lemmas.

XIX.5.2.27. Let $\epsilon > 0$. If G is continuous on $[a,b]$, there is a continuous function H on $[a,b]$ which is differentiable a.e. and $H' = 0$ a.e. which is uniformly within $\frac{\epsilon}{2}$ of G (XX.10.3.6.). If F_0 is a continuous function on $[a,b]$ which is differentiable a.e. with $F_0' = f$ a.e., with $|F_0(x)| < \frac{\epsilon}{2}$ for all $x \in [a,b]$, then $F = F_0 + H$ is the desired function. Thus it suffices to show that if $\epsilon > 0$, there is a continuous function F on $[a,b]$ which is differentiable a.e., with $F' = f$ a.e. and $|F(x)| < \epsilon$ for all $x \in [a,b]$.

If f is integrable, this is fairly easily accomplished with judicious use of the Cantor function (this is essentially an explicit version of XX.10.3.6., cf. Exercise XIX.5.8.6.):

XIX.5.2.28. Lemma. Let ϕ be an integrable function on an interval $[a,b]$. Then for any $\epsilon > 0$ there is a continuous function Φ on $[a,b]$ with the following properties:

(i) Φ is differentiable a.e. and $\Phi' = \phi$ a.e. on $[a,b]$.

(ii) $\Phi(a) = \Phi(b) = 0$.

(iii) $|\Phi(x)| < \epsilon$ for all $x \in [a,b]$.

Proof: For $x \in [a,b]$, set $\Psi(x) = \int_a^x \phi \, d\lambda$. Then Ψ is absolutely continuous on $[a,b]$, $\Psi' = \phi$ a.e. on $[a,b]$, and $\Psi(a) = 0$. Fix $\epsilon > 0$. Since Ψ is uniformly continuous on $[a,b]$, there is a $\delta > 0$ such that $|\Psi(x) - \Psi(y)| < \epsilon$ whenever $x, y \in [a,b]$ and $|x - y| < \delta$. Let $\mathcal{P} = \{a = x_0, \ldots, x_n = b\}$ be a partition of $[a,b]$ into subintervals of length $< \delta$. On $[x_{k-1}, x_k]$, let H_k be a copy of the Cantor Function scaled to $[x_{k-1}, x_k]$ with $H_k(x_{k-1}) = \Psi(x_{k-1})$ and $H_k(x_k) = \Psi_k(x_k)$. Define $\Phi(x) = \Psi(x) - H_k(x)$ if $x_{k-1} \le x \le x_k$. Then $\Phi(x_k) = 0$ for all k, and hence Φ is well defined and continuous; and $\Phi' = \Psi' = \phi$ a.e. since $H_k' = 0$ a.e. for each k. If $x \in [x_{k-1}, x_k]$, we have $|\Psi(x) - \Psi(x_{k-1})| < \epsilon$ and $|\Psi(x) - \Psi(x_k)| < \epsilon$; since $H_k(x)$ is between $H_k(x_{k-1}) = \Psi(x_{k-1})$ and $H_k(x_k) = \Psi(x_k)$, we also have $|\Phi(x)| = |\Psi(x) - H_k(x)| < \epsilon$. 👍

To handle the more difficult case where f is not integrable, we need the following technical lemma:

XIX.5.2.29. LEMMA. Let ϕ be an integrable function on an interval $[a,b]$, and C a closed subset of $[a,b]$. Then for any $\epsilon > 0$ there is a continuous function Φ on $[a,b]$ with the following properties:

(i) Φ is differentiable a.e. and $\Phi' = \phi$ a.e. on $[a,b] \setminus C$.

(ii) $\Phi(x) = \Phi'(x) = 0$ for all $x \in C$.

(iii) $|\Phi(x+h)| = |\Phi(x+h) - \Phi(x)| < \epsilon|h|$ whenever $x \in C$ and $x + h \in [a,b]$.

PROOF: Fix $\epsilon > 0$. The open set $(a,b) \setminus C$ is a disjoint union of open intervals (a_k, b_k). For each k, fix a set $\{x_{k,n} : n \in \mathbb{Z}\}$ with $x_{k,n} < x_{k,n+1}$ for all n, $\lim_{n\to-\infty} x_{k,n} = a_k$, and $\lim_{n\to+\infty} x_{k,n} = b_k$. For each k and n, let $\Phi_{k,n}$ be a continuous function on $[x_{k,n}, x_{k,n+1}]$ such that

(i) $\Phi_{k,n}$ is differentiable a.e. and $\Phi'_{k,n} = \phi$ a.e. on $[x_{k,n}, x_{k,n+1}]$.

(ii) $\Phi_{k,n}(x_{k,n}) = \Phi_{k,n}(x_{k,n+1}) = 0$.

(iii) $|\Phi_{k,n}(x)| < \min\left(\dfrac{\epsilon(x_{k,n} - a_k)}{k + |n|}, \dfrac{\epsilon(b_k - x_{k,n+1})}{k + |n|}\right)$ for all $x \in [x_{k,n}, x_{k,n+1}]$.

(Lemma XIX.5.2.28.). Define Φ on $[a,b]$ by $\Phi(a) = \Phi(b) = 0$, $\Phi(x) = 0$ for $x \in C$, and $\Phi(x) = \Phi_{k,n}(x)$ if $x \in [x_{k,n}, x_{k,n+1}]$. Then Φ is well defined and continuous, and it is routine to check it has the other required properties (Exercise XIX.5.8.5.). 👍

XIX.5.2.30. We now prove the theorem (in the reduced form $G = 0$). Fix $\epsilon > 0$. Inductively for $n \geq 0$ construct continuous functions Φ_n which are differentiable a.e. and closed subsets (C_n) of $[a,b]$ with $C_n \subseteq C_{n+1}$ for all n, such that, with $F_n = \sum_{k=1}^{n} \Phi_k$, the following conditions are satisfied for $n \geq 1$:

(i) F_n is differentiable at all points of C_n and $F'_n = f$ on C_n.

(ii) $\Phi_n(x) = \Phi'_n(x) = 0$ for all $x \in C_{n-1}$.

(iii) $|\Phi_n(x+h)| = |\Phi_n(x+h) - \Phi_n(x)| \leq \dfrac{\epsilon|h|}{2^n(b-a)}$ whenever $x \in C_{n-1}$ and $x + h \in [a,b]$.

(iv) $\lambda([a,b] \setminus C_n) < \frac{1}{n}$.

Begin with $C_0 = \emptyset$ and $\Phi_0 = 0$. Suppose Φ_k and C_k have been chosen for $k \leq n$. Since F'_n and f are defined a.e., for some sufficiently large M the measurable set

$$A_n = \{x \in [a,b] \setminus C_n : F'_n(x) \text{ and } f(x) \text{ exist and } |F'_n(x)|, |f(x)| \leq M\}$$

satisfies $\lambda([a,b] \setminus [C_n \cup A_n]) < \frac{1}{n+1}$.

Define ϕ_n on $[a,b]$ by $\phi_n = f - F'_n$ on A_n and $\phi_n = 0$ on $[a,b] \setminus A_n$. Then ϕ_n is bounded, hence integrable, so by Lemma XIX.5.2.29. there is a continuous function Φ_{n+1} on $[a,b]$ which is differentiable a.e., $\Phi'_{n+1}(x) = f(x) - F'_n(x)$ for almost all $x \in A_n$, $\Phi_{n+1}(x) = \Phi'_{n+1}(x) = 0$ for all $x \in C_n$, and

$$|\Phi_{n+1}(x+h)| < \frac{\epsilon|h|}{2^{n+1}(b-a)}$$

if $x \in C_n$ and $x + h \in [a, b]$.

By (), there is a closed subset B_n of A_n such that $\Phi'_{n+1}(x) = f(x) - F'_n(x)$ for all $x \in B_n$ and $\lambda([0,1] \setminus [C_n \cup B_n]) < \frac{1}{n+1}$. Set $C_{n+1} = C_n \cup B_n$. This completes the inductive construction, and properties (i)–(iv) obviously hold.

From (iii), we have $|\Phi_n(x)| < \frac{\epsilon}{2^n}$ for all $x \in [a, b]$; thus the series

$$\sum_{n=1}^{\infty} \Phi_n$$

converges uniformly to a continuous function F on $[a, b]$, and $|F(x)| < \epsilon$ for all $x \in [a, b]$. Set $C = \cup_{n=1}^{\infty} C_n$. Then $\lambda([a, b] \setminus C) = 0$, and if $x \in C$, then $x \in C_n$ for sufficiently large n; thus for any h with $x + h \in [a, b]$ we have

$$\left| \frac{F(x+h) - F(x)}{h} - \frac{F_n(x+h) - F_n(x)}{h} \right| = \left| \sum_{k=n+1}^{\infty} \frac{\Phi_k(x+h) - \Phi_k(x)}{h} \right|$$

$$\leq \sum_{k=n+1}^{\infty} \frac{|\Phi_k(x+h) - \Phi_k(x)|}{|h|} < \frac{\epsilon}{2^n(b-a)}$$

for sufficiently large n, and hence

$$\limsup_{h \to 0} \left| \frac{F(x+h) - F(x)}{h} - f(x) \right| \leq \frac{\epsilon}{2^n(b-a)}$$

for all sufficiently large n, i.e. $F'(x) = f(x)$.

This completes the proof of Theorem XIX.5.2.26.. 

XIX.5.2.31. What about uniqueness in XIX.5.2.25.? The F is clearly unique up to adding a constant among absolutely continuous functions. But from XIX.5.2.26. there are many continuous functions which are differentiable a.e. and whose derivative equals f a.e.; in fact such functions are uniformly dense in $C([a, b])$. Thus there is no reasonable uniqueness statement among continuous functions.

XIX.5.3. Signed and Complex Borel Measures on $\mathbb{R}$

We can extend the notions of cumulative distribution function, absolute continuity, and density function to signed and complex measures. The definition of cdf is straightforward:

XIX.5.3.1. Definition. Let μ be a finite signed or complex Borel measure on $\mathbb{R}$. For $x \in \mathbb{R}$, set $F_\mu(x) = \mu((-\infty, x])$. Then F_μ is the *cumulative distribution function (cdf)* for μ.

XIX.5.3.2. Proposition. The function F_μ has the following properties:

(i) F_μ has at most countably many discontinuities, each of which is a jump discontinuity.

(ii) F_μ is continuous from the right.

(iii) $\lim_{x\to-\infty} F_\mu(x) = 0$.

(iv) F_μ is bounded, and $\lim_{x\to+\infty} F_\mu(x) = \mu(\mathbb{R})$.

PROOF: Parts (ii), (iii), and (iv) follow immediately from XVII.9.1.11.. It also follows from XVII.9.1.11. and the Sequential Criterion that for any $x \in \mathbb{R}$,

$$\mu((-\infty, x)) = \lim_{y\to x-} \mu((-\infty, y]) = \lim_{y\to x-} F_\mu(y)$$

so $F_\mu(x-)$ and $F_\mu(x+) = F_\mu(x)$ exist for all $x \in \mathbb{R}$. Thus every discontinuity of F_μ is a jump discontinuity. Furthermore,

$$F_\mu(x+) - F_\mu(x-) = \mu((-\infty, x]) - \mu((-\infty, x)) = \mu(\{x\})$$

for all x. Since μ is finite, it can have only countably many atoms, and the sum of the measures of the atoms is finite. 👍

The direct proof of (i) above is elementary and does not use the Jordan Decomposition Theorem. But using this theorem we get an important result which gives (i) as an immediate corollary:

XIX.5.3.3. PROPOSITION. Let μ be a finite signed Borel measure on $\mathbb{R}$, and let (μ^+, μ^-) be the Jordan decomposition (XVII.9.1.3.) of μ. Then $F_\mu = F_{\mu^+} - F_{\mu^-}$. In particular, since F_{μ^+} and F_{μ^-} are bounded and nondecreasing, F_μ has bounded variation on $\mathbb{R}$. In particular, F_μ is differentiable a.e. and F'_μ is integrable on closed bounded subintervals of $\mathbb{R}$.

XIX.5.3.4. COROLLARY. Let μ be a finite signed Borel measure on $\mathbb{R}$, and let (μ^+, μ^-) be the Jordan decomposition of μ. If $|\mu|$ is the total variation measure of μ (), then $F_{|\mu|} = TF_\mu$, the total variation function of F_μ.

XIX.5.3.5. These results extend immediately to complex measures by considering real and imaginary parts: if μ is a complex Borel measure on $\mathbb{R}$, then F_μ has bounded variation and $F_{|\mu|} = TF_\mu$.

Absolute Continuity

It is slightly tricky to give the proper definition of absolute continuity for signed or complex measures, due to the distinction between null sets and sets of measure zero. The proper definition is:

XIX.5.3.6. DEFINITION. Let μ and ν be signed or complex measures on a measurable space $(X, \mathcal{A})$. Then ν is *absolutely continuous* with respect to μ, written $\nu \ll \mu$, if every μ-null set is a ν-null set, i.e. $|\nu| \ll |\mu|$ in the sense of ().

The next result is an immediate consequence of XIX.5.3.4., XIX.5.2.6., andXIX.5.2.17.:

XIX.5.3.7. PROPOSITION. Let μ be a finite signed or complex Borel measure on $\mathbb{R}$, then the following are equivalent:

(i) $\mu \ll \lambda$.

(ii) TF_μ is absolutely continuous on every closed bounded subinterval of $\mathbb{R}$.

(ii) F_μ is absolutely continuous on every closed bounded subinterval of $\mathbb{R}$.

Density Functions

XIX.5.3.8. DEFINITION. Let μ be a finite signed or complex Borel measure on $\mathbb{R}$. Then μ has a *density function* ϕ if

$$\mu(B) = \int_B \phi \, d\lambda$$

for every Borel set B in $\mathbb{R}$.

A density function, if it exists, is uniquely determined a.e., i.e. two density functions for a measure μ must be equal a.e.

XIX.5.3.9. THEOREM. Let μ be a finite signed or complex Borel measure on $\mathbb{R}$. The following are equivalent:

(i) μ has a density function.

(ii) F'_μ is a density function for μ.

(iii) $|\mu|$ has a density function.

(iv) $|F'_\mu|$ is a density function for $|\mu|$.

(v) $\mu \ll \lambda$.

PROOF: If μ has a density function ϕ, then for any closed bounded interval $[a, b]$,

$$F_\mu(x) = F(a) + \int_a^x \phi \, d\lambda$$

so F_μ is absolutely continuous on $[a, b]$, and hence $F'_\mu = \phi$ a.e. on $\mathbb{R}$. Thus (i) $\Rightarrow$ (ii). (ii) $\Rightarrow$ (i) is trivial, so (i) $\iff$ (ii). Similarly, (iii) $\iff$ (iv). By XIX.5.3.7., (v) is equivalent to $TF_\mu = F_{|\mu|}$ being absolutely continuous, which is equivalent to $|F'|$ being a density function for $|\mu|$ by XIX.5.2.6.; thus (v) is equivalent to (iii) and (iv). (i)–(ii) also imply that F_μ is absolutely continuous on any closed bounded interval, hence $TF_\mu = F_{|\mu|}$ is absolutely continuous by XIX.5.2.17.. We thus have (i) $\iff$ (ii) $\Rightarrow$ (iii) $\iff$ (iv) $\iff$ (v). We need to show (iii)-(v) $\Rightarrow$ (i)-(ii). If (v) holds, then F_μ is absolutely continuous by XIX.5.3.7., so we have $\mu((a, b)) = \mu([a, b]) = \int_a^b F' \, d\lambda$ for any $a, b \in \mathbb{R}$, $a < b$. 👍

XIX.5.4. The Bounded Variation Norm

XIX.5.4.10. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$. Let $BV([a,b])$ be the set of real-valued functions of bounded variation on $[a,b]$. Then $BV([a,b])$ is a real vector space, and the total variation $V_{[a,b]}$ is a seminorm on $BV([a,b])$ (). It is not a norm since the total variation of a constant function (and only a constant function) is 0. To make a norm, we must add on a "size" term which is another seminorm. It does not much matter what additional seminorm we use so long as it separates constant functions. Here are three standard possibilities:

$$\|f\|_{BV} = |f(a)| + V_{[a,b]}(f) \ .$$

$$\|f\|_{BV1} = \|f\|_1 + V_{[a,b]}(f) = V_{[a,b]}(f) + \int_a^b |f| \, d\lambda \ .$$

$$\|f\|_{BV\infty} = \|f\|_\infty + V_{[a,b]}(f) = V_{[a,b]}(f) + \sup_{a\le t\le b} |f(t)| \ .$$

(Note that a function of bounded variation is bounded and Lebesgue integrable, even Riemann integrable ().)

Note that if f is absolutely continuous on $[a,b]$, then by ()

$$\|f\|_{BV} = |f(a)| + \int_a^b |f'| \, d\lambda \ .$$

In this case, $\|f\|_{BV}$ is usually written $\|f\|_{AC}$. There are similar formulas for the other norms.

XIX.5.4.11. Proposition. The norms $\|\cdot\|_{BV}$, $\|\cdot\|_{BV1}$, and $\|\cdot\|_{BV\infty}$ are equivalent on $BV([a,b])$, and dominate $\|\cdot\|_\infty$, so if $f_n \to f$ in one (hence all) of these norms, then $f_n \to f$ uniformly on $[a,b]$.

Proof: If $f \in BV([a,b])$, we have $|f(a)| \le \|f\|_\infty \le |f(a)| + V_{[a,b]}(f)$, so $\|f\|_{BV} \le \|f\|_{BV\infty} \le 2\|f\|_{BV}$; thus $\|\cdot\|_{BV}$ and $\|\cdot\|_{BV\infty}$ are equivalent. We have $\|f\|_1 \le (b-a)\|f\|_\infty$, so $\|f\|_{BV1} \le K\|f\|_{BV\infty}$ for $K = \max(b-a, 1)$. Let $m = \inf_{a\le t\le b} |f(t)|$; then $|f(a)| \le m + V_{[a,b]}(f)$, so

$$(b-a)|f(a)| \le m(b-a) + (b-a)V_{[a,b]}(f) \le (b-a)V_{[a,b]}(f) + \int_{[a,b]} |f| \, d\lambda$$

and therefore $\|f\|_{BV} \le K\|f\|_{BV1}$ for $K = \max\left(2, \frac{1}{b-a}\right)$. 👍

XIX.5.4.12. It is conventional to take $\|\cdot\|_{BV}$ as the *bounded variation norm*, although we could equally well work with any of these norms. There are even other norms which are equivalent to these, such as

If $c \in [a,b]$, set $\|f\|_{BV(c)} = |f(c)| + V_{[a,b]}(f)$ (so $\|f\|_{BV} = \|f\|_{BV(a)}$).

If $1 < p < \infty$, set $\|f\|_{BVp} = \|f\|_p + V_{[a,b]}(f) = V_{[a,b]}(f) + \left[\int_{[a,b]} |f|^p \, d\lambda\right]^{1/p}$.

XIX.5.4.13. THEOREM. (i) The norm $\|\cdot\|_{BV}$ (hence all the other ones above) is a complete norm on $BV([a,b])$.
(ii) The set $BVC([a,b])$ of continuous functions of bounded variation on $[a,b]$ and the set $AC([a,b])$ of absolutely continuous functions functions on $[a,b]$ are closed subspaces of $BV([a,b])$.
Thus $(BV([a,b]),\|\cdot\|_{BV})$, $(BVC([a,b]),\|\cdot\|_{BV})$, and $(AC([a,b]),\|\cdot\|_{AC})$ are Banach spaces.

PROOF: (i): Let (f_n) be a Cauchy sequence in $BV([a,b])$ for $\|\cdot\|_{BV}$. Then (f_n) is a uniform Cauchy sequence by XIX.5.4.11., hence converges uniformly to a real-valued function f. We also have that the sequence $(V_{[a,b]}(f_n))$ is a Cauchy sequence in $\mathbb{R}$, hence is bounded. So it suffices to prove the following slightly stronger statement: if (f_n) is a sequence in $BV([a,b])$ converging uniformly to f on $[a,b]$, and $V_{[a,b]}(f_n) \le M$ for all n, then $f \in BV([a,b])$ (and $V_{[a,b]}(f) \le M$). To prove this, suppose $V_{[a,b]}(f) > M$, and fix a partition $\mathcal{P} = \{a = x_0, \dots, x_m = b\}$ of $[a,b]$ with $V(f,\mathcal{P}) > M$. Set

$$\epsilon = \frac{V(f,\mathcal{P}) - M}{2m}$$

and choose N such that $\|f_n - f\|_\infty < \epsilon$ for all $n \ge N$. If $n \ge N$, then

$$V(f,\mathcal{P}) = \sum_{k=1}^{m} |f(x_k) - f(x_{k-1})| \le \sum_{k=1}^{m} [|f(x_k) - f_n(x_k)| + |f_n(x_k) - f_n(x_{k-1})| + |f_n(x_{k-1}) - f(x_{k-1})|]$$
$$< \sum_{k=1}^{m} [\epsilon + |f_n(x_k) - f_n(x_{k-1})| + \epsilon] = V(f_n,\mathcal{P}) + 2m\epsilon \le M + (V(f,\mathcal{P}) - M) = V(f,\mathcal{P})$$

a contradiction.
(ii): Closure of $BVC([a,b])$ in $BV([a,b])$ follows immediately from XIX.5.4.11. and (). For $AC([a,b])$, suppose (f_n) is a sequence of absolutely continuous functions on $[a,b]$ which is Cauchy in the $\|\cdot\|_{BV}$ norm. We have

$$V_{[a,b]}(f_n - f_m) = \int_a^b |f_n' - f_m'|\, d\lambda$$

and, since (f_n) is a Cauchy sequence for $V_{[a,b]}$, it follows that (f_n') is a Cauchy sequence for the L^1-norm. Thus $f_n' \to g$ in the L^1-norm for some integrable function g on $[a,b]$. Also, $(f_n(a))$ is a Cauchy sequence in $\mathbb{R}$, hence converges to some y_0. (Actually, (f_n) converges to some $f \in BVC([a,b])$, and $y_0 = f(a)$.) For $x \in [a,b]$, set

$$G(x) = y_0 + \int_a^x g\, d\lambda\ .$$

Then G is absolutely continuous on $[a,b]$, and

$$\|G - f_n\|_{BV} = |y_0 - f_n(a)| + \int_a^b |g - f_n'|\, d\lambda$$

so $f_n \to G$ in the $\|\cdot\|_{BV}$ ($\|\cdot\|_{AC}$) norm. 👍

The following result is "almost" a strengthened form of V.8.5.5.. It is not exactly stronger since the hypotheses of V.8.5.5. on an interval I do not imply that the hypotheses of XIX.5.4.14. hold on closed bounded subintervals of I (under the hypotheses of V.8.5.5., the derivatives are not necessarily locally integrable), and the conclusion is only a.e.; but it is roughly a stronger statement since the convergence hypothesis on the derivatives is weaker and the concluded convergence of the functions is stronger.

XIX.5.4.14. THEOREM. Let (F_n) be a sequence of absolutely continuous functions on a closed bounded interval $[a,b]$, and (f_n) a sequence of integrable functions on $[a,b]$ with $f_n = F_n'$ a.e. Suppose

(a) The sequence (f_n) converges in the $\|\cdot\|_1$ norm on $[a,b]$ to a function f.

(b) There is an $x_0 \in [a,b]$ such that the sequence $(F_n(x_0))$ converges.

Then:

(i) The sequence (F_n) converges in the $\|\cdot\|_{AC}$ norm on $[a,b]$ to an (absolutely continuous) function F.

(ii) F is differentiable a.e. and $F'(x) = f(x)$ a.e. on $[a,b]$.

XIX.5.4.15. COROLLARY. Let $[a,b]$ be a closed bounded interval in $\mathbb{R}$, and (f_n) a sequence of integrable functions on $[a,b]$ converging in the L^1-norm to an (integrable) function f on $[a,b]$. Set

$$F_n(x) = \int_{[a,x]} f_n \, d\lambda \ , \quad F(x) = \int_{[a,x]} f \, d\lambda$$

for each n and each $x \in [a,b]$. Then the F_n and F are absolutely continuous on $[a,b]$, and $F_n \to F$ in the $\|\cdot\|_{AC}$ norm.

In other words, roughly speaking, convergence of a sequence of functions in a fairly weak sense implies convergence of the sequence of antiderivatives in a much stronger sense.

Corollary XIX.5.4.15. is clearly a special case of XIX.5.4.14.. Conversely, we will prove the Corollary and then reduce the proof of XIX.5.4.14. to XIX.5.4.15..

So we first prove the Corollary.

PROOF: Let $[a,b]$, (f_n), (F_n), f, F be as in the statement. We need only show $F_n \to F$ in the $\|\cdot\|_{AC}$ norm. We have $f_n(a) = F(a) = 0$ for all n, so

$$\|F_n - F\|_{AC} = \int_a^b |f_n - f| \, d\lambda$$

which goes to 0 by assumption. 👍

We now prove XIX.5.4.14.. The proof is some simple machinations to take care of the complication that x_0 is not necessarily a, and that the functions do not all agree at x_0.

PROOF: For each n, we have

$$F_n(a) = F_n(x_0) + \int_{x_0}^a f_n \, d\lambda = F_n(x_0) - \int_a^{x_0} f_n \, d\lambda \ .$$

If we set $G_n(x) = F_n(x) - F_n(a)$, so $G_n' = F_n' = f_n$ a.e. and $G_n(a) = 0$ for all n, then the f_n and G_n satisfy the hypotheses of XIX.5.4.15. with G_n in place of F_n, so (G_n) converges in the $\|\cdot\|_{AC}$ norm to the absolutely continuous function G, where

$$G(x) = \int_a^x f \, d\lambda$$

for $x \in [a, b]$. Set $F(x) = G(x) + [y_0 - G(x_0)]$, where

$$y_0 = \lim_{n\to\infty} F_n(x_0)$$

(so $F(x_0) = y_0$, F is absolutely continuous, and $F' = f$ a.e.). For all n we have that

$$\|F_n - F\|_{AC} = |F_n(a) - F(a)| + \int_a^b |f_n - f| \, d\lambda$$

$$= \left| F_n(x_0) - \int_a^{x_0} f_n \, d\lambda - F(x_0) + \int_a^{x_0} f \, d\lambda + \int_a^b |f_n - f| \, d\lambda \right|$$

$$\leq |F_n(x_0) - F(x_0)| + \int_a^{x_0} |f_n - f| \, d\lambda + \int_a^b |f_n - f| \, d\lambda$$

which goes to 0 as $n \to \infty$. 👍

XIX.5.4.16. There are a number of similar results; see () for one. The informal principle illustrated by all these results is that if $f_n \to f$ in a certain sense, and if F_n, F are antiderivatives adjusted by suitable constants so that $F_n(x_0) \to F(x_0)$ for a single x_0, then $F_n \to F$ in a considerably stronger sense than the sense in which $f_n \to f$.

XIX.5.5. Integration by Substitution and Integration by Parts

Using absolute continuity, we can give general versions of integration by substitution and by parts.

Integration by Parts

We first examine integration by parts. We begin with an observation of independent interest:

XIX.5.5.1. PROPOSITION. Let f and g be absolutely continuous functions on $[a,b]$. Then fg is also absolutely continuous on $[a,b]$.

PROOF: Since f and g are continuous on $[a,b]$, they are bounded. Fix positive constants M and N with $|f| \le M$ and $|g| \le N$. Let $\epsilon > 0$, and take δ_1 so that whenever $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta_1$, then $\sum_{k=1}^n |f(b_k) - f(a_k)| < \frac{\epsilon}{2N}$, and δ_2 so that whenever $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta_2$, then $\sum_{k=1}^n |g(b_k) - g(a_k)| < \frac{\epsilon}{2M}$. Set $\delta = \min(\delta_1, \delta_2)$. If $\{[a_k,b_k] : 1 \le k \le n\}$ is a set of nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, then, by the usual trick of adding and subtracting,

$$\begin{aligned}
\sum_{k=1}^n |f(b_k)g(b_k) - f(a_k)g(a_k)| &= \sum_{k=1}^n |f(b_k)g(b_k) - f(a_k)g(b_k) + f(a_k)g(b_k) - f(a_k)g(a_k)| \\
&\le \sum_{k=1}^n [|f(b_k)g(b_k) - f(a_k)g(b_k)| + |f(a_k)g(b_k) - f(a_k)g(a_k)|] \\
&= \sum_{k=1}^n |g(b_k)||f(b_k) - f(a_k)| + \sum_{k=1}^n |f(a_k)||g(b_k) - g(a_k)| \\
&\le N \sum_{k=1}^n |f(b_k) - f(a_k)| + M \sum_{k=1}^n |g(b_k) - g(a_k)| < N\frac{\epsilon}{2N} + M\frac{\epsilon}{2M} = \epsilon \ .
\end{aligned}$$

XIX.5.5.2. THEOREM. [INTEGRATION BY PARTS] Let F and G be absolutely continuous functions on $[a,b]$. Then $F'G$ and FG' (defined a.e. on $[a,b]$) are Lebesgue integrable on $[a,b]$, and

$$\int_a^b F'G \, d\lambda = F(b)G(b) - F(a)G(a) - \int_a^b FG' \, d\lambda \ .$$

PROOF: $F'G$ and FG' are integrable since F' and G' are integrable, and a product of an integrable function and a bounded measurable function is integrable. We have that F' exists on $[a,b] \setminus A$ and G' exists on $[a,b] \setminus B$, where A and B are null sets; thus F and G are differentiable on $[a,b] \setminus (A \cup B)$. Therefore by the Product Rule FG is differentiable on $[a,b] \setminus (A \cup B)$ and $(FG)' = F'G + FG'$ on $[a,b] \setminus (A \cup B)$, and in

particular $(FG)' = F'G + FG'$ a.e. (FG may be differentiable on a larger set, but this is irrelevant.) Since FG is absolutely continuous, we thus have

$$F(b)G(b) - F(a)G(a) = \int_a^b (FG)'\,d\lambda = \int_a^b (F'G + FG')\,d\lambda = \int_a^b F'G\,d\lambda + \int_a^b FG'\,d\lambda$$

and the result follows by moving the last term. 👍

XIX.5.5.3. Corollary. Let f and g be integrable functions on $[a,b]$. For $x \in [a,b]$, set $F(x) = \int_a^x f\,d\lambda$ and $G(x) = \int_a^x g\,d\lambda$. Then fG and Fg are integrable on $[a,b]$, and

$$\int_a^b fG\,d\lambda = F(b)G(b) - \int_a^b Fg\,d\lambda\ .$$

See XIX.1.8.2. for a more general Integration by Parts result.

XIX.5.5.4. The hypothesis that both F and G are absolutely continuous is essentially necessary: it is not enough that one is absolutely continuous and the other continuous and differentiable a.e. For example, let F be the constant function 1 and G the Cantor function. Then

$$\int_0^1 F'G\,d\lambda \neq F(1)G(1) - F(0)G(0) - \int_0^1 FG'\,d\lambda$$

since both integrals are zero. Other examples such as taking $F(x) = x$ and G the Cantor function also fail.

Integration by Substitution or Change of Variable

Now we examine integration by substitution. This also seems to have a straightforward argument along the same lines, based on the following general fact:

XIX.5.5.5. Proposition. Let ϕ be an absolutely continuous monotone function from $[a,b]$ to $[c,d]$, and let F be an absolutely continuous function on $[c,d]$. Then $F \circ \phi$ is absolutely continuous on $[a,b]$.

Proof: We prove the result for ϕ nondecreasing; the nonincreasing case is an easy notational modification. Let $\epsilon > 0$, and choose $\eta > 0$ such that if $[c_1,d_1],\ldots,[c_n,d_n]$ are nonoverlapping subintervals of $[c,d]$ with $\sum_{k=1}^n (d_k - c_k) < \eta$, then $\sum_{k=1}^n |F(d_k) - f(c_k)| < \epsilon$. Then choose $\delta > 0$ such that if $[a_1,b_1],\ldots,[a_n,b_n]$ are nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, then $\sum_{k=1}^n |\phi(b_k) - \phi(a_k)| < \eta$.

Let $[a_1,b_1],\ldots,[a_n,b_n]$ be nonoverlapping subintervals of $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$. Set $c_k = \phi(a_k)$ and $d_k = \phi(b_k)$. Since ϕ is nondecreasing, $[c_1,d_1],\ldots,[c_n,d_n]$ are nonoverlapping subintervals of $[c,d]$ (although some may be degenerate). We have $\sum_{k=1}^n (d_k - c_k) = \sum_{k=1}^n |\phi(b_k) - \phi(a_k)| < \eta$, so

$$\sum_{k=1}^n |F(\phi(b_k)) - F(\phi(a_k))| = \sum_{k=1}^n |F(d_k) - F(c_k)| < \epsilon\ .$$

It is not true that a composition of two absolutely continuous functions is necessarily absolutely continuous, even if the inside function is Lipschitz and the outside function is monotone (XIX.5.8.10.). Thus XIX.5.5.5. is rather special.

XIX.5.5.6. THEOREM. [INTEGRATION BY SUBSTITUTION OR CHANGE OF VARIABLE] Let ϕ be an absolutely continuous monotone function on an interval $[a,b]$, with $\phi([a,b]) = [c,d]$.

(i) If f is a nonnegative Lebesgue measurable function on $[c,d]$, then $(f\circ\phi)|\phi'|$ (which is defined a.e. on $[a,b]$, and nonnegative) is Lebesgue measurable on $[a,b]$, and

$$\int_a^b (f\circ\phi)|\phi'|\,d\lambda = \int_c^d f\,d\lambda\ .$$

(ii) If f is a Lebesgue integrable function on $[c,d]$, then $(f\circ\phi)|\phi'|$ (which is defined a.e. on $[a,b]$) is Lebesgue integrable on $[a,b]$, and

$$\int_a^b (f\circ\phi)|\phi'|\,d\lambda = \int_c^d f\,d\lambda\ .$$

XIX.5.5.7. The assumption that ϕ is absolutely continuous is essentially necessary: it is not enough that ϕ is continuous, and differentiable a.e. For example, let ϕ be the Cantor function, and f the constant function 1. Then

$$0 = \int_0^1 \phi'\,d\lambda = \int_0^1 (f\circ\phi)\phi'\,d\lambda \neq \int_0^1 f\,d\lambda = 1\ .$$

XIX.5.5.8. Here is a potential argument for (ii) which looks convincing at first glance: Set

$$F(x) = \int_c^x f\,d\lambda$$

for $x\in[c,d]$. Then F is absolutely continuous on $[c,d]$, so by XIX.5.5.5., $F\circ\phi$ is absolutely continuous, hence differentiable a.e. on $[a,b]$. By the Chain Rule, $(F\circ\phi)' = (F'\circ\phi)\phi'$, so since $F' = f$ a.e. we should have $(F\circ\phi)' = (f\circ\phi)\phi'$ a.e. Thus

$$\int_a^b (f\circ\phi)\phi'\,d\lambda = \int_a^b (F\circ\phi)'\,d\lambda = F(\phi(b)) - F(\phi(a)) = F(d) - F(c) = \int_c^d F'\,d\lambda = \int_a^b f\,d\lambda\ .$$

XIX.5.5.9. Unfortunately, there is a flaw in the argument. What the Chain Rule says is that *if* ϕ is differentiable at x and F is differentiable at $\phi(x)$, then $F\circ\phi$ is differentiable at x and

$$(F\circ\phi)'(x) = (F'\circ\phi)(x)\phi'(x)\ .$$

We have that F is differentiable on $[c,d]\setminus A$ for some null set A, and ϕ is differentiable on $[a,b]\setminus B$ for some null set B. Thus by the Chain Rule $F\circ\phi$ is differentiable on $[a,b]\setminus[\phi^{-1}(A)\cup B]$ and for x in this set,

$$(F\circ\phi)'(x) = (F'\circ\phi)(x)\phi'(x)\ .$$

The problem is that $\phi^{-1}(A)$ is not necessarily a null set! So even though $F\circ\phi$ is differentiable a.e. on $[a,b]$, we cannot conclude from the Chain Rule that it equals $(F'\circ\phi)\phi'$ a.e.

There is an additional problem: although $F'=f$ a.e. on $[c,d]$, it does not follow that $F'\circ\phi = f\circ\phi$ a.e. on $[a,b]$. Thus the argument has a seemingly irreparable flaw.

We will have to do considerable roundabout work to get past the flaw and prove the theorem. After the fact it will indeed turn out that $(F\circ\phi)' = (F'\circ\phi)\phi' = (f\circ\phi)\phi'$ a.e. on $[a,b]$ (XIX.5.8.8.).

We develop the proof of Theorem XIX.5.5.6. in a series of lemmas. We prove the theorem for ϕ nondecreasing (so $\phi'\geq 0$); the proof for ϕ nonincreasing is essentially identical, or see XIX.5.8.7..

XIX.5.5.10. LEMMA. Let ϕ be as in XIX.5.5.6., with ϕ nondecreasing. If D is a G_δ in $[c,d]$, then

$$\lambda(D) = \int_{\phi^{-1}(D)} \phi'\,d\lambda\ .$$

PROOF: First note that if (x,y) is an open interval in $[c,d]$, then $\phi^{-1}((x,y))$ is a relatively open subinterval of $[a,b]$ with endpoints u and v, $u<v$ (possibly containing u or v if $u=a$ or $v=b$), and

$$\int_{\phi^{-1}((x,y))} \phi'\,d\lambda = \int_u^v \phi'\,d\lambda = \phi(v)-\phi(u) = y-x = \lambda((x,y))\ .$$

A similar argument works for a relatively open subinterval with endpoints c and/or d. Thus the formula holds if D is a relatively open interval in $[c,d]$.

Now suppose U is a relatively open subset of $[c,d]$. Then U is a disjoint union of relatively open subintervals $\{I_n : n\in\mathbb{N}\}$. If $J_n=\phi^{-1}(I_n)$, then the J_n are pairwise disjoint relatively open subintervals of $[a,b]$ and $\phi^{-1}(U)=\cup_n J_n$, so

$$\int_{\phi^{-1}(U)} g'\,d\lambda = \int_{\cup_n J_n} \phi'\,d\lambda = \sum_{n=1}^\infty \left[\int_{J_n}\phi'\,d\lambda\right] = \sum_{n=1}^\infty \lambda(I_n) = \lambda(U)$$

so the formula holds if D is relatively open in $[a,b]$.

Finally, suppose $D=\cap_n U_n$ is a G_δ in $[a,b]$, with U_n relatively open and $U_{n+1}\subseteq U_n$ for all n. Then

$$\lambda(D) = \inf_n \lambda(U_n) = \lim_{n\to\infty}\lambda(U_n)$$

and $Y:=\phi^{-1}(D)=\cap_n Y_n$, where $Y_n=\phi^{-1}(U_n)$. If $g_n=\phi'\chi_{Y_n}$, then $g_n\to g:=\phi'\chi_Y$. Since all the g_n are dominated by ϕ', which is integrable, by the DCT (or downward MCT) we have

$$\int_Y \phi'\,d\lambda = \int_a^b g\,d\lambda = \lim_{n\to\infty}\int_a^b g_n\,d\lambda = \lim_{n\to\infty}\int_{Y_n}\phi'\,d\lambda = \lim_{n\to\infty}\lambda(U_n) = \lambda(D)\ .$$

We would like to get a similar formula for an arbitrary Lebesgue measurable subset A of $[c,d]$, but there is a technical difficulty: $\phi^{-1}(A)$ need not be Lebesgue measurable in general. To avoid the problem, we need to work more carefully. In the next lemmas, using the notation of XIX.5.5.6. with ϕ nondecreasing, set

$$P = \{x \in [a,b] : \phi'(x) \text{ exists and } \phi'(x) > 0\}$$

and $Q = [a,b] \setminus P$. Thus P and Q are Lebesgue measurable () and $\phi' = 0$ a.e. on Q.

XIX.5.5.11. LEMMA. Let ϕ be as in XIX.5.5.6., with ϕ nondecreasing. If $N \subseteq [c,d]$ with $\lambda(N) = 0$, then $\lambda^*(\phi^{-1}(N) \cap P) = 0$ (so in particular $\phi^{-1}(N) \cap P$ is Lebesgue measurable).

PROOF: Let D be a G_δ in $[c,d]$ with $N \subseteq D$ and $\lambda(D) = 0$. For $n \in \mathbb{N}$ set

$$P_n = \left\{ x \in [a,b] : \phi'(x) \text{ exists and } \phi'(x) \geq \frac{1}{n} \right\} .$$

Then P_n is measurable and $P = \cup_n P_n$. We have, for each n,

$$0 = \lambda(D) = \int_{\phi^{-1}(D)} \phi' \, d\lambda \geq \int_{\phi^{-1}(D) \cap P_n} \phi' \, d\lambda \geq \frac{1}{n} \lambda(\phi^{-1}(D) \cap P_n)$$

and thus $\lambda(\phi^{-1}(D) \cap P_n)$=0 for all n, and hence

$$\lambda^*(\phi^{-1}(N) \cap P) \leq \lambda(\phi^{-1}(D) \cap P) = \sup_n \lambda(\phi^{-1}(D) \cap P_n) = 0 .$$

XIX.5.5.12. LEMMA. Let ϕ be as in XIX.5.5.6., with ϕ nondecreasing. If $A \subseteq [c,d]$ is measurable, then $\phi^{-1}(A) \cap P$ is measurable, and

$$\int_{\phi^{-1}(A) \cap P} \phi' \, d\lambda = \lambda(A) .$$

PROOF: Write $A = D \setminus N$, where $D \subseteq [c,d]$ is a G_δ and N is a null set. Then $\phi^{-1}(A) = \phi^{-1}(D) \setminus \phi^{-1}(N)$. Since ϕ is continuous, $\phi^{-1}(D)$ is a Borel set (actually a G_δ), so $\phi^{-1}(D) \cap P$ is Lebesgue measurable. Also, $\phi^{-1}(N) \cap P$ is a null set by XIX.5.5.11., so $\phi^{-1}(A) \cap P = [\phi^{-1}(D) \cap P] \setminus [\phi^{-1}(N) \cap P]$ differs from $\phi^{-1}(D) \cap P$ by a null set, hence is Lebesgue measurable, and

$$\int_{\phi^{-1}(A) \cap P} \phi' \, d\lambda = \int_{\phi^{-1}(D) \cap P} \phi' \, d\lambda = \int_{\phi^{-1}(D) \cap P} \phi' \, d\lambda + \int_{\phi^{-1}(D) \cap Q} \phi' \, d\lambda = \int_{\phi^{-1}(D)} \phi' \, d\lambda = \lambda(D) = \lambda(A) .$$

XIX.5.5.13. LEMMA. Let ϕ be as in XIX.5.5.6., with ϕ nondecreasing. If f is a Lebesgue measurable function on $[c,d]$, then $g = (f \circ \phi)\phi'$ (defined a.e.) is Lebesgue measurable on $[a,b]$.

PROOF: We have that $g|_Q$ is zero a.e., hence Lebesgue measurable. If B is a Borel set in $\mathbb{R}$, then

$$(f \circ \phi)^{-1}(B) \cap P = \phi^{-1}(f^{-1}(B)) \cap P \ .$$

Since f is Lebesgue measurable, $f^{-1}(B)$ is a Lebesgue measurable subset of $[c,d]$, and hence $\phi^{-1}(f^{-1}(B)) \cap P$ is Lebesgue measurable by XIX.5.5.12.. Since this is true for all B, $(f \circ \phi)|_P$ is measurable. Since $\phi'|_P$ is also Lebesgue measurable, the product $g|_P$ is Lebesgue measurable by (). Since $g|_P$ and $g|_Q$ are both Lebesgue measurable, g is Lebesgue measurable by ().

XIX.5.5.14. We now finish the proof of XIX.5.5.6. (in the case where ϕ is nondecreasing). There is no issue of measurability by XIX.5.5.13.. We will bootstrap to prove the theorem. First suppose $f = \chi_A$ for some Lebesgue measurable set $A \subseteq [c,d]$. Then $f \circ \phi = \chi_{\phi^{-1}(A)}$, so

$$\int_a^b (f \circ \phi)\phi' \, d\lambda = \int_{\phi^{-1}(A) \cap P} \phi' \, d\lambda = \lambda(A) = \int_c^d \chi_A \, d\lambda = \int_c^d f \, d\lambda \ .$$

If f is a simple function, the formula also holds by linearity. If f is a nonnegative measurable function, then take an increasing sequence (f_n) of nonnegative simple functions converging pointwise to f; then by two applications of the MCT we have

$$\int_a^b (f \circ \phi)\phi' \, d\lambda = \sup_n \int_a^b (f_n \circ \phi)\phi' \, d\lambda = \sup_n \int_c^d f_n \, d\lambda = \int_c^d f \, d\lambda \ .$$

If f is an integrable function, the formula again holds by linearity.

This completes the proof of Theorem XIX.5.5.6. (for ϕ nondecreasing).

See [HS75, 20.5] for a somewhat quicker proof using the Radon-Nikodým Theorem.

XIX.5.6. L^p Versions, $p > 1$

There is an L^p analog of absolute continuity for $p > 1$. Surprisingly, the distinction between absolute continuity and bounded variation disappears in this case. This section is adapted from [**?**].

XIX.5.6.1. THEOREM. Let F be a function from $[a,b]$ to $\mathbb{R}$, and $p > 1$. The following are equivalent:

(i) There is a measurable function f on $[a,b]$ with $\displaystyle\int_a^b |f|^p \, d\lambda < \infty$ (i.e. $f \in L^p([a,b])$) and

$$F(x) = F(a) + \int_a^x f \, d\lambda$$

for all $x \in [a,b]$.

(ii) There is a constant M such that whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are nonoverlapping subintervals of $[a, b]$, with $h_k = b_k - a_k$ for each k, we have

$$\sum_{k=1}^{n} |F(b_k) - F(a_k)|^p h_k^{1-p} \leq M \ .$$

(iii) There is a constant M such that whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are nonoverlapping subintervals of $[a, b]$, with $h_k = b_k - a_k$ for each k, we have

$$\sum_{k=1}^{n} |F(b_k) - F(a_k)|^p h_k^{1-p} \leq M$$

and, for every $\epsilon > 0$, there is a $\delta > 0$ such that whenever $[a_1, b_1], \ldots, [a_n, b_n]$ are nonoverlapping subintervals of $[a, b]$, with $h_k = b_k - a_k$ for each k and $\sum_{k=1}^{n} h_k < \delta$,

$$\sum_{k=1}^{n} |F(b_k) - F(a_k)|^p h_k^{1-p} < \epsilon \ .$$

If these conditions are satisfied, the smallest M in (ii) and (iii) is precisely $\displaystyle\int_a^b |f|^p \, d\lambda$.

XIX.5.6.2. The conditions in (iii) are apparently more restrictive than in (ii). The condition in (i) implies that f is integrable on $[a, b]$, and hence that F is absolutely continuous on $[a, b]$.

If $p = 1$, the condition in (ii) is precisely that F have bounded variation on $[a, b]$, and the second condition in (iii) is that F be absolutely continuous on $[a, b]$, which implies that F has bounded variation and hence satisfies the first condition of (iii). Thus Theorem XIX.5.2.5. says that (i) and (iii) are equivalent also if $p = 1$; however, (ii) is not equivalent to (i) and (iii) in general if $p = 1$ (the Cantor function is a counterexample).

XIX.5.6.3. We now give the proof of XIX.5.6.1..

PROOF: (iii) $\Rightarrow$ (ii) is trivial.

(i) $\Rightarrow$ (iii): By (), for each k we have

$$|F(b_k) - F(a_k)| = \left| \int_{a_k}^{b_k} f \, d\lambda \right| \leq \left[\int_{a_k}^{b_k} |f|^p \, d\lambda \right]^{1/p} \left[\int_{a_k}^{b_k} 1 \, d\lambda \right]^{1-1/p} = h_k^{1-1/p} \left[\int_{a_k}^{b_k} |f|^p \, d\lambda \right]^{1/p}$$

so

$$\sum_{k=1}^{n} |F(b_k) - F(a_k)|^p h_k^{1-p} \leq \sum_{k=1}^{n} h_k^{1-p} \left[h_k^{1-1/p} \right]^p \left[\int_{a_k}^{b_k} |f|^p \, d\lambda \right] = \sum_{k=1}^{n} \left[\int_{a_k}^{b_k} |f|^p \, d\lambda \right] \leq \int_a^b |f|^p \, d\lambda \ .$$

Also, by (), for any $\epsilon > 0$ there is a $\delta > 0$ such that if $\sum_{k=1}^{n} h_k < \delta$,

$$\sum_{k=1}^{n} \left[\int_{a_k}^{b_k} |f|^p \, d\lambda \right] < \epsilon$$

so condition (iii) is satisfied with $M = \int_a^b |f|^p \, d\lambda$.

(ii) $\Rightarrow$ (i): Let $[a_1, b_1], \ldots, [a_n, b_n]$ be nonoverlapping subintervals of $[a, b]$. Then, by Hölder's Inequality (), we have

$$\sum_{k=1}^{n} |F(b_k) - F(a_k)| = \sum_{k=1}^{n} \left[|F(b_k) - F(a_k)| h_k^{1/p-1} \right] h_k^{1-1/p} = \sum_{k=1}^{n} \left[|F(b_k) - F(a_k)|^p h_k^{1-p} \right]^{1/p} h_k^{1-1/p}$$

$$\leq \left(\sum_{k=1}^{n} \left[|F(b_k) - F(a_k)|^p h_k^{1-p} \right] \right)^{1/p} \left(\sum_{k=1}^{n} h_k \right)^{1-1/p} \leq M^{1/p} \left(\sum_{k=1}^{n} h_k \right)^{1-1/p}$$

and thus, for any $\epsilon > 0$, there is a $\delta > 0$ such that if $\sum_{k=1}^{n} h_k < \delta$, $\sum_{k=1}^{n} |F(b_k) - F(a_k)| < \epsilon$. Thus F is absolutely continuous. Hence there is an integrable function f on $[a, b]$ with

$$F(x) = F(a) + \int_a^x f \, d\lambda$$

for all $x \in [a, b]$. It remains to show that $\int_a^b |f|^p \, d\lambda \leq M$.

For each n let $\mathcal{P}_n = \{a = x_{n,0}, x_{n,1}, \ldots, x_{n,n} = b\}$ be a partition of $[a, b]$ into n subintervals of equal length. For $x_{n,k} \leq x < x_{n,k+1}$, set

$$f_n(x) = \frac{F(x_{n,k+1}) - F(x_{n,k})}{x_{n,k+1} - x_{n,k}} \ .$$

Since F is absolutely continuous, it is differentiable a.e. and $F' = f$ a.e. Suppose $x_{n,k} < x < x_{n,k+1}$ and F is differentiable at x. Then

$$f_n(x) = \frac{F(x_{n,k+1}) - F(x)}{x_{n,k+1} - x} \frac{x_{n,k+1} - x}{x_{n,k+1} - x_{n,k}} + \frac{F(x) - F(x_{n,k})}{x - x_{n,k}} \frac{x - x_{n,k}}{x_{n,k+1} - x_{n,k}}$$

$$= (F'(x) + \eta_1) \frac{x_{n,k+1} - x}{x_{n,k+1} - x_{n,k}} + (F'(x) + \eta_2) \frac{x - x_{n,k}}{x_{n,k+1} - x_{n,k}} = F'(x) + \eta_3$$

where η_1, η_2, η_3 depend on x and n; for fixed x and n we have $|\eta_3| \leq |\eta_1| + |\eta_2|$, and for fixed x we have $\eta_1, \eta_2 \to 0$ as $n \to \infty$. So if x is not one of the $x_{n,k}$ for any n and F is differentiable at x, $f_n(x) \to F'(x)$. Thus $f_n \to f$ a.e. By Fatou's Lemma (),

$$\int_a^b |f|^p \, d\lambda \leq \liminf_{n \to \infty} \int_a^b |f_n|^p \, d\lambda = \liminf_{n \to \infty} \left[\sum_{k=1}^{n} \frac{|F(x_{n,k}) - F(x_{n,k-1})|^p}{(x_{n,k} - x_{n,k-1})^p} (x_{n,k} - x_{n,k-1}) \right] \leq M \ .$$

XIX.5.7. The Lebesgue Density Theorem

If A and B are Lebesgue measurable subsets of $\mathbb{R}$, with $0 < \lambda(B) < \infty$, it makes sense to interpret $\frac{\lambda(A \cap B)}{\lambda(B)}$ as the "fraction of the points of B which are in A." We will particularly make this interpretation when B is a finite interval.

If A is a measurable subset of $\mathbb{R}$ and $c \in \mathbb{R}$, we similarly want to carefully define "the fraction of points near c which are in A." There is a way to do this, and it leads to a remarkable result (XIX.5.7.3.): for almost all points x of A, almost all points near x are in A, and for almost all points y of A^c, almost all points near y are in A^c. As a consequence, a Lebesgue measurable subset of $\mathbb{R}$ is "lumpy" and cannot be distributed very evenly throughout $\mathbb{R}$, or an interval in $\mathbb{R}$, unless it is very densely or very sparsely distributed in that interval in a measure-theoretic sense.

XIX.5.7.1. DEFINITION. Let A be a Lebesgue measurable subset of $\mathbb{R}$. If $c \in \mathbb{R}$, define the left and right upper and lower densities of A at c to be

The *left lower density* $d_-(A,c) = \liminf\limits_{x \to c-} \dfrac{\lambda([x,c] \cap A)}{\lambda([x,c])}$.

The *left upper density* $d^-(A,c) = \limsup\limits_{x \to c-} \dfrac{\lambda([x,c] \cap A)}{\lambda([x,c])}$.

The *right lower density* $d_+(A,c) = \liminf\limits_{x \to c+} \dfrac{\lambda([c,x] \cap A)}{\lambda([c,x])}$.

The *right upper density* $d^+(A,c) = \limsup\limits_{x \to c+} \dfrac{\lambda([c,x] \cap A)}{\lambda([c,x])}$.

If $d_-(A,c) = d^-(A,c)$, the *left density* $d_l(A,c)$ is defined and equals the common value.
If $d_+(A,c) = d^+(A,c)$, the *right density* $d_r(A,c)$ is defined and equals the common value.
If $d_l(A,c) = d_r(A,c)$, the *density* $d(A,c)$ is defined and equals the common value.

Note that the left and right upper and lower densities are always defined and are between 0 and 1. It should not be expected that left or right densities, much less densities, always exist (indeed, see XIX.5.8.12.); in fact, it is remarkable that the density exists for almost all c (and even more is true).

Comparing definitions, we immediately get the following characterization of densities as Dini derivatives:

XIX.5.7.2. PROPOSITION. Let $A \subseteq \mathbb{R}$ be Lebesgue measurable, and $c \in \mathbb{R}$. Fix $a < c$, and define

$$F(x) = \int_a^x \chi_A \, d\lambda$$

for $x \geq a$. then $d_-(A,c)$, $d^-(A,c)$, $d_+(A,c)$, $d^+(A,c)$ are the Dini derivatives () $D_-F(c)$, $D^-F(c)$, $D_+F(c)$, $D^+F(c)$ respectively. Thus $d_l(A,c)$ $[d_r(A,c)]$ is defined if and only if F is left [right] differentiable at c, and $d_l(A,c) = D_LF(c)$ $[d_r(A,c) = D_RF(c)]$. The density $d(A,c)$ is defined if and only if F is differentiable at c, and $d(A,c) = F'(c)$.

XIX.5.7.3. THEOREM. [LEBESGUE DENSITY THEOREM] Let A be a Lebesgue measurable subset of $\mathbb{R}$. Then the density of A exists and is 1 for almost all points of A, and exists and is 0 for almost all points of A^c.

PROOF: This is an almost immediate corollary of XIX.5.2.15.. We may assume A is bounded, since for $x \in (a,b)$ the density of A at x is the same as the density of $A \cap (a,b)$ at x. So suppose $A \subseteq (a,b)$. Set

$$F(x) = \int_a^x \chi_A \, d\lambda$$

for $x \in [a,b]$. Then F is differentiable a.e. and $F' = \chi_A$ a.e. on $[a,b]$, i.e. $F'(x) = 1$ for almost all $x \in A$ and $F'(x) = 0$ for almost all $x \in [a,b] \setminus A$. 👍

The next Corollary makes precise the statement that Lebesgue measurable sets are "lumpy" and cannot be evenly distributed.

XIX.5.7.4. COROLLARY. Let A be a measurable subset of $\mathbb{R}$, and I a bounded interval in $\mathbb{R}$.
(i) If $\lambda(A \cap I) > 0$, then for every $\epsilon > 0$ there is a subinterval J of I such that

$$\frac{\lambda(A \cap J)}{\lambda(J)} \geq 1 - \epsilon \ .$$

(ii) If $\lambda(A \cap I) < \lambda(I)$, then for every $\epsilon > 0$ there is a subinterval J of I such that

$$\frac{\lambda(A \cap J)}{\lambda(J)} \leq \epsilon \ .$$

XIX.5.7.5. Thus, for example, there cannot be a Lebesgue measurable set A such that $\lambda(A \cap I) = \frac{1}{2}\lambda(I)$ for every bounded interval I, i.e. such that A contains "half" the points in every interval. In fact, if A is Lebesgue measurable, there is no $\epsilon > 0$ such that

$$\epsilon < \frac{\lambda(A \cap I)}{\lambda(I)} < 1 - \epsilon$$

for every bounded interval I.

XIX.5.7.4. does not hold for $\epsilon = 0$ in general (XIX.5.8.14.)

XIX.5.8. Exercises

XIX.5.8.1. Show that a function $f : [a,b] \to \mathbb{R}$ (or $\mathbb{C}$) is absolutely continuous if and only if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $[a_1,b_1], \ldots, [a_n,b_n]$ are finitely many pairwise disjoint intervals in $[a,b]$ with $\sum_{k=1}^n (b_k - a_k) < \delta$, we have $\sum_{k=1}^n |f(b_k) - f(a_k)| < \epsilon$.

This is the definition of absolute continuity in many references. The "nonoverlapping" version we use is more suitable for some arguments.

XIX.5.8.2. [Sak64] Let A be a Cantor set of positive measure in $[0,1]$, containing 0 and 1. Let $\{I_n : n \in \mathbb{N}\}$ be the open intervals in $[0,1] \setminus A$, in decreasing order of length, and c_n the midpoint of I_n. Let ℓ_n be the length of I_n, and r_n the length of the longest subinterval of $[0,1]$ disjoint from $I_1, \dots, I_n$. Define a continuous function $f : [0,1] \to \mathbb{R}$ by letting f be identically zero on A, $f(c_n) = \ell_n + r_n$, and f piecewise-linear on I_n.

(a) Show that $r_n \to 0$ (obviously $\ell_n \to 0$), and hence f is indeed continuous.

(b) Show that f is absolutely continuous on each $\bar{I}_n$.

(c) Show that f has Property (N). [If N is a null set in $[0,1]$, write $N = \cup_{k=0}^{\infty} N_k$, where $N_0 = N \cap A$ and $N_k = N \cap I_k$ for $k \geq 1$. Show that $\lambda(f(N_k)) = 0$ for all k.] This f is an *ACG function* in the language of [Sak64].

(d) If $x \in A$, show that $D_- f(x) \leq 0$ and $D^+ f(x) \geq 0$, and either $D_- f(x) \leq -1$ or $D^+ f(x) \geq 1$ (or both). Hence f fails to be differentiable anywhere on A, i.e. on a set of positive measure. Conclude that f does not have bounded variation.

(e) Show that if $x > 0$, then $N_f(x)$ is finite. Thus f has Property (S) by XIX.5.1.8..

XIX.5.8.3. Let f be as in XIX.5.8.2., and set $g(x) = x + f(x)$.

(a) Show that g has property (N). [See XIX.5.8.2.(c).]

(b) Show that for almost all $x \in A$, $N_g(x) = +\infty$. Thus N_g is not finite a.e.

(c) By Theorem XIX.5.1.8., g also does not satisfy property (S).

XIX.5.8.4. (a) Show that if f is continuous and real-valued on $[a,b]$, then $V_{[a,b]}(|f|) = V_{[a,b]}(f)$, and in particular f has bounded variation if and only if $|f|$ has bounded variation.

(b) Show that if f is continuous and real-valued, then f is absolutely continuous if and only if $|f|$ is absolutely continuous. [**Caution:** $|f|' \neq |f'|$ in general!]

(c) Show that the statements can fail if f is complex-valued [consider e^{ig} for g real-valued], or if f is not continuous.

XIX.5.8.5. Verify that the function Φ defined in the proof of Lemma XIX.5.2.29. has all the required properties.

XIX.5.8.6. Show that Lemma XIX.5.2.28. is essentially an explicit version of XX.10.3.6., as follows.

(a) Deduce XIX.5.2.28. from XX.10.3.6. by taking H to be a singular function approximating Ψ. [One additional correction with a scaled Cantor function is needed to make $H(a) = \Psi(a)$ and $H(b) = \Psi(b)$.]

(b) Slightly modify the proof of XIX.5.2.28. to obtain XX.10.3.6.. [If G is any continuous function, construct H to approximate $\Psi - G$.]

XIX.5.8.7. Deduce the Integration by Substitution formula for ϕ nonincreasing from the nondecreasing case. [Hint: if ϕ is nonincreasing, set $\psi(x) = \phi(a + b - x)$.]

XIX.5.8.8. Justify the statement in XIX.5.5.9. that $(F \circ \phi)' = (F' \circ \phi)\phi' = (f \circ \phi)\phi'$ a.e. on $[a,b]$ using XIX.5.5.5., XIX.5.5.6., and XIX.5.2.14..

XIX.5.8.9. Let f be continuous and monotone on $[a,b]$, and differentiable except possibly at countably many points of $[a,b]$. Show that f is absolutely continuous on $[a,b]$. [Use XIX.5.2.19..]

XIX.5.8.10. (a) Show that if $f:[a,b]\to[c,d]$ is absolutely continuous and g is Lipschitz on $[c,d]$, then $g\circ f$ is absolutely continuous on $[a,b]$.

(b) Show that $f(x)=x^2\sin^2\frac{1}{x}$ $(f(0)=0)$ is Lipschitz on $[0,1]$ and $f([0,1])\subseteq[0,1]$. [f is differentiable.]

(c) Let $g(x)=\sqrt{x}$. Show that g is absolutely continuous on $[0,1]$ (and obviously increasing). [Use XIX.5.8.9..]

(d) Show that $g\circ f$ does not have bounded variation on $[0,1]$, and in particular is not absolutely continuous. Contrast with (a) and XIX.5.5.5..

XIX.5.8.11. In the situation of XIX.5.5.10., if ϕ is strictly increasing, then

(a) Show that if E is a G_δ in $[a,b]$, then $\phi(E)$ is a G_δ in $[c,d]$.

(b) Show that if B is a Lebesgue measurable subset of $[a,b]$, then

$$\lambda(\phi(B))=\int_B \phi'\,d\lambda$$

(note that $\phi(B)$ is measurable by XIX.5.1.4.).

(c) Does this allow the argument following XIX.5.5.10. to be eliminated or simplified in the case where ϕ is strictly increasing? (In particular, if $A\subseteq[c,d]$ is Lebesgue measurable, is $\phi^{-1}(A)$ necessarily Lebesgue measurable? See XVII.5.2.3..)

XIX.5.8.12. Fix $c\in\mathbb{R}$ and m,M with $0\le m\le M\le 1$. Show that a sequence

$$a_1<b_1<a_2<b_2<\cdots<c$$

of real numbers converging to c can be constructed so that, if $A=\{c\}\cup\bigcup_{n=1}^{\infty}[a_n,b_n]$, then $d_-(A,c)=m$ and $d^-(A,c)=M$. Do the same for right densities. Thus the four densities at a given point can be arbitrary within the obvious constraints, even if A is a closed set (or open set).

XIX.5.8.13. Let A and B be Lebesgue measurable subsets of $\mathbb{R}$. If $\lambda(A\triangle B)=0$, show that $d_-(A,x)=d_-(B,x)$ for all $x\in\mathbb{R}$, and similarly for d^-, d_+, d^+.

XIX.5.8.14. Construct a Lebesgue measurable subset A of $\mathbb{R}$ such that

$$0<\frac{\lambda(A\cap I)}{\lambda(I)}<1$$

for every bounded interval I in $\mathbb{R}$. (Note that these numbers cannot be bounded away from 0 or 1 by XIX.5.7.4..) [Let f be a differentiable function which is not monotone on any interval (V.12.1.7., XIII.12.5.7.), and let $A=\{x:f'(x)>0\}$. Use XIX.5.1.19..]

XIX.5.8.15. Can there be a nonmeasurable subset A of $\mathbb{R}$ with the property that $\lambda^*(A\cap I)=\frac{1}{2}\lambda(I)$ for every bounded interval I?

XIX.6. Approximate Limits and Continuity

Our exposition in this section is largely adapted from [Fed69].

XIX.6.1. The Radon-Nikodým Theorem in $\mathbb{R}^n$

In what sense is the Radon-Nikodým "derivative" a true derivative? It is in the special case of a finite Borel measure on $\mathbb{R}$ absolutely continuous with respect to Lebesgue measure (XVIII.6.1.7.), and it has formal properties of a derivative (XVIII.6.1.8.). But in general it is not really a derivative in the calculus sense.

However, there is a nice interpretation of the Radon-Nikodým derivative in $\mathbb{R}^n$ as a sort of "derivative" in the literal sense, explained in this section. We will work more generally with Borel measures on a controlled (XVII.8.1.1.) metric space (X, ρ) so that the Besicovitch Covering Theorem (XVII.8.3.2.) can be used. Recall that $\mathbb{R}^n$ with the usual metric is controlled (XVII.8.1.6.), so the reader can just take $X = \mathbb{R}^n$ if desired. It is worthwhile to work in greater generality since the argument is identical and there are applications, e.g. in Geometric Measure Theory, where the greater generality is used. We denote by μ^* the outer measure induced by μ (XVII.6.1.6.) (so we can consider μ^*-measurable sets which are not Borel sets).

Recall the following (XVII.6.1.14.):

XIX.6.1.1. Proposition. Let μ be a Borel measure on a separable metrizable (i.e. second countable regular) space X. The following are equivalent:

(i) μ is neighborhood-finite, i.e. every point has a neighborhood of finite measure.

(ii) μ is locally finite (every set of finite measure is contained in an open set of finite measure), and μ has no infinite atoms.

(iii) μ is locally finite and σ-finite.

(iv) μ is regular and σ-finite.

If every closed bounded subset of X is compact (e.g. if $X = \mathbb{R}^n$ with the usual metric), these are also equivalent to

(v) μ is bounded-finite (XVII.6.1.15.).

XIX.6.1.2. Our setup is as follows. Let (X, ρ) be a separable controlled metric space, and μ and ν neighborhood-finite Borel measures on X, with $\nu \ll \mu$. (The reader can just take the most important case $X = \mathbb{R}^n$, or a subset, with the usual metric and μ Lebesgue measure for simplicity if desired.) Since μ and ν are σ-finite, there is a Radon-Nikodým derivative $\frac{d\nu}{d\mu}$. Fix a Radon-Nikodým derivative f, i.e. f is a nonnegative locally integrable Borel-measurable function on X with

$$\nu(E) = \int_E f \, d\mu$$

for each Borel set E.

XIX.6.1.3. There is a largest open set W in X with $\mu(W) = 0$: if $\mathcal{U}$ is a countable base for the topology, W is the union of all basic open sets of μ-measure 0. Since $\nu \ll \mu$, we also have $\nu(W) = 0$. Thus, for almost all $x \in X$ (for all $x \in W^c$), we have that every neighborhood of x has positive μ-measure (and sufficiently small neighborhoods have finite μ- and ν-measure).

XIX.6.1.4. If $x \in W^c$, and $\delta > 0$ is small enough that $\mu(B_\delta(x)) < \infty$, define

$$A(x,\delta) = \frac{\nu(B_\delta(x))}{\mu(B_\delta(x))} = \frac{1}{\mu(B_\delta(x))} \int_{B_\delta(x)} f \, d\mu$$

which can be interpreted as the "average value of f over $B_\delta(x)$." Thus for almost all $x \in X$, $A(x,\delta)$ is defined for all sufficiently small $\delta > 0$.

The main result is:

XIX.6.1.5. THEOREM. For μ-almost all $x \in X$, $\lim_{\delta\to 0} A(x,\delta)$ exists, is finite, and equals f.

XIX.6.1.6. One cannot expect better than an "almost everywhere" result, since for one thing f is only uniquely determined almost everywhere, and for another $A(x,\delta)$ is only defined for almost all x (and small δ), so the limit can only exist almost everywhere in general. There will be additional restrictions too.

The proof of the theorem requires several lemmas. We use the outer measures defined by μ and ν: if $Y \subseteq X$, set

$$\mu^*(Y) = \inf\{\mu(B) : B \text{ Borel}, Y \subseteq B\} = \min\{\mu(B) : B \text{ Borel}, Y \subseteq B\}$$

and similarly for ν^*.

XIX.6.1.7. LEMMA. Use the above notation.
(i) If $c > 0$ and

$$A \subseteq \{x \in W^c : \liminf_{\delta\to 0} A(x,\delta) < c\}$$

then $\nu^*(A) \le c\mu^*(A)$.
(ii) If $c \ge 0$ and

$$A \subseteq \{x \in W^c : \limsup_{\delta\to 0} A(x,\delta) > c\}$$

then $\nu^*(A) \ge c\mu^*(A)$.

PROOF: (i) If $\mu^*(A)$ is ∞ (or 0) there is nothing to prove, so assume $\mu^*(A) < \infty$. Let $\epsilon > 0$. Since μ is regular (XVII.6.1.10.), there is an open set V with $A \subseteq V$ and $\mu(V) < \mu^*(A) + \epsilon$. If $x \in A$, then for sufficiently many small $\delta > 0$, $A(x,\delta) < c$. If $\mathcal{C}$ is the set of open balls B contained in V for which $\nu(B) < c\mu(B)$, then $\mathcal{C}$ is a Besicovitch covering of A, so there is a sequence (B_k) of pairwise disjoint open balls in $\mathcal{C}$ which μ-almost cover A and hence ν-almost cover A. Then we have

$$\nu^*(A) \le \nu(\cup_k B_k) = \sum_{k=1}^{\infty} \nu(B_k) < c \sum_{k=1}^{\infty} \mu(B_k) = c\mu(\cup_k B_k) \le c\mu(V) < c\mu^*(A) + c\epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, the result follows. The proof of (ii) is almost identical (assuming $\mu(A) > 0$). 👍

XIX.6.1.8. LEMMA. With notation as above, $\lim_{\delta\to 0} A(x,\delta)$ exists and is finite for almost all $x \in X$.

PROOF: The set of x such that the limit does not exist or is infinite is the union of the following sets:

The open set W.

The set $Z = \{x \in W^c : \limsup_{\delta>0} A(x,\delta) = +\infty\}$.

The sets $R_{a,b}$ for $a, b \in \mathbb{Q}$, $0 < a < b$, where

$$R_{a,b} = \{x \in W^c : \liminf_{\delta\to 0} A(x,\delta) < a < b < \limsup_{\delta\to 0} A(x,\delta)\} \ .$$

We have $\mu(W) = 0$. To show that $\mu^*(Z) = 0$, let $x \in Z$; then there is a $\delta > 0$ with $\nu(B_\delta(x)) < \infty$. By Lemma XIX.6.1.7., for any c we have $\nu^*(Z \cap B_\delta(x)) \geq c\mu^*(Z \cap B_\delta(x))$; thus $\mu^*(Z \cap B_\delta(x)) = 0$. So, if $\mathcal{U}$ is a countable base for the topology of X, the set of $U \in \mathcal{U}$ such that $\mu^*(Z \cap U) = 0$ covers X. Thus $\mu^*(Z) = 0$. Now fix $a, b \in \mathbb{Q}$, $0 < a < b$. If $x \in R_{a,b}$ and δ is small enough that $0 < \mu(B_\delta(x)) < \infty$, we have by Lemma XIX.6.1.7.

$$b\mu^*(R_{a,b} \cap B_\delta(x)) \leq \nu^*(R_{a,b} \cap B_\delta(x)) \leq a\mu^*(R_{a,b} \cap B_\delta(x))$$

which implies that $\mu^*(R_{a,b} \cap B_\delta(x)) = 0$ (and $\nu^*(R_{a,b} \cap B_\delta(x)) = 0$) since $a < b$. Thus, for the countable base $\mathcal{U}$, the $U \in \mathcal{U}$ such that $\mu^*(R_{a,b} \cap U) = 0$ cover X, and therefore $\mu^*(R_{a,b}) = 0$. 👍

XIX.6.1.9. LEMMA. Using the above notation, set $g(x) = \lim_{\delta>0} A(x,\delta)$ (defined μ-a.e.). Then g is μ^*-measurable, hence equal μ-a.e. to a Borel measurable function.

PROOF: Let $a, b \in \mathbb{R}$, $a < b$, $Y \subseteq X$ with $\mu^*(Y) < \infty$, $A_1 = Y \cap \{x \in X : g(x) < a\}$, $A_2 = Y \cap \{x \in X : g(x) > b\}$. Let B_1 and B_2 be Borel sets with $A_j \subseteq B_j$, $\mu^*(A_j) = \mu(B_j)$, and $\nu^*(A_j) = \nu(B_j)$ for $i = 1, 2$. Then

$$\nu(B_1 \cap B_2) = \nu(A_1 \cap B_2) \leq a\mu(A_1 \cap B_2) \leq a\mu(B_1 \cap B_2)$$

$$\nu(B_1 \cap B_2) \geq \nu(B_1 \cap A_2) \geq b\mu(B_1 \cap A_2) = b\mu(B_1 \cap B_2)$$

by () and XIX.6.1.7., and it follows that $\mu(B_1 \cap B_2) = 0$. Thus

$$\mu^*(Y) = \mu^*(Y \cap B_1) + \mu^*(Y \cap B_2) \geq \mu^*(A_1) + \mu^*(A_2)$$

and so g is μ^*-measurable by XVII.5.1.4.. 👍

XIX.6.1.10. LEMMA. Using the above notation, $g = f$ μ-a.e. on X.

This completes the proof of Theorem XIX.6.1.5.. 💪

XIX.6.1.11. Essentially the same proof works in the case where ν and even μ are signed Borel measures. Details are left to the reader.

XIX.6.1.12. As a closely related matter, if μ is a neighborhood-finite Borel measure on X and f is a locally μ-integrable Borel measurable function on X, define

$$\nu(E) = \int_E f\, d\mu$$

for E a Borel subset of X. Then $\nu \ll \mu$ and f is a Radon-Nikodým derivative for ν with respect to μ. So we have:

XIX.6.1.13. Corollary. Let μ be a neighborhood-finite Borel measure on a separable controlled metric space (X, ρ), and f a locally μ-integrable Borel measurable function on X. Then

$$\lim_{\delta \to 0} \frac{1}{\mu(B_\delta(x))} \int_{B_\delta(x))} f\, d\mu$$

exists and equals $f(x)$ for μ-almost all $x \in X$.

We can actually do a little better:

XIX.6.1.14. Theorem. Let μ be a neighborhood-finite Borel measure on separable controlled metric space (X, ρ), and f a locally μ-integrable Borel measurable function on X. Then

$$\lim_{\delta \to 0} \frac{1}{\mu(B_\delta(x))} \int_{B_\delta(x))} |f(z) - f(x)|\, d\mu(z)$$

exists and equals 0 for μ-almost all $x \in X$. The set of such x is called the *Lebesgue set* of f.

Proof: For each $x \in X$ let $T(x)$ be the set of $y \in \mathbb{R}$ such that

$$\lim_{\delta \to 0} \frac{1}{\mu(B_\delta(x))} \int_{B_\delta(x))} |f(z) - y|\, d\mu(z)$$

exists and equals $|f(x) - y|$. If $y, y' \in \mathbb{R}$, then for any x

$$\frac{1}{\mu(B_\delta(x))} \int_{B_\delta(x))} |f(z) - y|\, d\mu(z) - \frac{1}{\mu(B_\delta(x))} \int_{B_\delta(x))} |f(z) - y'|\, d\mu(z) \leq |y - y'|$$

for any δ, and it follows that $T(x)$ is closed in $\mathbb{R}$. For each $y \in \mathbb{R}$, $y \in T(x)$ for μ-almost all x by XIX.6.1.13. applied to $g(x) = |f(x) - y|$. For $y \in \mathbb{Q}$ let A_y be the set of x for which $y \in T(x)$; then $\mu^*(X \setminus A_y) = 0$. If $A = \cap_{y \in \mathbb{Q}} A_y$, then $\mu^*(X \setminus A) = 0$ and $\mathbb{Q} \subseteq T(x)$ for all $x \in A$. Since $T(x)$ is closed in $\mathbb{R}$, $T(x) = \mathbb{R}$ for all $x \in A$, and in particular $f(x) \in T(x)$ for $x \in A$. 👍

XIX.6.1.15. REMARKS. (i) These results apply in the case $X = \mathbb{R}^n$ (or any subset) and $\mu = \lambda^{(n)}$, and in the case where μ is s-dimensional Hausdorff measure $\mathscr{H}^s$ on $\mathbb{R}^n$ and X is a set with $\mathscr{H}^s(X) < \infty$.

(ii) $\mathbb{R}^n$ with the Euclidean metric can be replaced by any finite-dimensional normed vector space.

(iii) Any separable metrizable topological space (any second countable regular space) X of finite topological dimension can be embedded homeomorphically in a Euclidean space (). But the statement of XIX.6.1.5. requires use of a specific metric. The relative metric from any embedding of X into a finite-dimensional normed vector space will do. However, it is unclear whether an arbitrary metric on X giving its topology will work (to get a version of the Besicovitch Covering Theorem as in XVII.8.3.1., the metric must be "controlled" in a certain sense).

(iv) If X is not finite-dimensional, the statement is probably false in general: the examples in [**?**] likely give, or lead to, a counterexample.

XIX.6.2. Lebesgue Density in $\mathbb{R}^n$

The Lebesgue Density Theorem (XIX.5.7.3.) is more delicate in $\mathbb{R}^n$ for $n > 1$, but is an immediate consequence of the results of the previous section. We work more generally with a neighborhood-finite Borel measure μ on a separable controlled metric space (X, ρ); we assume for simplicity that $\mu(B) > 0$ for every open ball B.

XIX.6.2.1. DEFINITION. Let A be a μ^*-measurable subset of X and $\boldsymbol{c} \in X$ (not necessarily in A). The *μ-upper density* of A at $\boldsymbol{c}$ is

$$\overline{d}(A, \boldsymbol{c}) = \limsup_{\delta \to 0} \frac{\mu(A \cap B_\delta(\boldsymbol{c}))}{\mu(B_\delta(\boldsymbol{c}))} \ .$$

The *μ-lower density* of A at $\boldsymbol{c}$ is

$$\underline{d}(A, \boldsymbol{c}) = \liminf_{\delta \to 0} \frac{\mu(A \cap B_\delta(\boldsymbol{c}))}{\mu(B_\delta(\boldsymbol{c}))} \ .$$

If $\overline{d}(A, \boldsymbol{c}) = \underline{d}(A, \boldsymbol{c})$, i.e. if

$$\lim_{\delta \to 0} \frac{\mu(A \cap B_\delta(\boldsymbol{c}))}{\mu(B_\delta(\boldsymbol{c}))}$$

exists, it is called the *μ-density* of A at $\boldsymbol{c}$, denoted $d(A, \boldsymbol{c})$.

XIX.6.2.2. The $\bar{d}$, etc., depends on μ and should really be written $\bar{d}_\mu$, etc. But the overwhelmingly important case is when μ is Lebesgue measure on $\mathbb{R}^n$, so we generally omit the μ; we even just say "density" instead of "μ-density" when μ is understood (e.g. Lebesgue measure).

We have $0 \le \underline{d}(A, \boldsymbol{c}) \le \overline{d}(A, \boldsymbol{c}) \le 1$ for any A and $\boldsymbol{c}$. Density is monotone: if $A \subseteq B$, then $\overline{d}(A, \boldsymbol{c}) \le \overline{d}(B, \boldsymbol{c})$ and $\underline{d}(A, \boldsymbol{c}) \le \underline{d}(B, \boldsymbol{c})$.

It is not obvious that $d(A, \boldsymbol{c})$ is defined for any $\boldsymbol{c}$ for general A. But we have:

XIX.6.2.3. THEOREM. [LEBESGUE DENSITY THEOREM, GENERAL VERSION] Let μ be a neighborhood-finite Borel measure on a separable controlled metric space (X, ρ), and let A be a μ^*-measurable subset of X. Then the density of A exists and is 1 for almost all points of A, and exists and is 0 for almost all points of A^c.

PROOF: This is an immediate consequence of XIX.6.1.13., taking $f = \chi_A$. 👍

XIX.6.2.4. COROLLARY. [LEBESGUE DENSITY THEOREM IN $\mathbb{R}^n$] Let A be a Lebesgue measurable subset of $\mathbb{R}^n$. Then the density of A exists and is 1 for almost all points of A, and exists and is 0 for almost all points of A^c.

An obvious but useful observation:

XIX.6.2.5. PROPOSITION. Let A and B be μ^*-measurable subsets of X. If $\mu(A \triangle B) = 0$, then $\overline{d}(A, \boldsymbol{c}) = \overline{d}(B, \boldsymbol{c})$ and $\underline{d}(A, \boldsymbol{c}) = \underline{d}(B, \boldsymbol{c})$ for any $\boldsymbol{c}$. Thus the densities of A and B coincide at any point where one (hence the other) is defined.

In particular, if A is μ^*-measurable, there is a Borel set B with the same density at every point.

There is no simple relation for the density of a union or intersection. But we have:

XIX.6.2.6. PROPOSITION. Let A and B be μ^*-measurable subsets of X, and $\boldsymbol{c} \in X$. Then

$$\overline{d}(A \cup B, \boldsymbol{c}) \leq \overline{d}(A, \boldsymbol{c}) + \overline{d}(B, \boldsymbol{c}) \ .$$

If A and B are disjoint (or, more generally, if $\mu(A \cap B) = 0$), and $d(A, c)$ and $d(B, c)$ exist, then $d(A \cup B, \boldsymbol{c})$ exists and equals $d(A, \boldsymbol{c}) + d(B, \boldsymbol{c})$.

Sets of density 0 will be particularly important. The density $d(A, \boldsymbol{c})$ exists and equals 0 if and only if $\overline{d}(A, \boldsymbol{c}) = 0$. Of course $d(A, \boldsymbol{c}) = 0$ if $\boldsymbol{c} \notin \bar{A}$, but this is far from necessary.

XIX.6.2.7. COROLLARY. If $d(A, \boldsymbol{c}) = d(B, \boldsymbol{c}) = 0$, then $d(A \cup B, \boldsymbol{c}) = 0$.

XIX.6.2.6. and XIX.6.2.7. obviously extend to finite unions (but not countable unions in general: in $\mathbb{R}$, let $A_n = [1/n, 1]$ and consider $d(A_n, 0)$).

XIX.6.3. Approximate Limits

We continue to work in the setting where μ is a neighborhood-finite Borel measure on a separable controlled metric space (e.g. Lebesgue measure on $\mathbb{R}^n$ with the usual metric).

XIX.6.3.1. DEFINITION. Let $A \subseteq X$, $f : A \to \mathbb{R}$ a function, and $L \in \mathbb{R}$. Then the *approximate limit* of f at $\boldsymbol{c} \in X$ is L if $d(A, \boldsymbol{c}) > 0$ and, for every $\epsilon > 0$. the set

$$A_\epsilon = \{x \in A : |f(\boldsymbol{x}) - L| \geq \epsilon\}$$

has density 0 at $\boldsymbol{c}$. We write

$$L = ap \lim_{\boldsymbol{x} \to \boldsymbol{c}} f(\boldsymbol{x}) = L$$

if L is the approximate limit of f at $\boldsymbol{c}$.

The approximate limit of course need not exist. For the approximate limit to be defined, $\boldsymbol{c}$ must be in $\bar{D}$ (but this is not sufficient). The definition is most commonly used if D is a neighborhood or deleted neighborhood of $\boldsymbol{c}$.

XIX.6.3.2. EXAMPLE. (i) Let $f : \mathbb{R} \to \mathbb{R}$ be

$$f(x) = \begin{cases} 1 & \text{if} \quad x \geq 0 \\ 0 & \text{if} \quad x < 0 \end{cases} .$$

Then $ap \lim_{x \to c} f(x) = 1$ if $c > 0$, $ap \lim_{x \to c} f(x) = 0$ if $c < 0$, and $ap \lim_{x \to 0} f(x)$ does not exist.
(ii) The Lebesgue Density Theorem can be rephrased: if $A \subseteq X$ is μ^*-measurable, then $ap \lim_{\boldsymbol{x} \to \boldsymbol{c}} \chi_A(x)$ exists and equals 1 for almost all $\boldsymbol{c} \in A$, and exists and is 0 for almost all $\boldsymbol{c} \in A^c$.

We have said "the" approximate limit. This is justified by the next observation (this is where $d(D, \boldsymbol{c}) > 0$ is needed):

XIX.6.3.3. PROPOSITION. Let $A \subseteq X$, $f : A \to \mathbb{R}$, and $\boldsymbol{c} \in X$ with $d(A, \boldsymbol{c}) > 0$. If $ap \lim_{\boldsymbol{x} \to \boldsymbol{c}} f(\boldsymbol{x}) = L$ and $ap \lim_{\boldsymbol{x} \to \boldsymbol{c}} f(\boldsymbol{x}) = L'$, then $L = L'$.

PROOF: If $L \neq L'$ and $\epsilon < |L - L'|$, $A_\epsilon = \{x \in A : |f(\boldsymbol{x}) - L| \geq \epsilon\}$, and $B_\epsilon = \{x \in A : |f(\boldsymbol{x}) - L'| \geq \epsilon\}$, then $A_\epsilon \cup B_\epsilon = A$. Apply XIX.6.2.7. to obtain a contradiction. 👍

The definition is consistent with the usual limit definition:

XIX.6.3.4. PROPOSITION. Let $A \subseteq X$, $f : A \to \mathbb{R}$, and $\boldsymbol{c} \in X$ with $d(A, \boldsymbol{c}) > 0$. If $\lim_{\boldsymbol{x} \to \boldsymbol{c}} f(x)$ exists and equals L, then $ap \lim_{\boldsymbol{x} \to \boldsymbol{c}} f(x)$ exists and equals L.

PROOF: For any $\epsilon > 0$, a deleted neighborhood of $\boldsymbol{c}$ is contained in A_ϵ^c. 👍

XIX.6.3.5. Approximate limits are much more general than (ordinary) limits: in $\mathbb{R}$, $\lim_{x \to c} \chi_{\mathbb{Q}}(x)$ does not exist anywhere, but $ap \lim_{x \to c} \chi_{\mathbb{Q}}(x)$ exists (and is 0) for every c. But there is a useful rephrasing of the definition:

XIX.6.3.6. PROPOSITION. Let $A \subseteq \mathbb{R}^n$, $f : A \to \mathbb{R}$, and $\boldsymbol{c} \in \mathbb{R}^n$ with $d(A, \boldsymbol{c}) > 0$. Then $ap\lim_{\boldsymbol{x}\to\boldsymbol{c}} f(x)$ exists and equals L if and only if there is a $B \subseteq A$ with $d(A \setminus B, \boldsymbol{c}) = 0$ such that $\lim_{\boldsymbol{x}\to\boldsymbol{c}} f|_B(x)$ exists and equals L.

XIX.6.4. Approximate Continuity

The main goal of this section is to make precise and justify the statement that "a Lebesgue measurable function is almost continuous almost everywhere."

The next definition is obvious:

XIX.6.4.1. DEFINITION. Let (X, ρ) be a metric space, μ a neighborhood-finite Borel measure on X, $A \subseteq X$, and $f : A \to \mathbb{R}$ a function. Then f is *approximately continuous* at $a \in A$ if $d(A, a) > 0$ and $ap\lim_{x\to a} f(x)$ exists and equals $f(a)$.

Luzin's Theorem

We begin with a theorem of LUZIN (who considered only the case of $\mathbb{R}$ with Lebesgue measure):

XIX.6.4.2. THEOREM. [LUZIN'S THEOREM] Let X be a metrizable space, and μ a locally finite Borel measure on X. Let f be a real-valued μ^*-measurable function on X, and $A \subseteq X$ a μ^*-measurable set with $\mu(A) < \infty$. Then for any $\epsilon > 0$ there is a closed set $E \subseteq A$ with $\mu(A \setminus E) < \epsilon$ and $f|_E$ continuous on E. If X is Polish (e.g. $\mathbb{R}^n$), then E can be chosen compact.

PROOF: Let $\epsilon > 0$. For each $n \in \mathbb{N}$ and $k \in \mathbb{Z}$ let $I_{n,k} = \left[\frac{k}{n}, \frac{k+1}{n}\right)$ and set

$$A_{n,k} = A \cap f^{-1}(I_{n,k}) \ .$$

For fixed n the sets $A_{n,k}$ are pairwise disjoint μ^*-measurable sets of finite measure. Thus by () for each n and k there is a closed set $E_{n,k} \subseteq A_{n,k}$ with

$$\mu(A_{n,k} \setminus E_{n,k}) < 2^{-n-|k|-1}\epsilon \ .$$

For each n we have that

$$A \setminus \bigcup_{k\in\mathbb{Z}} E_{n,k} = \bigcup_{k\in\mathbb{Z}} (A_{n,k} \setminus E_{n,k})$$

(disjoint union), so

$$\mu\left(A \setminus \bigcup_{k\in\mathbb{Z}} E_{n,k}\right) = \sum_{k\in\mathbb{Z}} \mu(A_{n,k} \setminus E_{n,k}) < \sum_{k\in\mathbb{Z}} 2^{-n-|k|-1}\epsilon < 2^{-n}\epsilon \ .$$

Thus there is an $m(n)$ such that, if

$$E_n = \bigcup_{k=-m(n)}^{m(n)} E_{n,k}$$

we have

$$\mu(A \setminus E_n) < 2^{-n}\epsilon$$

and the set E_n is closed.

Define a continuous function g_n on E_n by setting $g_n(x) = \frac{k}{n}$ if $x \in E_{n,k}$ ($|k| \leq m(n)$). We have $|f(x) - g_n(x)| < \frac{1}{n}$ for $x \in E_n$.

Set $E = \cap_n E_n$. Then E is closed, and

$$\mu(A \setminus E) \leq \sum_{n=1}^{\infty} \mu(A \setminus E_n) < \sum_{n=1}^{\infty} 2^{-n}\epsilon = \epsilon \ .$$

The continuous functions g_n converge uniformly to $f|_E$ on E, hence $f|_E$ is continuous on E.

For the last statement, if X is Polish, find E as above with $\mu(A \setminus E) < \epsilon/2$; then by XVII.6.1.22. there is a compact $K \subseteq E$ with $\mu(E \setminus K) < \epsilon/2$, so $\mu(A \setminus K) < \epsilon$ and $f|_K$ is continuous on K. 👍

XIX.6.4.3. **Caution:** The theorem asserts that the *restriction* of f to E is continuous, not that f is continuous at points of E. In fact a measurable function need not be continuous anywhere, such as $\chi_{\mathbb{Q}}$ (in this case the E would be a subset of the irrational numbers). ✋

Approximate Continuity of Measurable Functions

XIX.6.4.4. THEOREM. Let (X, ρ) be a separable controlled metric space and μ a neighborhood-finite Borel measure on X. If $f : X \to \mathbb{R}$ is μ^*-measurable, then f is approximately continuous μ-almost everywhere on X.

PROOF: By XIX.6.4.2. there is a sequence (E_n) of closed sets in X with $\mu(X \setminus \cup E_n) = 0$ and $f|_{E_n}$ continuous on E_n. By XIX.6.2.3. we have $d(X \setminus E_n, x) = 0$ for almost all $x \in E_n$, so f is approximately continuous at almost all points of E_n for each n. 👍

XIX.6.4.5. Since the conclusion does not depend on the metric, the result holds for any separable metrizable space which is topologically finite-dimensional (XVII.8.1.15.).

XIX.6.5. Approximate Differentiability

In this section, we consider only $\mathbb{R}^n$ with Lebesgue measure $\lambda = \lambda^{(n)}$.

XIX.6.5.1. DEFINITION. Let A be a subset of $\mathbb{R}^n$, $\boldsymbol{a} \in A$ with $d(A, \boldsymbol{a}) > 0$. A function $\boldsymbol{f} : A \to \mathbb{R}^m$ is *approximately differentiable* at $\boldsymbol{a}$ if there is a linear map $L : \mathbb{R}^n \to \mathbb{R}^m$ with

$$ap \lim_{\boldsymbol{x} \to \boldsymbol{a}} \frac{\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{a}) - L(\boldsymbol{x} - \boldsymbol{a})\|}{\|x - a\|} = 0 \ .$$

The linear map L, if it exists, is the *approximate derivative* of $\boldsymbol{f}$ at $\boldsymbol{a}$, denoted *ap-*$d\boldsymbol{f}(\boldsymbol{a})$ and its matrix denoted *ap-*$D\boldsymbol{f}(\boldsymbol{a})$.

XIX.6.5.2. The L, if it exists, is unique, just as in the differentiable case () (since $d(A, \boldsymbol{a}) > 0$). If $\boldsymbol{f} = (f_1, \ldots, f_m)$ with $f_k : A \to \mathbb{R}$, then $\boldsymbol{f}$ is approximately differentiable at $\boldsymbol{a}$ if and only if each f_k is approximately differentiable at $\boldsymbol{a}$, and the linear functions $ap\text{-}df_k$ are the coordinate functions of $ap\text{-}d\boldsymbol{f}$. Thus in analyzing approximate differentiability we may assume $m = 1$.

XIX.6.5.3. PROPOSITION. Let A be a subset of $\mathbb{R}^n$, $\boldsymbol{a} \in A$ with $d(A, \boldsymbol{a}) > 0$, $\boldsymbol{f} : A \to \mathbb{R}^m$. If $\boldsymbol{f}$ is approximately differentiable at $\boldsymbol{a}$, then $\boldsymbol{f}$ is approximately continuous at $\boldsymbol{a}$.

There is a companion definition:

XIX.6.5.4. DEFINITION. Let A be a subset of $\mathbb{R}^n$, $\boldsymbol{a} = (a_1, \ldots, a_n) \in A$, and $\boldsymbol{f} : A \to \mathbb{R}$ a function. Fix k, $1 \leq k \leq n$. If the intersection

$$A_k = A \cap \{(a_1, \ldots, a_{k-1}, t, a_{k+1}, \ldots, a_n) : t \in \mathbb{R}$$

has positive 1-dimensional density at $\boldsymbol{a}$ and

$$ap \lim_{t \to a_k} \frac{f(a_1, \ldots, a_{k-1}, t, a_{k+1}, \ldots, a_n) - f(a_1, \ldots, a_{k-1}, a_k, a_{k+1}, \ldots, a_n)}{t}$$

exists, it is called the *approximate partial derivative* of f at $\boldsymbol{a}$ with respect to x_k, denoted $ap\frac{\partial f}{\partial x_k}$.

XIX.6.5.5. If f is approximately differentiable at $\boldsymbol{a}$, the domain A of f need not have any point other than $\boldsymbol{a}$ itself on lines in coordinate directions through $\boldsymbol{a}$ (although A must contain points near $\boldsymbol{a}$ very close to such a coordinate line by density), so the approximate partial derivatives of f at $\boldsymbol{a}$ do not necessarily exist and the geometric interpretation of the entries of the matrix entries of $ap\text{-}Df(\boldsymbol{a})$ is a little obscure. But if the 1-dimensional densities of the intersection of A with lines in the coordinate directions through $\boldsymbol{a}$ are positive, the approximate partial derivatives exist at $\boldsymbol{a}$ and they are the entries of $ap\text{-}Df(\boldsymbol{a})$. Thus, modulo delicate questions about domain, approximate differentiability implies existence of approximate partial derivatives in each coordinate direction.

The converse is false in general, just as in the exact case: existence of approximate partial derivatives is not sufficient for approximate differentiability. Even existence and approximate continuity of exact partial derivatives is not sufficient:

XIX.6.5.6. EXAMPLE. Theorem VIII.3.3.1. does not hold if continuity and differentiability are replaced by approximate continuity and approximate differentiability. In $\mathbb{R}^2$, let L_1 and L_2 be rays from $\mathbf{0}$ in the open first quadrant with L_2 above L_1, and C_1 and C_2 curves in the open first quadrant below L_1 and above L_2 respectively and asymptotic to the rays approaching $\mathbf{0}$ (see Figure ()). Let $f : \mathbb{R}^2 \to [0, 1]$ be a function which is 1 at points of the open first quadrant between L_1 and L_2, zero on points of the first quadrant below C_1 or above C_2 and on the other three (closed) quadrants, and $\mathcal{C}^\infty$ on $\mathbb{R}^2 \setminus \{\mathbf{0}\}$. Then the first-order partials of f exist everywhere and are continuous everywhere except at $\mathbf{0}$. If C_1 and C_2 approach L_1 and L_2 rapidly enough, the first-order partials are approximately continuous at $\mathbf{0}$ since they are identically 0 except between C_1 and L_1 and between L_2 and C_2. But f is not even approximately continuous at $\mathbf{0}$, hence not approximately differentiable at $\mathbf{0}$.

XIX.7. Lipschitz Functions

Lipschitz functions are especially important on Euclidean spaces, where they are a generalization of smooth functions with many of the same properties; but the definition makes sense for functions between general metric spaces

XIX.7.1. Lipschitz Functions Between Metric Spaces

XIX.7.1.1. DEFINITION. Let (X, ρ) and (Y, σ) be metric spaces. A function $f : X \to Y$ is *Lipschitz* if there is a constant C such that $\sigma(f(x_1), f(x_2)) \leq C\rho(x_1, x_2)$ for all $x_1, x_2 \in X$. The smallest such C is called the *Lipschitz constant* of f, denoted $Lip(f)$.

XIX.7.1.2. Note that this is strictly a metric notion, and depends on the choice of metrics (which should technically be incorporated into the notation and terminology, although it is conventional not to since generally only Lipschitz functions between subsets of Euclidean spaces are considered).

XIX.7.1.3. PROPOSITION. A Lipschitz function is uniformly continuous (and in particular continuous).

PROOF: Let $f : (X, \rho) \to (Y, \sigma)$ be Lipschitz, with constant C; we may assume $C > 0$. Let $\epsilon > 0$, and set $\delta = \frac{\epsilon}{C}$. Then if $\rho(x_1, x_2) < \delta$, then $\sigma(f(x_1), f(x_2)) < \epsilon$. 👍

XIX.7.2. Locally Lipschitz Functions

DEFINITION. Let (X, ρ) and (Y, σ) be metric spaces. A function $f : X \to Y$ is *locally Lipschitz (LL)* if, for each $a \in X$, there is an open ball B around a such that $f|_B$ is a Lipschitz function from B to Y. In other words, for every $a \in X$ there is an $\epsilon > 0$ and $C \geq 0$ such that $\sigma(f(x_1), f(x_2)) \leq C\rho(x_1, x_2)$ for any $x_1, x_2 \in X$ with $\rho(x_1, a), \rho(x_2, a) < \epsilon$.

An LL function is locally uniformly continuous, hence continuous.

XIX.7.2.1. EXAMPLE. In this definition it is not enough that $\sigma(f(x), f(a)) \leq C\rho(x, a)$ whenever $\rho(x, a) < \epsilon$. Let $X \subseteq \mathbb{R}^2$ consist of $\mathbf{0}$ and two sequences (p_n), (q_n), where

$$p_n = \left(\frac{1}{n}, \frac{1}{n^2}\right) \quad , \quad q_n = \left(\frac{1}{n}, -\frac{1}{n^2}\right) .$$

Define $f : X \to \mathbb{R}$ by $f(\mathbf{0}) = 0$, $f(p_n) = \frac{1}{n}$, and $f(q_n) = -\frac{1}{n}$ for all n. Then f is continuous,

$$|f(p_n) - f(\mathbf{0})| = |f(q_n) - f(\mathbf{0})| < \|p_n - \mathbf{0}\| = \|q_n - \mathbf{0}\|$$

for all n, but f is not Lipschitz on any neighborhood of $\mathbf{0}$ since $|f(p_n) - f(q_n)| = n\|p_n - q_n\|$ for all n.

XIX.7.2.2. PROPOSITION. Let (X, ρ) and (Y, σ) be metric spaces. A function $f : X \to Y$ is LL if and only if its restriction to every compact subset of X is Lipschitz.

PROOF: Suppose f is not LL. Then there is an $a \in X$ and sequences (p_n), (q_n) in X with $\rho(p_n, a), \rho(q_n, a) < \frac{1}{n}$ and $\sigma(f(p_n), f(q_n)) > n\rho(p_n, q_n)$ for each n. Then $K = \{p_n, q_n : n \in \mathbb{N}\} \cup \{a\}$ is a compact subset of X, and $f|_K$ is not Lipschitz.

Conversely, suppose f is LL, and let K be a compact subset of X. Since f is continuous, the function $g(x_1, x_2) = \sigma(f(x_1), f(x_2))$ is a continuous function from $X \times X$ to $\mathbb{R}$, hence is bounded on the compact set $K \times K$, say by M.

For each $a \in K$, there is an open ball B_a around a such that $f|_{B_a}$ is Lipschitz. The B_a cover K, so there is a finite subcover $\{B_{a_1}, \ldots, B_{a_n}\}$. Let C_i be the Lipschitz constant of $f|_{B_{a_i}}$, and $C = \max(C_1, \ldots, C_n)$.

Let δ be a Lebesgue number () for the open cover $\{B_{a_1}, \ldots, B_{a_n}\}$ of K, and set $N = \max\left(C, \frac{M}{\delta}\right)$. Let $x_1, x_2 \in K$. If $\rho(x_1, x_2) < \delta$, then x_1 and x_2 are in the same B_{a_i} and hence

$$\sigma(f(x_1), f(x_2)) \leq C_i \rho(x_1, x_2) \leq N\rho(x_1, x_2) \ .$$

And if $\rho(x_1, x_2) \geq \delta$, then

$$\frac{\sigma(f(x_1), f(x_2))}{\rho(x_1, x_2)} \leq \frac{\sigma(f(x_1), f(x_2))}{\delta} \leq \frac{M}{\delta} \leq N \ .$$

XIX.7.3. Lipschitz Functions Between Euclidean Spaces

We now specialize to the case of functions from $\mathbb{R}^n$, or an open set in $\mathbb{R}^n$, to $\mathbb{R}^m$. We will always use the standard metric on a Euclidean space, or any subset.

XIX.7.3.1. PROPOSITION. Let U be a subset of $\mathbb{R}^n$, and $\boldsymbol{f} : U \to \mathbb{R}^m$ a function. Then $\boldsymbol{f}$ is Lipschitz (resp. LL) if and only if each coordinate function is a Lipschitz (resp. LL) function from U to $\mathbb{R}$.

XIX.7.3.2. If $[a, b]$ is a closed bounded interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$ is a function, then by () f is Lipschitz if and only if it is absolutely continuous and f' (defined a.e.) is bounded. Thus, if I is an interval in $\mathbb{R}$ and $f : I \to \mathbb{R}$, then f is LL if and only if f is locally absolutely continuous (absolutely continuous on each closed bounded subinterval of I) and f' is locally bounded. In particular, if f is $\mathcal{C}^1$ on I, then f is LL on I. Thus LL functions are slightly more special than locally absolutely continuous functions, and generalize smooth functions. By considering coordinate functions, the same statements apply to functions from I to $\mathbb{R}^m$ for any m.

XIX.7.3.3. There is no natural analog of absolute continuity for real-valued functions on a rectangle in $\mathbb{R}^n$ for $n > 1$. But Lipschitz makes sense and is a good substitute. So if $U \subseteq \mathbb{R}^n$ is open, it makes sense to consider LL functions from U to $\mathbb{R}$, or to $\mathbb{R}^m$ for any m.

XIX.7.3.4. PROPOSITION. Let $U \subseteq \mathbb{R}^n$ be open, and $\boldsymbol{f} : U \to \mathbb{R}^m$ be a $\mathcal{C}^1$ function. Then $\boldsymbol{f}$ is LL on U.

PROOF: Let $\boldsymbol{a} \in U$, and B an open ball around $\boldsymbol{a}$ whose closure is contained in U. Then all partial derivatives of $\boldsymbol{f}$ are bounded on B. Set

$$M = \sum_{i,j} \sup_{\boldsymbol{x} \in B} \left| \frac{\partial f_i}{\partial x_j}(\boldsymbol{x}) \right| .$$

If $\boldsymbol{x}, \boldsymbol{y} \in B$, set $\boldsymbol{g}(t) = \boldsymbol{f}(t\boldsymbol{x} + (1-t)\boldsymbol{y})$. Then $\boldsymbol{g}$ is $\mathcal{C}^1$ on $[0,1]$, and $\|\boldsymbol{g}'(t)\| \leq M\|\boldsymbol{x} - \boldsymbol{y}\|$ a.e. by the Chain Rule, so $\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| = \|\boldsymbol{g}(1) - \boldsymbol{g}(0)\| \leq M\|\boldsymbol{x} - \boldsymbol{y}\|$. 👍

The next key result extending () is a partial converse.

XIX.7.3.5. THEOREM. [RADEMACHER'S THEOREM] Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ a Lipschitz function from U to $\mathbb{R}^m$. Then $\boldsymbol{f}$ is differentiable a.e., and $\|D\boldsymbol{f}(\boldsymbol{a})\| \leq Lip(\boldsymbol{f})$ for all $\boldsymbol{a}$ for which it is defined (i.e. for all $\boldsymbol{a}$ where $\boldsymbol{f}$ is differentiable).

XIX.7.3.6. COROLLARY. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ an LL function from U to $\mathbb{R}^m$. Then $\boldsymbol{f}$ is differentiable a.e., and $D\boldsymbol{f}$ is locally bounded.

The proof of Rademacher's Theorem uses several lemmas.

XIX.7.3.7. LEMMA. Let U be an open set in $\mathbb{R}^n$, and $\boldsymbol{f}$ a Lipschitz function from U to $\mathbb{R}^m$. Then, for each i, $\frac{\partial \boldsymbol{f}}{\partial x_i}$ exists a.e. on U with respect to n-dimensional Lebesgue measure; in fact, on any line segment within U parallel to $\boldsymbol{e}_i$, $\frac{\partial \boldsymbol{f}}{\partial x_i}$ exists a.e. with respect to 1-dimensional Lebesgue measure (1-dimensional Hausdorff measure on $\mathbb{R}^n$). $\left\| \frac{\partial \boldsymbol{f}}{\partial x_i}(\boldsymbol{a}) \right\| \leq Lip(\boldsymbol{f})$ whenever it is defined.

PROOF: We may assume $m = 1$ by considering coordinate functions (so we write f instead of $\boldsymbol{f}$). Let $\boldsymbol{a}_0 \in U$, and set $g(t) = f(\boldsymbol{a}_0 + t\boldsymbol{e}_i)$. Then g is a Lipschitz function on the largest open interval I around 0 whose image is contained in U, whose constant is less than or equal to $Lip(f)$, and hence is differentiable a.e. in I with $|g'(t)| \leq Lip(f)$ wherever defined. We have

$$\frac{\partial f}{\partial x_i}(\boldsymbol{a}_0 + t\boldsymbol{e}_i) = g'(t)$$

for any t where the right side is defined. It then follows from the Junior Tonelli Theorem that $\frac{\partial f}{\partial x_i}$ exists a.e. on U with respect to n-dimensional Lebesgue measure (the set on which it is defined is a Borel set, cf. ()). 👍

XIX.7.3.8. More generally, if L is any line in $\mathbb{R}^n$ and I is any line segment contained in U parallel to L, then the directional derivative of $\boldsymbol{f}$ in the direction L is defined a.e. on I with respect to 1-dimensional Hausdorff measure, with the same bound $Lip(\boldsymbol{f})$. With care, this can be extended to any smooth curve in U.

XIX.7.4. Hölder Continuity

Chapter XX

Functional Analysis

Functional analysis is, in a nutshell, the study of topological vector spaces and operators on them. The name arose in part because historically (and still today), many of the most important topological vector spaces are function spaces. Another reason for the name is that one of the main underlying philosophies of functional analysis is that spaces can be studied by means of real- or complex-valued functions (*functionals*) on them.

Functional analysis combines linear algebra with analysis and topology. The interaction makes for a rich and powerful theory that has many applications in both pure and applied mathematics.

If a scientist wants to study a forest, there are two principal ways to examine it. First, one can walk through the forest. In this way, the small-scale properties can be studied in detail: the individual trees and the local ecosystems. But it is not easy to determine the large-scale structure of the forest this way, its overall extent and shape, the relative distribution of tree species and open spaces, etc. The other way is to fly over the forest (https://www.youtube.com/watch?v=EqFmLgK_H9s). If you do this, the small-scale structure is not easily seen: individual trees are almost reduced to points in the overall picture; but the large-scale structure can be easily seen at a glance. A full study of the forest would undoubtedly include both methods.

Calculus and traditional elementary analysis give the basic tools necessary to do small-scale analysis of spaces of functions. Functional analysis provides the way to do large-scale analysis. Individual functions, although they have a rich internal structure of their own, are just points in an overall space from the functional analysis point of view. Flying over the function space using functional analysis is often a powerful way of easily establishing properties or solving problems related to function spaces.

Functional analysis reaches its full development and power in combination with measure theory, and many if not most texts assume familiarity with measure theory. But the essentials of functional analysis can be done without measure theory, so we will go as far as possible without weaving measure theory into the picture.

XX.1. Preliminaries

In this section, we introduce important results and terminology about vector spaces over the real or complex numbers. No topology will be used in this section (actually there is some, hidden in the structure of the field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$); the subject matter is entirely linear algebra. However, the applications will be (almost) all to the study of topological vector spaces.

In this section, all vector spaces will be vector spaces over $\mathbb{F}$, which could be either $\mathbb{R}$ or $\mathbb{C}$. We will specify which when it matters.

XX.1.1. Properties of Subsets

Set Arithmetic

XX.1.1.1. If A and B are subsets of a vector space X, then

$$A + B = \{x + y : x \in A, y \in B\} \ .$$

If $y \in X$, then $A + y = A + \{y\} = \{x + y : x \in A\}$.

If $\alpha \in \mathbb{F}$, define $\alpha A = \{\alpha x : x \in A\}$.

These can be combined, e.g. $\alpha A + \beta B$, $A - B$, etc.

Special Types of Sets

XX.1.1.2. Definition. Let X be a vector space over $\mathbb{F}$, and $A \subseteq X$. Then

(i) A is *balanced* if $\lambda \in \mathbb{F}$, $|\lambda| \leq 1$, implies $\lambda A \subseteq A$, i.e. $x \in A$ and $|\lambda| \leq 1$ imply $\lambda x \in A$.

(ii) A is *radial* if for every $x \in X$ there is an $M > 0$ such that $x \in \lambda A$ for all $\lambda \in \mathbb{F}$, $|\lambda| \geq M$. Equivalently, A is radial if for every $x \in X$ there is an $\epsilon > 0$ such that $\alpha x \in A$ for all $\alpha \in \mathbb{F}$, $0 < |\alpha| \leq \epsilon$.

(iii) More generally, A *absorbs* a subset B of X if there is an $M > 0$ such that $B \subseteq \lambda A$ for all $\lambda \in \mathbb{F}$, $|\lambda| \geq M$.

XX.1.1.3. Being balanced depends on the field $\mathbb{F}$ (being radial or absorbing does not). In some references the term *circled* is used to mean *balanced*. The term *balanced* seems better, especially if $\mathbb{F} = \mathbb{R}$ (not to mention that there is a mathematical distinction between a circle and a disk). A balanced set, if nonempty, always contains 0.

A set is radial if and only if it absorbs every singleton (hence every finite subset of X). A radial set always contains 0 (thus is nonempty); hence A is radial if and only if for every $x \in X$, there is an $\epsilon > 0$ such that $\alpha x \in A$ for all $\alpha \in \mathbb{F}$, $|\alpha| \leq \epsilon$.

Being balanced and being radial seem like similar concepts, but neither implies the other. For example, a proper vector subspace is balanced but not radial. And since any set containing a radial set is radial, a set can be radial but not balanced.

XX.1.1.4. Definition. Let A be a subset of a vector space A. Then x is an *internal point* of A if the set $A - x$ is radial, i.e. for every $y \in X$ there is an $\epsilon > 0$ such that $x + \alpha y \in A$ for all α, $|\alpha| \leq \epsilon$.

An internal point is an algebraic version of an interior point (note that there is no topology involved).

Balanced Hull

XX.1.1.5. Let X be a vector space. Any intersection of balanced subsets of X is balanced. Thus if $A \subseteq X$, there is a smallest balanced set B containing A, the intersection of all balanced sets containing A (there is at least one such set, X). B is called the *balanced hull* (some references call it the *circled hull*) of A, denoted $bal(A)$. We have $bal(A) = \cup_{|\alpha| \leq 1} \alpha A$.

XX.1.2. Convex Sets

Convexity is hard to categorize mathematically. It is essentially a purely algebraic concept; but it is defined only in vector spaces, or affine spaces, over the real numbers (or over a field containing the real numbers, notably $\mathbb{C}$), and the primary applications are in analysis.

XX.1.2.1. Definition. Let X be a vector space over $\mathbb{R}$. If $x, y \in X$, the *interval* between x and y is

$$[x, y] = \{\alpha x + (1 - \alpha)y : 0 \leq \alpha \leq 1\} .$$

A subset C of X is *convex* if, whenever $x, y \in C$, then $[x, y] \subseteq C$.

XX.1.2.2. Geometrically, the interval $[x, y]$ is the "straight line segment" between x and y ($[x, y] = [y, x]$; $[x, y] = \{x\}$ if $x = y$). Thus a set is convex if and only if it contains the straight line segment between any two of its points; thus it can have no "dents" or "holes."

XX.1.2.3. Examples. (i) Any singleton set is convex.

(ii) An affine subspace is convex.

(iii) In a normed vector space, any ball (open or closed) is convex by the triangle inequality.

(iv) In $\mathbb{R}^2$, any regular polygon (including the interior!) is convex. Many nonregular polygons are convex: for example, any triangle or trapezoid is convex. But not every quadrilateral is convex (the interior angles must be less than 180°). In $\mathbb{R}^3$, (solid) regular polyhedra are convex, as are many, but far from all, nonregular polyhedra. A convex set in $\mathbb{R}^2$ contained in a two-dimensional affine subspace of $\mathbb{R}^3$ is convex (more generally, the property of being convex is intrinsic and largely independent of the containing vector space, e.g. if Y is an affine subspace of X, a subset of Y is convex as a subset of X if and only if it is convex as a subset of Y.)

(v) The empty set is vacuously convex. (Not all authors call $\emptyset$ convex, but it is convenient to allow it.) ✋

XX.1.2.4. Proposition. Let X be a vector space, and A and B convex subsets of X. Then αA is convex for any $\alpha \in \mathbb{F}$, and $A + B$ is convex. Hence $A - B$, etc., are convex.

Convex Combinations

Convex Hull

XX.1.2.5. Let X be a vector space. Any intersection of convex subsets of X is convex. Thus if $A \subseteq X$, there is a smallest convex set C containing A, the intersection of all convex sets containing A (there is at least one such set, X). C is called the *convex hull* of A, denoted $co(A)$. We have that $co(A)$ is the set of all convex combinations of elements of A.

XX.1.2.6. If A is a subset of a vector space X, there is also a smallest balanced convex set containing A, called the *balanced convex hull* of A, denoted $bc(A)$. It can be shown that

$$bc(A) = co(bal(A)) = \left\{ \sum_{k=1}^{n} \alpha_k x_k \right\}$$

where n can vary, $x_1, \ldots, x_n \in A$, $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$, $\sum_{k=1}^{n} |\alpha_k| \leq 1$. (It is not $bal(co(A))$ in general, e.g. if A consists of two linearly independent points in $\mathbb{R}^2$.)

Affine Functions

XX.1.2.7. DEFINITION. Let C and D be convex sets in vector spaces X and Y respectively. An *affine map* from C to D is a function $f : C \to D$ such that

$$f(\alpha x + (1-\alpha)y) = \alpha f(x) + (1-\alpha) f(y)$$

for all $x, y \in C$ and all α, $0 \leq \alpha \leq 1$.

XX.1.2.8. An affine map from C to D is a function which preserves the convex structure. If f is an affine map from X to Y in the usual sense of (), and $f(C) \subseteq D$, then $f|_C$ is an affine map from C to D in the sense of XX.1.2.7.. It can be shown that every affine map from C to D arises in this way (), so the terminology is consistent.

XX.1.2.9. DEFINITION. Let C and D be convex sets in vector spaces X and Y respectively. An *isomorphism* from C to D is an affine bijection from C to D.

Thus an isomorphism from C to D is a bijection which preserves the convex structure.

XX.1.3. Faces and Extreme Points

The crucial notion underlying the whole analysis of convex sets is that of a face.

XX.1.3.1. DEFINITION. Let K be a convex set in a (real) vector space. A nonempty subset F of K is a *face* of K if, whenever $x, y \in K$ and $\alpha x + (1-\alpha)y \in F$ for some α, $0 < \alpha < 1$, then both x and y are in F. A one-point face is called an *extreme point* of K. More precisely, p is an extreme point of K if $\{p\}$ is a face of K. Thus an extreme point of K is a point which cannot be written as a nontrivial convex combination of two distinct points in K.

Note that this definition makes no use of any topology. However, faces and extreme points are most interesting for closed convex sets in topological vector spaces. A face is automatically convex.

In some references, faces are called *extreme subsets* or *supports*.

XX.1.3.2. K itself is a face of K; we are mainly interested in *proper faces* of K. Roughly (imprecisely, and not quite correctly) speaking, a proper face of K is a subset of the "boundary" of K.

XX.1.3.3. EXAMPLES. (i) In $\mathbb{R}$, convex subsets are intervals. If I is an interval, the only proper faces are the extreme points; the extreme points of I are the endpoints of I which are contained in I (by definition, a face of a convex set is a subset of the convex set). Thus a nondegenerate closed bounded interval has two extreme points, an open interval has no extreme points, and other intervals have one extreme point.

(ii) Let K be a closed convex (solid) polygon in $\mathbb{R}^2$, e.g. a (solid) triangle or rectangle. Each closed edge of K is a face of K, and each vertex is an extreme point, and these are all the proper faces.

(iii) Let K be a closed convex (solid) polyhedron in $\mathbb{R}^3$, e.g. a tetrahedron or cube. There are two-dimensional faces, the closed bounding (solid) polygons (which are usually called the "faces" of the polyhedron); one-dimensional faces, the closed edges; and 0-dimensional faces (extreme points), the vertices. These are all the proper faces. This example is undoubtedly the origin of the term "face" for general convex sets.

(iv) Let K be the closed unit ball in $\mathbb{R}^n$. Then the only proper faces of K are the extreme points, which are precisely all the points on the unit sphere. Thus if $n > 1$ there are uncountably many extreme points.

(v) An open ball in $\mathbb{R}^n$ has no extreme points. A nontrivial subspace of $\mathbb{R}^n$ has no extreme points.

XX.1.3.4. PROPOSITION. Let K be a convex set in a vector space, and F a nonempty subset of K. The following are equivalent:

(i) Whenever $x, y \in K$ and $\frac{1}{2}(x+y) \in F$, then $x, y \in F$.

(ii) Whenever $x_1, \ldots, x_n \in K$, $0 < \alpha_1, \ldots, \alpha_n < 1$, $\sum_{k=1}^n \alpha_k = 1$, and $\sum_{k=1}^n \alpha_k x_k \in F$, then $x_1, \ldots, x_n$ are all in F.

(iii) F is a face of K.

In particular, F is automatically convex.

PROOF: (ii) $\Rightarrow$ (iii) $\Rightarrow$ (i) are trivial.

(i) $\Rightarrow$ (iii): Suppose $x, y \in K$, $0 < \alpha < 1$, $z_0 := \alpha x + (1-\alpha)y \in F$. We may assume without loss of generality that $\alpha \leq \frac{1}{2}$. Set $z_1 = 2\alpha x + (1-2\alpha)y$. Then $z_1 \in K$ and $z_0 = \frac{1}{2}(y + z_1) \in F$ and thus $y, z_1 \in F$. If $2\alpha \geq \frac{1}{2}$, repeat the argument with x and y reversed and α replaced by 2α to conclude that $x \in F$. If $2\alpha < \frac{1}{2}$, set $z_2 = 4\alpha x + (1-4\alpha)y \in K$; then $z_1 = \frac{1}{2}(y + z_2)$, so $z_2 \in F$. Repeat the argument as many times as necessary; after a finite number of steps (at most $-\log_2 \alpha$) we obtain $x \in F$.

(iii) $\Rightarrow$ (ii): Suppose $x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_n$ are as in (ii), and set

$$z = \sum_{k=1}^{n} \alpha_k x_k \ .$$

For $1 \leq k \leq n-1$ set $\beta_k = \frac{\alpha_k}{1-\alpha_n}$. Then $\sum_{k=1}^{n-1} \beta_k = 1$. Set

$$y = \sum_{k=1}^{n-1} \beta_k x_k \in K$$

so that $z = \alpha_n x_n + (1-\alpha_n)y$. Since K is a face, x_n and y are in F. Repeat the argument with y in place of z, and iterate $n-1$ times (or prove by induction on n). 👍

XX.1.3.5. Corollary. Let K be a convex set in a vector space, and $z \in K$. The following are equivalent:

(i) Whenever $x, y \in K$ and $\frac{1}{2}(x+y) = z$, then $x = y = z$.

(ii) Whenever $x_1, \ldots, x_n \in K$, $0 < \alpha_1, \ldots, \alpha_n < 1$, $\sum_{k=1}^n \alpha_k = 1$, and $\sum_{k=1}^n \alpha_k x_k = z$, then $x_k = z$ for all k.

(iii) z is an extreme point of K.

XX.1.3.6. Proposition. Let K be a convex set in a vector space.

(i) Let $\{F_i : i \in I\}$ be a collection of faces of K. If $\cap_i F_i \neq \emptyset$, then it is a face of K.

(ii) A face of a face of K is a face of K.

(iii) If C is a convex subset of K, F is a face of K, and $F \cap C \neq \emptyset$, then $F \cap C$ is a face of C.

Proof: Straightforward and left to the reader. 👍

XX.1.3.7. Thus, if S is any nonempty subset of K, there is a smallest face $F(S)$ containing S, called the *face generated by* S. $F(S)$ is the intersection of all faces containing S.

We can define an *inner point* of K to be any point p of K such that $F(\{p\}) = K$. Note that such a p is not necessarily an interior point in the topological sense in general (indeed, there may be no topology in the picture!); but if K is a convex set in a topological vector space which has a nonempty interior, then points in the topological interior are inner points (XX.9.1.3.).

A nonempty convex need not have any inner points (). But if it has one, it has a lot:

XX.1.3.8. Proposition. Let C be a convex set in a vector space X. If x is an inner point of C, then for any $y \in C$ the set $[x, y) = \{\alpha x + (1-\alpha)y : 0 < \alpha \leq 1\}$ consists entirely of inner points.

Images and Preimages

XX.1.3.9. Proposition. Let X and Y be vector spaces, and $T : X \to Y$ a linear transformation. If K is a convex set in Y and F is a face of K, then $T^{-1}(F)$ is a face in $T^{-1}(K)$.

The simple proof is again left to the reader.

XX.1.3.10. The same result does not hold for images in general: if $T : X \to Y$ is linear, K is a convex set in X, and F is a face in K, then $T(F)$ is not necessarily a face in $T(K)$ (consider the case where Y is one-dimensional). ✋

XX.1.3.11. A particularly important case of XX.1.3.9. is when $Y = \mathbb{R}$. If ϕ is a linear functional on X, K is a convex subset of X, $I = \phi(K)$ (an interval in $\mathbb{R}$), and c is an endpoint of I contained in I, then the intersection of K with the hyperplane $\phi^{-1}(\{c\})$ is a face of K (cf. XX.1.3.6.(iii)). In other words, if $\phi(x) \geq c$ (or $\phi(x) \leq c$) for all $x \in K$, and there is an $x \in K$ with $\phi(x) = c$, then the set

$$\{x \in K : \phi(x) = c\}$$

is a face of K. In this case, the hyperplane $\phi^{-1}(\{c\})$ is called a *support hyperplane* for K, and the intersection with K a *support face*. Such a support face is frequently, although far from always, an extreme point of K.

XX.1.4. Minkowski Functionals

XX.1.4.1. Definition. Let X be a vector space, and A a subset of X. The *Minkowski functional*, or *gauge*, of A is the function $m_A : X \to [0, \infty]$ defined by

$$m_A(x) = \inf\{\alpha > 0 : x \in \alpha A\} = \inf\left\{\alpha > 0 : \frac{1}{\alpha}x \in A\right\}$$

(where $\inf \emptyset = \infty$).

XX.1.4.2. Proposition. Let A be a subset of a vector space X, and m_A its Minkowski functional. Then

(i) If $x \in X$, then $m_A(\alpha x) = \alpha m_A(x)$ for $\alpha > 0$. If $0 \in A$, this holds also for $\alpha = 0$.

(ii) If A is radial, then $m_A(x) < \infty$ for all $x \in X$.

(iii) If A is balanced, then $m_A(\alpha x) = |\alpha| m_A(x)$ for all $x \in X$, $\alpha \in \mathbb{F}$.

(iv) If A is convex, then $m_A(x + y) \leq m_A(x) + m_A(y)$ for all $x, y \in X$.

Proof: (i)–(iii) are obvious; only (iv) needs proof. Suppose A is convex, and let $x, y \in X$. Let $\alpha > m_A(x)$ and $\beta > m_A(y)$ and $x \in \alpha A$, $y \in \beta A$. We have

$$\alpha A + \beta A = (\alpha + \beta)\left[\frac{\alpha}{\alpha + \beta}A + \frac{\beta}{\alpha + \beta}A\right] \subseteq (\alpha + \beta)A$$

since A is convex, so $x + y \in (\alpha + \beta)A$ and $m_A(x + y) \leq \alpha + \beta$. It follows that m_A is subadditive. 👍

XX.2. Normed Vector Spaces

XX.2.1. Completeness

XX.2.1.1. Proposition. Let X be a normed vector space and Y a (necessarily closed) subspace of X. If Y is complete, and X/Y is complete in the quotient norm, then X is complete.

Proof: Let $\bar{X}$ be the completion of X. Then Y is closed in $\bar{X}$ since Y is complete. Thus the quotient map from X to X/Y extends to a (bounded) quotient map π from $\bar{X}$ to $\bar{X}/Y$; so X/Y is naturally a subspace of $\bar{X}/Y$. On the one hand, X/Y is dense in $\bar{X}/Y$ since X is dense in $\bar{X}$ and π is continuous; on the other hand, X/Y is complete, hence closed in $\bar{X}/Y$. Thus $X/Y = \bar{X}/Y$, and hence $X = \bar{X}$ [if $z \in \bar{X}$, let $x \in X$ with $\pi(x) = \pi(z)$; then $z - x \in Y \subseteq X$, so $z = x + (z - x) \in X$]. 👍

Since finite-dimensional normed spaces are complete, we obtain:

XX.2.1.2. Corollary. Let X be a normed vector space and Y a (necessarily closed) subspace of X. If Y is complete, and X/Y is finite-dimensional, then X is complete.

XX.2.2. Isometries of Normed Vector Spaces

In this section, we prove the following theorem due to Mazur and Ulam []:

XX.2.2.1. Theorem. Let $\mathcal{V}$ and $\mathcal{W}$ be normed vector spaces over $\mathbb{R}$. Then any isometry from $\mathcal{V}$ onto $\mathcal{W}$ is an $\mathbb{R}$-affine function.

XX.2.2.2. This result is false if $\mathbb{R}$ is replaced by $\mathbb{C}$. For example, $\alpha \mapsto \bar{\alpha}$ is an isometry on the complex vector space $\mathbb{C}$ which is not $\mathbb{C}$-affine (it is, of course, $\mathbb{R}$-affine).

We begin with a simple observation:

XX.2.2.3. Proposition. Let $\mathcal{V}$ and $\mathcal{W}$ be Hausdorff topological vector spaces over $\mathbb{R}$, and $f : \mathcal{V} \to \mathcal{W}$ a continuous function. Then f is $\mathbb{R}$-affine if and only if, for every $x, y \in \mathcal{V}$, we have

$$f\left(\frac{x+y}{2}\right) = \frac{f(x)+f(y)}{2}$$

(i.e. f preserves midpoints).

Proof: It is obvious that an affine function preserves midpoints. Conversely, suppose f preserves midpoints. By composing f with a translation, we may assume $f(0) = 0$. It then follows that

$$f\left(\frac{x}{2}\right) = \frac{f(x)}{2}$$

for every $x \in \mathcal{V}$ by taking $y = 0$ in the midpoint property. Thus $f(2x) = 2f(x)$ for every $x \in \mathcal{V}$, and thus, for any $x, y \in \mathcal{V}$, we have

$$f(x+y) = f\left(\frac{2x+2y}{2}\right) = \frac{f(2x)+f(2y)}{2} = \frac{2f(x)+2f(y)}{2} = f(x)+f(y)$$

so f is additive. Then by iteration we get that $f(\alpha x) = \alpha f(x)$ for any dyadic rational α and any x. If $\alpha \in \mathbb{R}$ is arbitrary, let (α_n) be a sequence of dyadic rationals converging to α. Then, for any x,

$$f(\alpha x) = \lim_{n\to\infty} f(\alpha_n x) = \lim_{n\to\infty} \alpha_n f(x) = \alpha f(x)$$

since $\mathcal{W}$ is Hausdorff, i.e. the (translated) function f is linear. 👍

The special case of XX.2.2.1. for $\mathcal{V} = \mathcal{W} = \mathbb{R}^n$ follows almost immediately:

XX.2.2.4. THEOREM. Every isometry of $\mathbb{R}^n$ (with the usual Euclidean norm) is affine.

PROOF: Let f be an isometry of $\mathbb{R}^n$, and let $x, y \in \mathbb{R}^n$. Let $z = \frac{x+y}{2}$ be the midpoint of x and y. Then z is the unique point of $\mathbb{R}^n$ with the property that

$$\rho_2(x,z) = \rho_2(z,y) = \frac{1}{2}\rho_2(x,y)$$

where ρ_2 is the standard metric coming from the norm. It follows that $f(z)$ must satisfy

$$\rho_2(f(x),f(z)) = \rho_2(f(z),f(y)) = \frac{1}{2}\rho_2(f(x),f(y))$$

since f is an isometry. Thus $f(z)$ must be the midpoint of $f(x)$ and $f(y)$. Since f is continuous, it is affine by XX.2.2.3.. 👍

XX.2.2.5. This proof works more generally for XX.2.2.1. if $\mathcal{V}$ and $\mathcal{W}$ are strictly convex (). For the general case of XX.2.2.1., we need a subtler argument:

PROOF: Write ρ and σ for the norm metrics on $\mathcal{V}$ and $\mathcal{W}$ respectively. Fix x and y in $\mathcal{V}$, and set $z = \frac{x+y}{2}$. By XX.2.2.3. it suffices to show that

$$f(z) = \frac{f(x)+f(y)}{2} \ .$$

Let A_1 be the set of all $a \in \mathcal{V}$ such that

$$\rho(x,a) = \rho(a,y) = \frac{1}{2}\rho(x,y) \ .$$

Then A_1 is a subset of $\mathcal{V}$ containing z. We show that A_1 is symmetric around z, i.e. if $a \in A_1$, then $b = a + 2(z-a) = 2z - a$ is also in A_1 (b is the point for which z is the midpoint of a and b). We have $2z = x + y$, and thus $b - x = y - a$ and $b - y = x - a$, so

$$\rho(x,b) = \rho(y,a) = \rho(a,x) = \rho(b,y) = \frac{1}{2}\rho(x,y)$$

and $b \in A_1$.

Let δ_1 be the diameter of A_1. Then δ_1 is finite since if $a, b \in A_1$, we have

$$\rho(a,b) \leq \rho(a,x) + \rho(x,b) = \rho(x,y) \ .$$

For any $a \in A_1$, set $b = 2z - a \in A_1$. Since z is the midpoint of a and b, we have

$$\rho(a,z) = \frac{1}{2}\rho(a,b) \leq \delta_1 \ .$$

Now let A_2 be the set of $c \in A_1$ such that $\rho(a,c) \leq \frac{1}{2}\delta_1$ for all $a \in A_1$. Then $z \in A_2$. If $c \in A_2$, set $d = 2z - c$. If $a \in A_1$, set $b = 2z - a \in A_1$. We then have $2z = a + b$, so $d - a = c - b$ and hence $\rho(a,d) = \rho(b,c) \leq \frac{1}{2}\delta_1$, so $d \in A_2$ and A_2 is also symmetric around z. Also, if $c, d \in A_2$, then $c \in A_1$ so $\rho(c,d) \leq \frac{1}{2}\delta_1$. Thus the diameter δ_2 of A_2 satisfies $\delta_2 \leq \frac{1}{2}\delta_1$.

If $c \in A_2$, then $d = 2z - c \in A_2$, so we have

$$\rho(c,z) = \rho(z,d) = \frac{1}{2}\rho(c,d) \leq \frac{1}{2}\delta_2 \ .$$

Let A_3 be the set of $v \in A_2$ for which $\rho(c,v) \leq \frac{1}{2}\delta_2$ for all $c \in A_2$. Then $z \in A_3$ and, as before, A_3 is symmetric around z and the diameter δ_3 of A_3 satisfies $\delta_3 \leq \frac{1}{2}\delta_2$.

Iterating the process, we obtain a decreasing sequence (A_n) of sets containing z, symmetric around z, with diameter

$$\delta_n \leq \frac{1}{2}\delta_{n-1} \leq \frac{1}{2^{n-1}}\delta_1$$

and hence $\cap_n A_n = \{z\}$.

Let w be the midpoint of $f(x)$ and $f(y)$ in $\mathcal{W}$. Similarly, using σ construct a decreasing sequence (B_n) of subsets of $\mathcal{W}$ containing w. Since the A_n and B_n are constructed using just the metrics ρ and σ, and f is an isometry, we must have $f(A_n) = B_n$ for all n. Thus we have that

$$f(\{z\}) = f(\cap_n A_n) = \cap_n B_n = \{w\} \ .$$

XX.2.3. Exercises

XX.2.3.1. Use the Baire Category Theorem to give an alternate proof that that no Banach space has vector space dimension $\aleph_0$ (). [If $\{\xi_n : n \in \mathbb{N}\}$ is a Hamel basis for X, consider $X_n = span\{\xi_1, \ldots, \xi_n\}$.]

XX.3. Banach Spaces

"Banach started modestly enough from a class consisting of at least two completely arbitrary elements, and postulated that the class is closed under addition and under multiplication by real numbers subject to certain simple, almost trivial postulates. Again it seemed superficially to be the device of putting in what you want to get out. He also postulated an absolute value for which the triangular [sic] inequality holds. [This description leaves out the crucial notion of completeness.] It might seem that from such slavish copying of the rudiments of elementary algebra and elementary geometry nothing really new could issue. But this appears to have been contradicted by experience. In spite of Banach's professional rivals who say he put into his space the trivialities he wished to get out of it, 'Banach space' has at least unified the underlying concepts of several 'spaces' of interest to geometers."

E. T. Bell[1]

[1] [Bel87, p. 161-162].

XX.4. Topological Vector Spaces

Just as topological spaces are a generalization of metric spaces, topological vector spaces are a generalization of normed vector spaces. Throughout this section, $\mathbb{F}$ will denote either $\mathbb{R}$ or $\mathbb{C}$ (we will be specific in the few cases where it matters which), and "vector space" will mean "vector space over $\mathbb{F}$".

XX.4.1. Basic Definitions and Properties

XX.4.1.1. DEFINITION. A *topological vector space* (over $\mathbb{F}$) is a vector space X over $\mathbb{F}$ with a topology $\mathcal{T}$ satisfying

(i) $(x, y) \mapsto x+y$ is a continuous function from $(X, \mathcal{T}) \times (X, \mathcal{T})$ to $(X, \mathcal{T})$ ("addition is jointly continuous").

(i) $(\lambda, x) \mapsto \lambda x$ is a continuous function from $(\mathbb{F}, \mathcal{S}) \times (X, \mathcal{T})$ to $(X, \mathcal{T})$, where $\mathcal{S}$ is the usual topology on $\mathbb{F}$ ("scalar multiplication is jointly continuous").

In both cases, the topology on the product is the product topology. A topology on a vector space making it a topological vector space is called a *TVS topology*. Thus a topological vector space is a pair $(X, \mathcal{T})$, where X is a vector space and $\mathcal{T}$ a TVS topology on X.

XX.4.1.2. The topology $\mathcal{T}$ is part of the definition of a topological vector space (as is the field $\mathbb{F}$); thus it is technically incorrect to say "Let X be a topological vector space." But, as usual, we frequently abuse language by saying this when the topology is understood.

XX.4.1.3. EXAMPLES. (i) If $\|\cdot\|$ is a norm on a vector space X, the topology $\mathcal{T}$ defined by the corresponding metric $\rho(x, y) = \|x - y\|$ is a TVS topology on X, which is Hausdorff and even normal. [Note that it is technically incorrect to say a normed vector space *is* a topological vector space, just as it is incorrect to say that a metric space *is* a topological space. For one thing, different norms can define the same topology. But we frequently abuse language here too.] In particular, the ordinary topology on $\mathbb{F}$, or $\mathbb{F}^n$ for any $n \in \mathbb{N}$, is a TVS topology.

(ii) Let $\{(X_i, \mathcal{T}_i) : i \in I\}$ be a family of topological vector spaces (over the same field $\mathbb{F}$). Let $X = \prod_i X_i$. Then X is a vector space under coordinatewise operations, and the product topology on X is a TVS topology. In particular, if I is any set, then $\mathbb{F}^I$ is a topological vector space under the product topology from the ordinary topology on $\mathbb{F}$. If I is infinite, this topology does not come from a norm (XX.4.4.15.). If I is uncountable, this topology is not even first countable.

(iii) Let X be any vector space. Then the indiscrete topology on X is a (not very interesting) TVS topology. (We will primarily be concerned with Hausdorff topological vector spaces.)

We will see many more interesting examples.

XX.4.1.4. PROPOSITION. Let $(X, \mathcal{T})$ be a topological vector space, and Y a subspace of X. Then the relative topology $\tilde{\mathcal{T}}$ on Y is a TVS topology on Y.

XX.4.1.5. PROPOSITION. Let $(X, \mathcal{T})$ be a topological vector space, and Y a subspace of X. Then the closure $\bar{Y}$ of Y is also a subspace.

This is an easy consequence of the continuity of addition and scalar multiplication, and is left to the reader.

XX.4.1.6. If the scalar multiplication is ignored, a vector space is an abelian group under addition. If $(X, \mathcal{T})$ is a topological vector space, then by (i) it is a topological group under addition. Thus all results about abelian topological groups apply to topological vector spaces. For one thing, translation is a homeomorphism, i.e. the topology on a topological vector space is homogeneous (XXIII.1.2.9.). In fact:

XX.4.1.7. PROPOSITION. Let X be a topological vector space over $\mathbb{F}$. Then
(i) For any $y \in X$, the translation function $\tau_y : x \mapsto x + y = y + x$ is a homeomorphism from X to X.
(ii) For any $\lambda \in \mathbb{F}$, $\lambda \neq 0$, the scaling function $\sigma_\lambda : x \mapsto \lambda x$ is a homeomorphism from X to X.

PROOF: The functions τ_y and σ_λ are continuous by joint continuity (which implies separate continuity), and have inverses τ_{-y} and $\sigma_{\lambda^{-1}}$ respectively. 👍

In contrast to general abelian topological groups, we have:

XX.4.1.8. PROPOSITION. A topological vector space is path-connected.

PROOF: Let X be a topological vector space, and $x, y \in X$. Then the function $[0,1] \to X$ given by $\alpha \mapsto \alpha y + (1-\alpha)x$ is continuous and gives a path from x to y. 👍

This can be taken much further. The path is the "straight line segment" from x to y, and is the beginning of the theory of convex sets (), which is a crucial part of topological vector space theory. See also XX.4.1.17..

In fact, a topological vector space is contractible: the function $h : [0,1] \times X \to X$, $h(\alpha, x) = \alpha x$, is a contraction.

XX.4.1.9. If X is a topological vector space, $A \subseteq X$, an interior point of A in an internal point, by continuity of addition and scalar multiplication. In particular, any neighborhood of 0 is radial. The converse is false in general (XX.4.8.4.(b)).

We have (cf. XXIII.1.2.26.):

XX.4.1.10. PROPOSITION. Let $(X, \mathcal{T})$ be a topological vector space. Then the following are equivalent:

(i) $(X, \mathcal{T})$ is T_0.

(ii) $(X, \mathcal{T})$ is Hausdorff.

(iii) $(X, \mathcal{T})$ is completely regular.

(iv) For any $x \in X$, $x \neq 0$, there is an open neighborhood U of 0 with $x \notin U$.

We also have, in analogy with XXIII.1.2.29.:

XX.4.1.11. PROPOSITION. Let $(X, \mathcal{T})$ be a topological vector space, and Y a subspace of X. Let $\bar{\mathcal{T}}$ be the quotient topology () on the vector space X/Y. Then

(i) $\bar{\mathcal{T}}$ is a TVS topology on X/Y.

(ii) $(X/Y, \bar{\mathcal{T}})$ is Hausdorff if and only if Y is closed in X.

We will primarily be interested in Hausdorff topological vector spaces. Indeed, some authors require topological vector spaces to be Hausdorff. We will not since some natural topological vector spaces are not Hausdorff (). But there is a painless way to reduce to the Hausdorff case, an analogy (almost a special case) of XXIII.1.2.31.:

XX.4.1.12. PROPOSITION. Let X be a topological vector space, and $Y = \overline{\{0\}}$. Then Y is a (closed) subspace of X, and the quotient topology on X/Y is Hausdorff. X/Y is the maximal T_0 quotient of X in the sense of XIII.8.3.10., and the maximal Hausdorff quotient in the sense of XIII.8.5.6..

In fact, see XX.4.8.1..

Specification by Fundamental Neighborhoods

Just as in a topological group, TVS topology on a vector space can be described by giving a fundamental family of neighborhoods of 0.

XX.4.1.13. DEFINITION. Let X be a vector space. A *fundamental system of neighborhoods of* 0 is a nonempty collection $\mathcal{F}$ of subsets of X, satisfying:

(i) Each $U \in \mathcal{F}$ is radial and balanced (and thus contains 0).

(ii) If $U, V \in \mathcal{F}$, there is a $W \in \mathcal{F}$ with $W \subseteq U \cap V$.

(iii) For each $U \in \mathcal{F}$ there is a $V \in \mathcal{F}$ with $V + V \subseteq U$.

Note that in this definition, there is no specified topology on X, so the sets in $\mathcal{F}$ are not "neighborhoods" of 0 in X in the usual sense of the term. But the terminology is justified by:

XX.4.1.14. THEOREM. Let X be a vector space, and let $\mathcal{F}$ be a fundamental system of neighborhoods of 0 in X. Then there is a (unique) TVS topology $\mathcal{T}$ on X such that $\mathcal{F}$ forms a local base for $\mathcal{T}$ at 0 (consisting of not necessarily open sets). Conversely, if $\mathcal{T}$ is a TVS topology on X, then there is a fundamental system of neighborhoods of 0 in X forming a local base for $\mathcal{T}$; in fact, the set of all balanced neighborhoods of 0 is a fundamental system of neighborhoods of 0 forming a local base for $\mathcal{T}$.

PROOF: If $\mathcal{F}$ is a fundamental system of neighborhoods of 0 in X, then $\mathcal{F}$ satisfies the conditions of XXIII.1.2.18., and thus is a local base for a topology $\mathcal{T}$ on X making X into a topological group under addition, i.e. addition is jointly continuous for $\mathcal{T}$.

To show that scalar multiplication is jointly continuous, fix $\lambda_0 \in \mathbb{F}$ and $x_0 \in X$. If $U \in \mathcal{F}$, then $\lambda_0 x_0 + U$ is a basic neighborhood of $\lambda_0 x_0$. Choose $V \in \mathcal{F}$ with $V + V \subseteq U$. Since V is radial, there is an $\epsilon > 0$ such that $(\lambda - \lambda_0)x \in V$ whenever $|\lambda - \lambda_0| \leq \epsilon$. Fix $n \in \mathbb{N}$ with $n \geq |\lambda_0| + \epsilon$, and by repeatedly using XX.4.1.13., there is a $W \in \mathcal{F}$ with $W + \cdots + W$ (n terms) contained in V; so $nW \subseteq V$. Then $x - x_0 \in W$ and $|\lambda - \lambda_0| \leq \epsilon$ imply $\lambda(x - x_0) \in V$ since V is balanced; so

$$\lambda x = \lambda_0 x_0 + (\lambda - \lambda_0)x_0 + \lambda(x - x_0) \in \lambda_0 x_0 + V + V \subseteq \lambda_0 x_0 + U \ .$$

For the converse, let $\mathcal{T}$ be a TVS topology on X. Any neighborhood of 0 is radial by (separate) continuity of scalar multiplication. Let U be a neighborhood of 0. By continuity of scalar multiplication, there is an $\epsilon > 0$ and a neighborhood V of 0 such that $\lambda V \subseteq U$ whenever $|\lambda| \leq \epsilon$. Set $W = \cup_{|\lambda| \leq \epsilon} \lambda V$; then W is a balanced neighborhood of 0 contained in U. Thus the set $\mathcal{F}$ of balanced neighborhoods of 0 is a local base for $\mathcal{T}$, and satisfies XX.4.1.13.(i) and (ii). If $U \in \mathcal{F}$, by continuity of addition there is a $\mathcal{T}$-neighborhood V of 0 with $V + V \subseteq U$, and V contains a balanced neighborhood W of 0, which satisfies $W + W \subseteq U$. Thus $\mathcal{F}$ satisfies XX.4.1.13.(iii). 👍

This proof is adapted from [SW99], one of the standard references for topological vector spaces.

XX.4.1.15. Corollary. Let $\mathcal{T}$ be a translation-invariant topology on a vector space X (i.e. for which translations are homeomorphisms). If there is a fundamental system $\mathcal{F}$ of neighborhoods of 0 in X forming a local base at 0 for $\mathcal{T}$, then $\mathcal{T}$ is a TVS topology on X.

Proof: There is a TVS topology $\mathcal{S}$ on X for which $\mathcal{F}$ forms a local base at 0; so $\mathcal{S}$ and $\mathcal{T}$ have identical local bases at every point since $\mathcal{T}$ is translation-invariant. Thus $\mathcal{S} = \mathcal{T}$. 👍

Finite-Dimensional Topological Vector Spaces

We also have the following generalization of (). We have designated this as a theorem since it is a little more than just a corollary of XX.4.1.14., although XX.4.1.14. (converse direction) is the crucial tool in the proof.

XX.4.1.16. Theorem. Let $n \in \mathbb{N}$. Then the ordinary topology on $X = \mathbb{F}^n$ is the unique Hausdorff TVS topology on $\mathbb{F}^n$.

Proof: We may assume $\mathbb{F} = \mathbb{R}$ and $X = \mathbb{R}^n$. Let $\mathcal{T}$ be a Hausdorff TVS topology on $\mathbb{R}^n$, and $\mathcal{S}$ the ordinary topology, i.e. $\mathcal{S}$ comes from a norm on $\mathbb{R}^n$ (any norm will do, but it will be convenient to use $\| \cdot \|_1$. Reference to balls in this proof will be with respect to this norm.)

Let $\mathcal{F}$ be a fundamental system of neighborhoods of 0 for $\mathcal{T}$, and $U \in \mathcal{F}$. Let B be the set consisting of the standard basis vectors in $\mathbb{R}^n$ and their negatives. Then B is a finite set with $2n$ elements. The convex hull C of B is an $\mathcal{S}$-neighborhood of 0 (it is the closed unit ball). By repeated use of XX.4.1.14.(iii), there is a $V \in \mathcal{F}$ with $V + \cdots + V$ ($2n$ terms) contained in U. Since V is radial, V absorbs B, i.e. there is an $\epsilon > 0$ such that $\epsilon B \subseteq V$. Since V is balanced, it follows that ϵC, which is an $\mathcal{S}$-neighborhood of 0, is contained in U. So every $\mathcal{T}$-neighborhood of 0 contains an $\mathcal{S}$-neighborhood, i.e. $\mathcal{S}$ is stronger than $\mathcal{T}$.

Let S be the unit sphere in $\mathbb{R}^n$. Then S is $\mathcal{S}$-compact, hence $\mathcal{T}$-compact, so S is $\mathcal{T}$-closed since $\mathcal{T}$ is Hausdorff. Thus $Z = \mathbb{R}^n \setminus S$ is a $\mathcal{T}$-open neighborhood of 0, so contains a $W \in \mathcal{F}$. Let $x \in W$. Since W is

balanced, the line segment $\{\lambda x : 0 \leq \lambda \leq 1\}$ is contained in W. This line segment cannot cross S, so x is in the open unit ball of $\mathbb{R}^n$, i.e. W is contained in the open unit ball. Thus every $\mathcal{S}$-neighborhood of 0 contains a $\mathcal{T}$-neighborhood (cf. XX.4.1.7.(ii)), i.e. $\mathcal{T}$ is stronger than $\mathcal{S}$. Therefore $\mathcal{T} = \mathcal{S}$. 👍

XX.4.1.17. Corollary. Let X be a Hausdorff topological vector space, and Y a finite-dimensional affine subspace of X. Then there is an affine homeomorphism from Y to $\mathbb{F}^n$, where $n = dim(Y)$. In particular, if X is finite-dimensional, there is a unique Hausdorff TVS topology on X.

Thus, although a vector space can have many Hausdorff TVS topologies, they cannot be drastically different: they must all agree on every finite-dimensional affine subspace.

XX.4.2. Locally Convex Topological Vector Spaces

> "The theory of general topological vector spaces ... remains on the whole an unsatisfactory theory, devoid of a great number of valuable results both from pure and applied viewpoints. The concept of topological vector space ... is too general to support a rich theory just as, on the other hand, the concept of Banach space is too narrow. The notion on which a satisfactory and applicable theory can be built is that of local convexity ..."
>
> *H. Schaefer*[2]

As pointed out in [KN76], the main extra feature in topological vector spaces distinguishing them from general (abelian) topological groups is convexity. See XX.1.2.1..

XX.4.2.1. Definition. Let $(X, \mathcal{T})$ be a topological vector space. Then $(X, \mathcal{T})$ is *locally convex* if it is Hausdorff and there is a base for $\mathcal{T}$ consisting of convex sets.

XX.4.2.2. Some authors do not require a locally convex topological vector space to be Hausdorff, but it will be convenient for us to make this assumption to avoid verbosity, and since essentially all interesting topological vector spaces are Hausdorff. We call a locally convex TVS topology an *LCS topology*.

XX.4.2.3. Examples. (i) Any normed vector space is locally convex (balls are convex by the triangle inequality). In particular, the unique TVS topology on a finite-dimensional vector space is locally convex.

(ii) If X is a normed vector space and Y a total subspace of X^*, then the $\sigma(X, Y)$ topology () is an LCS topology on X.

(iii) The inductive limit topology on a directed inductive limit (XX.4.8.2.) of LCS topologies is LCS (if Hausdorff). In particular, the maximal TVS topology on a vector space is locally convex.

[2] [SW99, p. 36].

XX.4.2.4. PROPOSITION. Let $(X, \mathcal{T})$ be a Hausdorff topological vector space. The following are equivalent:

(i) $(X, \mathcal{T})$ is locally convex.

(ii) There is a local base at 0 consisting of convex sets.

(iii) There is a local base at 0 consisting of balanced convex open sets.

XX.4.2.5. The importance of local convexity is, in part, due to the fact that the following are equivalent:

(i) $(X, \mathcal{T})$ is locally convex.

(ii) The topology $\mathcal{T}$ is defined by a separating family of seminorms (XX.4.4.5.).

(iii) There are "sufficiently many" continuous linear functionals on X ().

In fact, a Hausdorff topological vector space which is not locally convex can fail to have any nonzero continuous linear functionals (XX.4.5.5.). Important theorems such as the Krein-Milman Theorem () can fail in Hausdorff topological vector spaces which are not locally convex.

XX.4.3. Paranorms and Fréchet Spaces

There is a useful extension of the notion of a norm:

XX.4.3.1. DEFINITION. Let X be a vector space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. A function $p : X \to [0, \infty)$ is a *paranorm* if it satisfies:

(i) $p(x) = 0$ if and only if $x = 0$.

(ii) $p(x + y) \leq p(x) + p(y)$ for all $x, y \in X$.

(iii) $p(-x) = p(x)$ for all $x \in X$.

(iv) Scalar multiplication in X is jointly continuous for the metric ρ defined by $\rho(x, y) = p(x - y)$.

The paranorm p is *balanced* if $p(\alpha x) \leq p(x)$ for all $x \in X$ and all $\alpha \in \mathbb{F}$, $|\alpha| \leq 1$. This implies that $p(\alpha x) \leq p(\beta x)$ for all $x \in X$ and all $\alpha, \beta \in \mathbb{F}$, $|\alpha| \leq |\beta|$ (see the proof of XX.4.3.5.). As with norms, we say a paranorm is *complete* if its corresponding metric is complete.

It follows from (iv) that X with the metric topology is a Hausdorff topological vector space (from (ii), addition is jointly continuous).

XX.4.3.2. A paranorm is really the same thing as a translation-invariant metric satisfying (iv): if p is a paranorm, then ρ defined by $\rho(x, y) = p(x - y)$ is a translation-invariant metric. Conversely, if ρ is a translation-invariant metric satisfying (iv), set $p(x) = \rho(x, 0)$; then p is a paranorm defining ρ. Thus, if $(X, \mathcal{T})$ is a topological vector space, then the topology $\mathcal{T}$ comes from a paranorm if and only if it is Hausdorff and first countable (XX.4.4.1.).

XX.4.3.3. The main difference between a paranorm and a norm is that it is not assumed that $p(\alpha x) = |\alpha| p(x)$ for $x \in X$, $\alpha \in \mathbb{F}$. In fact a norm is precisely a paranorm with this additional homogeneity property (from which (iv) is automatic for scalar multiplication, and which is automatically balanced).

It turns out that paranorms are more common than norms, and there are many important topological vector spaces whose topology is given by a paranorm but not a norm. **Caution:** because of the lack of homogeneity in general, a topological vector space whose topology is defined by a paranorm is not necessarily locally convex (XX.4.3.7.(iii)).

XX.4.3.4. A function satisfying (ii)–(iv), and

(i′) $p(0) = 0$.

is called an *indefinite paranorm* (we will avoid the terms *semiparanorm* or *paraseminorm*). A standard example is a seminorm.

Most natural paranorms are balanced, and any paranorm is equivalent to a balanced one (in the sense that the corresponding metrics are uniformly equivalent):

XX.4.3.5. Proposition. Let p be a paranorm on a vector space X. Then there is a balanced paranorm $\tilde{p}$ on X inducing the same topology as p.

Proof: Since $\mathbb{D} := \{\alpha \in \mathbb{F} : |\alpha| \leq 1\}$ is compact and p is continuous for the induced metric, for any $x \in X$ there is an M_x such that $p(\alpha x) \leq M_x$ for all $\alpha \in \mathbb{D}$. So we set

$$\tilde{p}(x) = \sup_{\alpha \in \mathbb{D}} p(\alpha x) \ .$$

It is easily checked that $\tilde{p}$ satisfies (i)–(iii), and clearly $\tilde{p}$ is balanced and $p(x) \leq \tilde{p}(x)$ for any $x \in X$.

We show that if (x_n) is a sequence in X and $p(x_n) \to 0$, then $\tilde{p}(x_n) \to 0$. Suppose not; passing to a subsequence, there is an $\epsilon > 0$ and, for each n, an $\alpha_n \in \mathbb{D}$, with $p(\alpha_n x_n) \geq \epsilon$. Passing to a further subsequence, we may assume $\alpha_n \to \alpha$ for some $\alpha \in \mathbb{D}$; then $p(\alpha_n x_n) \to \alpha \cdot 0 = 0$ by joint continuity of scalar multiplication for p, a contradiction.

If $|\alpha| \leq |\beta|$, $\beta \neq 0$, and $x \in X$, then

$$\tilde{p}(\alpha x) = \sup_{|\gamma| \leq 1} p(\gamma \cdot \alpha x) = \sup_{|\gamma| \leq |\frac{\alpha}{\beta}|} p\left(\gamma \cdot \beta x\right) \leq \sup_{|\gamma| \leq 1} p\left(\gamma \cdot \beta x\right) = \tilde{p}(\beta x) \ .$$

It remains to note that scalar multiplication is jointly continuous for $\tilde{p}$. It suffices to show that if $\alpha_n \to \alpha$ and $x_n \to 0$, then $p(\alpha_n x_n) \to 0$. For any $m \in \mathbb{N}$ and $x \in X$, we have $p(x) \leq m p\left(\frac{x}{m}\right)$ by the triangle inequality. Choose m so that $m \geq |\alpha_n|$ for all n. Then we have

$$\tilde{p}(\alpha_n x_n) = \sup_{|\gamma| \leq 1} p(\gamma \alpha_n x_n) \leq m \sup_{|\gamma| \leq 1} p\left(\gamma \frac{\alpha_n}{m} x_n\right) \leq m \tilde{p}(x_n) \to 0 \ .$$

XX.4.3.6. From now on, for economy of exposition we will take "paranorm" to mean "balanced paranorm" unless specifically stated otherwise. Many of the results also apply to unbalanced paranorms (which rarely arise in practice); the interested reader can work out the details.

XX.4.3.7. EXAMPLES. (i) Any norm is a paranorm. Any positive multiple of a paranorm is a paranorm. Any (finite) sum of paranorms is a paranorm.

(ii) If p is a paranorm, so is $\sqrt{p}$ (or p^α for $0 < \alpha < 1$), defining the same topology as p. If p is a norm, then p^α $(0 < \alpha < 1)$ is not a norm.

(iii) Similarly, if $0 < q < 1$, the vector space ℓ^q of q'th power summable sequences has a paranorm

$$p((x_k)) = \sum_{k=1}^{\infty} |x_k|^q \ .$$

The corresponding TVS topology on ℓ^q is not locally convex. More generally, if $(X, \mathcal{A}, \mu)$ is a measure space, then $L^q(X, \mathcal{A}, \mu)$ $(0 < q < 1)$ has a paranorm

$$p(f) = \int_X |f(x)|^q \, d\mu(x) \ .$$

The corresponding TVS topology on $L^q(X, \mathcal{A}, \mu)$ is rarely locally convex.

(iv) Here is a crucial construction which can be done with paranorms but not norms. Let p be a paranorm on X, and $\lambda > 0$. For $x \in X$ set

$$\tilde{p}_\lambda(x) = \min(p(x), \lambda)$$

$$\bar{p}(x) = \frac{p(x)}{1 + p(x)}$$

(cf. ()). Then p_λ and $\bar{p}$ are paranorms on X defining the same topology as p. These paranorms are bounded, hence can never be norms. These constructions can also be done with indefinite paranorms.

Here is an important construction which follows from (iv). The proof is essentially the same as the proof of ().

XX.4.3.8. PROPOSITION. Let (p_k) be a sequence of paranorms on a vector space X, and for $x \in X$ set

$$p(x) = \sum_{k=1}^{\infty} 2^{-k} \bar{p}_k(x) \ .$$

Then p is a paranorm on X. The p-topology is the weakest topology stronger than the p_k-topology for all k, i.e. $x_n \to x$ in the p-topology if and only if $x_n \to x$ in the p_k-topology for all k.

XX.4.3.9. Instead of $\sum_k 2^{-k} \bar{p}_k(x)$, we could have used $\sum_k 2^{-k} (\widetilde{p_k})_1$ or $\sum_k (\widetilde{p_k})_{2^{-k}}$, which are also paranorms giving the same topology. More generally, instead of paranorms we could use a separating sequence of indefinite paranorms, e.g. a separating sequence of seminorms.

Completion

XX.4.3.10. If p is a paranorm on a vector space X, then the completion $\bar{X}$ of X with respect to the corresponding metric has a natural structure as a vector space, and p extends naturally to a function $\bar{p}$ on $\bar{X}$. It is routine to show that the completion with respect to the metric is the same (isomorphic to) the completion as a topological vector space (), and hence (using ()) that $\bar{p}$ is a paranorm on $\bar{X}$ giving its topology (only joint continuity of multiplication is an issue; a direct proof is somewhat simpler than the general proof of ()).

F-Spaces and Fréchet Spaces

XX.4.3.11. Definition. An *F-space* is a topological vector space whose topology is given by a complete paranorm.

A *Fréchet space* is a locally convex F-space.

XX.4.3.12. **Caution:** Terminology in this subject is not standardized. Some references use the terms F-space and Fréchet space interchangeably, with or without the locally convex requirement. It is also not standard whether the paranorm is part of the definition of an F-space or Fréchet space (i.e. there are "isometric" and "isomorphic" theories). Even the term *paranorm* is not entirely standard. Check the definitions in any reference.



XX.4.3.13. Examples. (i) Any Banach space is a Fréchet space (more precisely, the underlying topological vector space of a Banach space is a Fréchet space).

(ii) If $0 < q < 1$, ℓ^q (XX.4.3.7.(iii)) is an F-space which is not a Fréchet space.

(iii) A Hausdorff topological vector space is an F-space if and only if it is complete and first countable (). It is a Fréchet space if and only if it is complete, first countable, and locally convex. Thus the completion of any first countable Hausdorff topological vector space is an F-space, and the completion of any Hausdorff first countable locally convex topological vector space is a Fréchet space.

There are many Fréchet spaces whose topology does not come from a norm, so Fréchet spaces are considerably more general than Banach spaces.

XX.4.3.14. Remark. We can relax the definition of paranorm. Say a function p on a vector space X is a *weak paranorm* (nonstandard term) if it satisfies XX.4.3.1.(i)–(iii) and

(iv′) Scalar multiplication is separately continuous for the corresponding metric.

It turns out that a complete weak paranorm is automatically a paranorm (Exercise XX.4.8.3.). However, completeness seems to be necessary: it is unclear whether XX.4.3.10. holds for a weak paranorm (the issue is whether the scalar multiplication is separately continuous in the completion). The notion of weak paranorm has limited application in practice. See [DS88].

XX.4.3.15. Here is an important standard construction of Fréchet spaces. Let X be a vector space, and (p_k) an increasing sequence of paranorms on X. Let X_1 be the completion of X with respect to p_1. Since the p_k are stronger than p_1, the completion X_k of X with respect to p_k can be identified with a subspace of X_1. The X_k form a decreasing sequence of subspaces of X_1, and $X \subseteq \cap_k X_k$. Form the paranorm p on X from the p_k as in XX.4.3.7.(iv). Then the completion $\bar{X}$ of X is naturally a subspace of X_1.

XX.4.3.16. Proposition. We have $\bar{X} = \cap_k X_k$, and $\bar{X}$ is an F-space. If all the X_k are Fréchet spaces (e.g. if the p_k are norms), then $\bar{X}$ is a Fréchet space. In particular, if $X = \cap X_k$, then X is an F-space from p.

Proof: We have $\bar{X} \subseteq \cap_k X_k$. Conversely, if $x \in \cap_k X_k$, then for each k there is a sequence x_{kn} in X with $p_k(x_{kn} - x) < 2^{-n}$. The diagonal subsequence (x_{kk}) is a p-Cauchy sequence which converges to x, so $x \in \bar{X}$. 👍

XX.4.3.17. Example. Let $[a, b]$ be an interval in $\mathbb{R}$, and let $X = \mathcal{C}^\infty([a, b])$. For $f \in X$, let

$$p_k(f) = \max_{0 \le r \le k} \max_{x \in [a,b]} |f^{(r)}(x)| \ .$$

Then $X_k = \mathcal{C}^k([a, b])$, and $X = \cap_k X_k$. Each p_k is a norm, and X_k is a Banach space; so X is a Fréchet space. We have that $f_n \to f$ in $\mathcal{C}^\infty([a, b])$ if and only if $f_n^{(r)} \to f^{(r)}$ uniformly on $[a, b]$ for each $r \ge 0$. This is the standard topology on $\mathcal{C}^\infty([a, b])$. This topology is also the topology determined by the seminorms $s_k(f) = \|f^{(k)}\|_\infty$, or the paranorm constructed from these. It turns out that this topology does not come from a norm (XX.4.4.14.).

XX.4.3.18. A somewhat similar construction is also common. Let X be a vector space, and (p_k) a sequence of indefinite paranorms (e.g. seminorms) on X which is separating, i.e. $p_k(x) = 0$ for all k implies $x = 0$. Replacing p_k by the supremum or sum of the first k, we may assume the p_k are increasing. Let $Y_k = \{x \in X : p_k(x) = 0\}$. Then Y_k is a subspace of X, and p_k drops to a paranorm $\bar{p}_k$ on X/Y_k. The Y_k are decreasing, and the quotient map π_k from X/Y_j to X/Y_k is continuous (for the induced topologies) for $j > k$. The p_k define a paranorm p on X by XX.4.3.7.(iv).

XX.4.3.19. Proposition. Using the above notation, suppose

(a) X/Y_k is $\bar{p}_k$-complete for each k.

(b) If $(x^{(k)})$ is a coherent sequence, i.e. $x^{(k)} \in X/Y_k$ and $x^{(k)} = \pi_k(x^{(j)})$ for all $j > k$, there is a (necessarily unique) $x \in X$ with $x^{(k)} = \pi_k(x)$ for all k (in other words, $X = \lim_{\leftarrow} X/Y_k$ ()).

Then X is complete under p, i.e. X is an F-space. If the p_k are seminorms, X is a Fréchet space. A sequence (x_n) is X is p-Cauchy if and only if $(\pi_k(x_n))$ is a $\bar{p}_k$-Cauchy sequence in X/Y_k for all k.

Proof: Let (x_n) be a p-Cauchy sequence in X. Set $x_n^{(k)} = \pi_k(x_n)$ for all n and k. Then $(x_n^{(k)})$ is a $\bar{p}_k$-Cauchy sequence in X/Y_k, hence converges to some $x^{(k)} \in X/Y_k$. Since the quotient maps are continuous, the $x^{(k)}$ form a coherent sequence, so there is a (unique) $x \in X$ with $\pi(x) = x^{(k)}$ for all k; and $x_n \to x$. 👍

XX.4.3.20. Condition (b) is not automatic, even if condition (a) holds: for example, let $X = C([0,1])$, and (t_k) a dense sequence in $[0,1]$. Set $Y_k = \{f \in X : f(t_j) = 0 \text{ for } 1 \le j \le k\}$ and $p_k(f) = \max\{|f(t_1)|, \dots, |f(t_k)|\}$. Any function from the t_k to $\mathbb{F}$ gives a coherent sequence, but most such functions do not extend to a continuous function on $[0,1]$. The p-topology on X is the topology of pointwise convergence on the t_k.

XX.4.3.21. EXAMPLE. Let s be the vector space of all sequences in $\mathbb{F}$, i.e. $s = \mathbb{F}^{\mathbb{N}}$. The sequence (p_k) of seminorms defined by $p_k((x_n)) = |x_k|$ defines a paranorm p on s as in XX.4.3.7.(iv). It is easily checked directly that s is complete with respect to p (a sequence in s is p-Cauchy if and only if the sequence of k'th terms is Cauchy for each k), so is a Fréchet space. The topology on s is the topology of coordinatewise convergence (the product topology on $\mathbb{F}^{\mathbb{N}}$). We could replace $\mathbb{F}$ here by any Banach space (or, more generally, any Fréchet spaces X_k, which could vary from coordinate to coordinate, i.e. a countable product of Fréchet spaces is a Fréchet space with the product topology).

XX.4.3.22. EXAMPLE. Let Z be a locally compact, σ-compact Hausdorff space, e.g. $Z = \mathbb{R}^n$. Let (U_k) be an increasing sequence of open sets in Z with compact closure E_k, with $Z = \cup_k U_k$. On the vector space $X = C(Z)$ of all continuous functions from Z to $\mathbb{F}$, set

$$p_k(f) = \max_{x \in E_k} |f(x)| \ .$$

Then $Y_k = \{f \in C(Z) : f = 0 \text{ on } E_k\}$. We have $\cap_k Y_k = \{0\}$ and $X/Y_k \cong C(E_k)$, and $\bar{p}_k = \| \cdot \|_\infty$; thus X/Y_k is a Banach space. If $(f^{(k)})$ is a coherent sequence, i.e. $f^{(k)}$ is a continuous function on E_k which is the restriction of a continuous function on E_j for $j > k$, there is a continuous function f on Z whose restriction to E_k is $f^{(k)}$ for all k. Thus the hypotheses of XX.4.3.19. are satisfied, so $C(Z)$ with the p-paranorm is a Fréchet space. The topology on $C(Z)$ is the topology of u.c. convergence.

The two constructions can be combined:

XX.4.3.23. EXAMPLE. Let $X = \mathcal{C}^\infty(\mathbb{R}^m)$ for some m. For each k let

$$p_k(f) = \max_{0 \le |\alpha| \le k} \max_{\|\boldsymbol{x}\| \le k} |f^{(\alpha)}(\boldsymbol{x})| \ .$$

Then (p_k) is an increasing sequence of seminorms on X which is separating. If p is the induced paranorm on X, a p-Cauchy sequence (f_n) gives a $\bar{p}_k$-Cauchy sequence in $\mathcal{C}^k(B_k)$ for each k, where B_k is the closed ball of radius k in $\mathbb{R}^m$, which converges to some f^k, and the f^k form a coherent sequence. There is a unique $f \in X$ which restricts to f^k for each k, and $f_n \to f$. Thus X is p-complete, hence a Fréchet space. A sequence (f_n) in $\mathcal{C}^\infty(\mathbb{R}^m)$ converges to f in this topology if and only if $(f_n^{(\alpha)})$ converges to $f^{(\alpha)}$ u.c. for all α.

Many similar examples and constructions are possible.

XX.4.4. Metrizability and Normability

In this section, we describe the conditions under which a topological vector space is metrizable and/or normable.

Metrizability

The next theorem is a special case of (), but we give a proof here to keep this section independent of ().

XX.4.4.1. THEOREM. A topological vector space which is Hausdorff and first countable has a paranorm (equivalently, an invariant metric) inducing its topology and uniform structure.

We will give an alternate proof in the locally convex case in XX.4.4.6..

Minkowski Functionals

In the locally convex case, the key tool for further analysis is Minkowski functionals ().

The next result is an immediate corollary of XX.1.4.2.:

XX.4.4.2. COROLLARY. Let X be a vector space, and $A \subseteq X$ be radial, balanced, and convex. Then the Minkowski functional m_A is a seminorm on X.

In general, under the hypotheses m_A is only a seminorm.

EXAMPLE. Let X be a normed vector space, and A the unit ball of X (open or closed). Then $m_A(x) = \|x\|$ for all $x \in X$. More generally, if A is the ball of radius r, then $m_A(x) = \frac{1}{r}\|x\|$.

XX.4.4.3. If A is radial, balanced, and convex, we have

$$\{x \in X : m_A(x) < 1\} \subseteq A \subseteq \{x \in X : m_A(x) \leq 1\}$$

but this is the best that can be said in general: if X is a topological vector space, we have $m_A = m_{\bar{A}}$, and if A has nonempty interior (which is not automatic), then $m_A = m_{A^\circ}$. (And note that the definition of Minkowski functional applies in any vector space X and does not presuppose a topology on X.)

XX.4.4.4. PROPOSITION. Let $(X, \mathcal{T})$ be a topological vector space, and A a radial, balanced, convex subset of X. Then the following are equivalent:

(i) $m_A : X \to [0, \infty)$ is $\mathcal{T}$-continuous at 0.

(ii) m_A is $\mathcal{T}$-uniformly continuous.

(iii) A is a $\mathcal{T}$-neighborhood of 0 in X.

(iv) $U_A := \{x \in X : m_A(x) < 1\}$ is a $\mathcal{T}$-neighborhood of 0 in X.

(v) $E_A := \{x \in X : m_A(x) \leq 1\}$ is a $\mathcal{T}$-neighborhood of 0 in X.

PROOF: (ii) ⇒ (iv) since $U_A = m_A^{-1}([0,1))$, which is open in $[0,\infty)$. (iv) ⇒ (iii) ⇒ (v) by the previous remark.

(i) ⇒ (ii): If m_A is continuous at 0, then for any $\epsilon > 0$ there is a $\mathcal{T}$-neighborhood V of 0 such that $m_A(x) < \epsilon$ for all $x \in V$. Then, for any $x, y \in X$, $m_A(x-y) < \epsilon$ whenever $x - y \in V$, so m_A is uniformly continuous.

(v) ⇒ (i): If m_A is not continuous at 0, there is a net (x_i) $\mathcal{T}$-converging to 0 such that $m_A(x_i) \not\to 0$. Passing to a subnet, there is an $\epsilon > 0$ such that $m_A(x_i) > \epsilon$ for all i. Then $\frac{1}{\epsilon}x_i \to 0$, but $m_A\left(\frac{1}{\epsilon}x_i\right) > 1$ for all i, so no x_i is in E_A. Thus E_A is not a $\mathcal{T}$-neighborhood of 0. 👍

XX.4.4.5. THEOREM. Let $(X, \mathcal{T})$ be a (Hausdorff) topological vector space. Then the topology $\mathcal{T}$ is determined by a (separating) family of seminorms if and only if X is locally convex. The topology $\mathcal{T}$ is determined by a countable (separating) family of seminorms if and only if X is locally convex and first countable.

PROOF: If $\mathcal{T}$ is determined by a family (s_i) of seminorms, then finite intersections of sets of the form $\{x \in X : s_i(x) < \epsilon\}$ form a local base for $\mathcal{T}$ at 0 of convex sets, so $(X, \mathcal{T})$ is locally convex.

Conversely, suppose X is locally convex, and let $\mathcal{F}$ be a fundamental system of neighborhoods of 0 consisting of convex sets (). For each $V \in \mathcal{F}$ let m_V be the corresponding Minkowski functional. Then the m_V are seminorms on X, and we claim $\mathcal{T}$ is the topology determined by $\{m_V : V \in \mathcal{F}\}$. Each m_V is $\mathcal{T}$-continuous, so if $x_i \to 0$, then $m_V(x_i) \to 0$. On the other hand, suppose (x_i) is a net in X and $m_V(x_i) \to 0$ for all $V \in \mathcal{F}$. Let U be a $\mathcal{T}$-neighborhood of 0. There is a $V \in \mathcal{F}$ with $V \subseteq U$. There is an i_0 such that $m_V(x_i) < 1$ for all $i > i_0$, hence $x_i \in V \subseteq U$ for $i > i_0$; thus $x_i \to 0$ in $(X, \mathcal{T})$.

If $(X, \mathcal{T})$ is Hausdorff, for each $x \in X$, $x \neq 0$, there is a $V \in \mathcal{F}$ with $x \notin V$, so $m_V(x) \geq 1$ and $m_V(x) \neq 0$. Thus the m_V are a separating family of seminorms on X.

If X is first countable, we may take $\mathcal{F}$ to be countable. Conversely, if $\mathcal{T}$ is determined by a countable family (s_k) of seminorms, finite intersections of sets of the form $U_{k,n} = \left\{x \in X : s_k(x) < \frac{1}{n}\right\}$ form a countable local base for $\mathcal{T}$ at 0. 👍

XX.4.4.6. Since a separating sequence of seminorms can be combined into a single paranorm by XX.4.3.8., we obtain another proof of XX.4.4.1. in the locally convex case (and even a somewhat better result).

Bounded Sets

XX.4.4.7. DEFINITION. Let X be a topological vector space. A subset A of X is *bounded* ($\mathcal{T}$-bounded) if every $\mathcal{T}$-neighborhood of 0 absorbs A, i.e. for every $\mathcal{T}$-neighborhood U of 0 there is a λ with $A \subseteq \lambda U$, i.e. $\epsilon A \subseteq U$ for some $\epsilon > 0$.

Clearly, to check that A is bounded it suffices to let U run over a base for the topology. Any subset of a bounded set is bounded.

XX.4.4.8. EXAMPLES. (i) Let $(X, \|\cdot\|)$ be a normed vector space. Then, since the balls B_r around 0 are a base for the topology, and $B_r = rB_1$ for any $r > 0$, the following are equivalent for a subset A of X:

(a) A is bounded (in the sense of XX.4.4.7.).

(b) B_r absorbs A for every $r > 0$.

(c) B_1 absorbs A.

(d) A is bounded in the norm sense, i.e. there is an M such that $\|x\| \leq M$ for all $x \in A$.

(ii) Similarly, let $0 < q < 1$ and $X = \ell^q$, with topology from the paranorm p of XX.4.3.7.(iii). Let $B_r = \{x \in X : p(x) < r\}$. Then the B_r form a base for the topology, and $B_r = r^{1/q}B_1$ for any r. Then a subset A of X is bounded if and only if it is contained in B_r for some r. In particular, the B_r are bounded.

XX.4.4.9. PROPOSITION. Let $(X, \mathcal{T})$ be a topological vector space, and s a $\mathcal{T}$-continuous seminorm on X. If A is a bounded subset of X, then $s(A)$ is a bounded subset of $[0, \infty)$, i.e. s is bounded on A.

PROOF: $\{x \in X : s(x) < 1\}$ is a $\mathcal{T}$-neighborhood of 0 in X, so absorbs A. 👍

There is a simple criterion for normability of a topological vector space:

XX.4.4.10. THEOREM. A topological vector space is normable if and only if it is Hausdorff and has a bounded convex neighborhood of 0.

PROOF: If X is a normable topological vector space, then X is Hausdorff and the open unit ball is a bounded convex neighborhood of 0.

Conversely, suppose X is Hausdorff and U is a bounded convex neighborhood of 0. There is a balanced neighborhood V of 0 contained in U, and the convex hull W of V is a balanced convex neighborhood of 0 contained in U since U is convex. Then the sets $\left\{\frac{1}{n}W : n \in \mathbb{N}\right\}$ form a base for the topology of X since W is bounded. If $x \in X$, $x \neq 0$, there is a neighborhood Z of 0 not containing x, and Z absorbs W, so there is an $\epsilon > 0$ with $x \notin \epsilon W$, i.e. $m_W(x) \geq \epsilon$. It follows that the seminorm m_W is a norm. If $W_n = \frac{1}{n}W$, then the m_{W_n} define the topology of X. But m_{W_n} is just nm_W, so the norm m_W determines the topology of X. 👍

XX.4.4.11. Note that convexity of the bounded neighborhood cannot be eliminated: ℓ^q $(0 < q < 1)$ is Hausdorff and has a neighborhood base of bounded open sets, but is not normable since it is not locally convex.

XX.4.4.12. It follows from (the easy direction of) XX.4.4.10. that if X is a normable topological vector space and $\mathcal{B}$ is a local base for the topology at 0, then $\mathcal{B}$ contains a bounded neighborhood. (**Caution:** this only applies to a *base*, not a subbase.) ✋

Nonnormability

We can then get a criterion which shows that many natural topological vector spaces are not normable:

XX.4.4.13. PROPOSITION. Let $(X, \mathcal{T})$ be a topological vector space, and (s_k) an increasing sequence of seminorms defining $\mathcal{T}$. Set $U_k = \{x \in X : s_k(x) < 1\}$. If for every k there is an $n > k$ such that s_n is unbounded on U_k, then $(X, \mathcal{T})$ is not normable. [If $\mathcal{T}$ is not defined by a sequence of seminorms, then $(X, \mathcal{T})$ is trivially not normable; if it is defined by a sequence of seminorms, we may assume the s_k are increasing by replacing s_k by the supremum or sum of the first k for each k.]

PROOF: The sets ϵU_k for $\epsilon > 0$ and $k \in \mathbb{N}$ form a local base for $\mathcal{T}$ at 0 since the s_k are increasing. If $(X, \mathcal{T})$ is normable, one of these is bounded, so for some ϵ and k, s_n is bounded on ϵU_k for all n by XX.4.4.9.. 👍

XX.4.4.14. EXAMPLE. Let $X = \mathcal{C}^\infty([a, b])$, with the seminorms (actually norms) p_k of XX.4.3.17.. For each k, it is easy to construct an $f \in \mathcal{C}^\infty([a, b])$ with $|f^{(r)}(x)| < 1$ for $0 \leq r \leq k$ and all $x \in [a, b]$, and $\max_{x \in [a,b]} |f^{(k+1)}(x)|$ arbitrarily large. Thus p_{k+1} is not bounded on $U_k = \{x \in X : p_k(x) < 1\}$ for any k, and therefore $\mathcal{C}^\infty([a, b])$ is not normable.

XX.4.4.15. Similarly, the examples from XX.4.3.21. (if all the X_k are nonzero), XX.4.3.22., and XX.4.3.23. are not normable.

XX.4.5. Continuous Linear Functionals and the Dual Space

In keeping with the basic philosophy of functional analysis, we want to describe and analyze topological vector spaces by means of their continuous linear functionals. This can be done satisfactorily in the (Hausdorff) locally convex case (and only in this case).

We will need to use the AC, or at least the Hahn-Banach Theorem, in this section. In fact, without the AC even the algebraic dual of a vector space can be $\{0\}$ ().

XX.4.5.1. DEFINITION. Let $(X, \mathcal{T})$ be a topological vector space. The *(topological) dual space* X^* of X is the set of all continuous linear functionals on X.

XX.4.5.2. The space X^* is a subspace of the algebraic dual X^d (). It is typically much smaller, but in good cases is large enough to reflect the TVS structure of X.

XX.4.5.3. EXAMPLES. (i) If X is a normed vector space, then X^* is the set of bounded linear functionals $\mathcal{B}(X, \mathbb{F})$ (). (So the notation is consistent.)

(ii) If $\mathcal{T}$ is the largest TVS topology on a vector space X (), then every linear functional on X is continuous. So $X^* = X^d$.

(iii) Let X be a vector space with the indiscrete topology. Then $X^* = \{0\}$.

XX.4.5.4. Because of XX.4.5.3.(iii), the theory of dual spaces is only good in the Hausdorff case. Thus for the rest of this section we will assume all topological vector spaces are Hausdorff.

The dual space can be degenerate even in the Hausdorff case:

XX.4.5.5. EXAMPLE. Fix q, $0 < q < 1$, and let $X = L^q([0,1])$. Then X is an F-space, and in particular Hausdorff. ($L^q([0,1])$ is not locally convex; in fact, see XX.4.5.10.).

We have that $L^1([0,1]) \subseteq L^q([0,1])$, and the 1-topology on $L^1([0,1])$ is stronger than the q-topology. Thus the restriction to $L^1([0,1])$ of any linear functional ϕ on $L^q([0,1])$ which is q-continuous is 1-continuous, hence is given by integration against an L^∞ function g, i.e.

$$\phi(f) = \int_0^1 fg\, d\lambda$$

for $f \in L^1([0,1])$ by the Riesz Representation Theorem ().

If p is the paranorm defining the topology of $L^q([0,1])$ (), and for $r > 0$

$$B_r^p = \{f \in L^q([0,1]) : p(f) \leq r\} = \left\{f \in L^q([0,1]) : \int_0^1 |f|^q\, d\lambda \leq r\right\}$$

is the closed ball of radius r around 0, we have that the B_r^p form a local base for the q-topology at 0, and $B_r^p = r^q B_1^p$ for any $r > 0$. It follows that if ϕ is a q-continuous linear functional on $L^q([0,1])$, then ϕ must be bounded on each B_r^p, and in particular on B_1^p.

Now suppose ϕ is a nonzero q-continuous linear functional on $L^q([0,1])$, and let g be the corresponding L^∞ function for the restriction of ϕ to $L^1([0,1])$. Then there is an $\epsilon > 0$ such that $A = \{t \in [0,1] : |g(t)| \geq \epsilon\}$ has $\lambda(A) > 0$. Let (A_n) be a sequence of measurable subsets of A with $\lambda(A_n) > 0$ for all n and $\lambda(A_n) \to 0$. Set

$$f_n = \frac{|g|}{g}[\lambda(A_n)]^{-1/q}\chi_{A_n} \in L^1([0,1]) \subseteq L^q([0,1])$$

for each n (this is well defined since $g \neq 0$ on A_n). Then

$$p(f_n) = \int_0^1 |f_n|^q\, d\lambda = \int_{A_n} [\lambda(A_n)]^{-1}\, d\lambda = 1$$

for all n. But

$$|\phi(f_n)| = \left|\int_0^1 g\frac{|g|}{g}[\lambda(A_n)]^{-1/q}\chi_{A_n}\, d\lambda\right| = \int_0^1 |g|[\lambda(A_n)]^{-1/q}\chi_{A_n}\, d\lambda \geq \epsilon[\lambda(A_n)]^{1-1/q} \to \infty$$

which is a contradiction.

Thus $L^q([0,1])$ has no nonzero q-continuous linear functionals. This argument works more generally in any (nonzero) finite, or even semifinite, measure space with no atoms, i.e. where every measurable set of positive measure contains measurable subsets of arbitrarily small (finite) positive measure. (Conversely, if $(X, \mathcal{A}, \mu)$ has an atom A, then $L^q(X, \mathcal{A}, \mu)$ has a one-dimensional topological direct summand spanned by χ_A and hence a nonzero continuous linear functional.)

However, a topological vector space which is not locally convex can have a lot of continuous linear functionals (although not "enough" – see ()):

XX.4.5.6. Example. Fix q, $0 < q < 1$, and let $X = \ell^q$. Then X is an F-space, and in particular Hausdorff. (ℓ^q is not locally convex; cf. ().)

We have $\ell^q \subseteq \ell^1$, and the q-topology is stronger than the 1-topology. Thus any 1-continuous linear functional is q-continuous. There are a lot of 1-continuous linear functionals, one for each sequence $(y_n) \in \ell^\infty$, given by

$$\phi((x_n)) = \sum_{n=1}^{\infty} x_n y_n \ .$$

In particular, the q-continuous linear functionals on ℓ^q separate points (and even more).

Actually, these are all the q-continuous linear functionals on ℓ^q. The linear combinations of the standard basis vectors $\boldsymbol{e}_n$ are dense in ℓ^q, so any q-continuous linear functional ϕ is determined by the sequence $(\phi(\boldsymbol{e}_n))$. But as in XX.4.5.5., every q-continuous linear functional is bounded on

$$B_1^p = \left\{ (x_n) \in \ell^q : \sum_{n=1}^{\infty} |x_n|^q \leq 1 \right\}$$

and $\boldsymbol{e}_n \in B_1^p$ for all n, so the sequence $(\phi(\boldsymbol{e}_n))$ is bounded, i.e. ϕ is 1-continuous, not just q-continuous. Thus there are not enough q-continuous linear functionals on ℓ^q to distinguish the q-topology from the 1-topology.

The main result of this section is:

XX.4.5.7. Theorem. [Geometrical Hahn-Banach Theorem] Let X be a topological vector space over $\mathbb{F}$, and A and B disjoint nonempty convex sets in X. If one of A, B has nonempty interior, then there is a nonzero continuous linear functional ϕ on X separating A and B. Equivalently in the real case, there is a closed hyperplane in X separating A and B.

Proof: Since an interior point is an internal point (), the theorem follows immediately from XX.6.4.6., or XX.6.4.9. in the complex case, and the following result. 👍

XX.4.5.8. Proposition. Let X be a topological vector space, and ϕ a linear functional on X. If $Re(\phi)$ is bounded above or below on a neighborhood of 0, then ϕ is continuous.

Proof: Let $Re(\phi)$ be bounded above on a neighborhood U of 0, i.e. there is a $c > 0$ such that $Re(\phi(x)) \leq c$ for all $c \in U$ (if $Re(\phi)$) is bounded below, replace ϕ by $-\phi$). There is a balanced neighborhood V of 0 contained in U. If $x \in V$, there is an $\alpha \in \mathbb{F}$, $|\alpha| = 1$, with $\phi(\alpha x) \geq 0$. Then $\alpha x \in V \subseteq U$, so

$$|\phi(x)| = |\alpha||\phi(x)| = |\phi(\alpha x)| = Re(\phi(\alpha x)) \leq c \ .$$

If $\epsilon > 0$, set $\beta = \frac{\epsilon}{c}$. Then βV is a neighborhood of 0 and if $x \in \beta V$, $|\phi(x)| \leq \epsilon$. 👍

XX.4.5.9. Corollary. Let X be a locally convex topological vector space over $\mathbb{R}$. Then the continuous linear functionals on X separate points of X, i.e. if $x, y \in X$, $x \neq y$, there is a continuous linear functional ϕ on X with $\phi(x) \neq \phi(y)$.

XX.4.5.10. Example. This is a continuation of (). The F-space $L^q([0,1])$ is not locally convex. In fact, if U is any neighborhood of 0, set $C = \overline{co}(U)$. If $C \neq X$ and $x \in X \setminus C$, then there is a nonzero continuous linear functional ϕ with $\phi(x) \geq \sup_{y \in C} \phi(y)$, a contradiction. Thus $C = X$, i.e. $co(U)$ is dense in X.

The next result is sometimes called *Mazur's Theorem*. It has a direct proof using Zorn's Lemma (consider a maximal subspace containing Y disjoint from C; cf. e.g. [SW99, II.3.1]), and can be used to prove the standard form of the Hahn-Banach Theorem XX.6.2.1. (Exercise ()).

XX.4.5.11. Corollary. Let X be a topological vector space over $\mathbb{R}$, Y an affine subspace, and C a nonempty convex open set in X disjoint from Y. Then there is a closed hyperplane H in X containing Y and disjoint from C.

Proof: There is a nonzero continuous linear functional ϕ separating Y and C. Then ϕ is bounded on Y. But a linear functional bounded on an affine subspace is constant on that subspace, i.e. ϕ is identically c on Y for some C. Set $H = \{x \in X : \phi(x) = c\}$. Since ϕ is not constant, $\phi(C)$ is open in $\mathbb{R}$, and contained in $[c, \infty)$, so is contained in (c, ∞) and C is disjoint from H. 👍

There is another separation theorem in the locally convex case, with compact subsets:

XX.4.5.12. Theorem. Let X be a locally convex topological vector space, and A and B disjoint nonempty closed convex subsets. If at least one of A, B is compact, then A and B are strictly separated.

Proof: Strictly separating A and B is the same as strictly separating 0 and $C = A - B$. C is closed by XXIII.1.2.20., and convex, and $0 \notin C$. Thus $X \setminus C$ is a neighborhood of 0. There is a balanced convex neighborhood V of 0 contained in $X \setminus C$ since X is locally convex, and there is a nonzero continuous linear functional ϕ with $\sup_{x \in V} Re(\phi(x)) \leq \inf_{y \in C} Re(\phi(y))$ by XX.4.5.7.. Then ϕ is not identically 0 on V, so since V is balanced there is an $z \in V$ with $\phi(z) > 0$. Then

$$\phi(0) = 0 < \phi(z) \leq \sup_{x \in V} Re(\phi(x)) \leq \inf_{y \in C} Re(\phi(y)) \ .$$

👍

XX.4.6. Duality and the Weak and Weak-* Topologies

We can define the weak topology on any topological vector space and the weak-* topology on its dual in a manner identical to the normed case. These topologies have reasonable properties only in the locally convex case.

The Weak Topology

XX.4.6.1. Let $(X, \mathcal{T})$ be a topological vector space, and X^* its topological dual. We define the *weak topology* $\mathcal{T}_w$ on X to be the topology of pointwise convergence for X^*, i.e. $x_i \to x$ if and only if $\phi(x_i) \to \phi(x)$ for all $\phi \in X^*$. The sets

$$U_{\{\phi_1,\ldots,\phi_n\},\epsilon} = \{x \in X : |\phi_k(x)| < \epsilon \text{ for } 1 \leq k \leq n\}$$

form a fundamental system of neighborhoods of 0 for $\mathcal{T}_w$.

XX.4.6.2. The weak topology $\mathcal{T}_w$ is clearly weaker than $\mathcal{T}$. If $\mathcal{T}$ is an LCS topology, so is $\mathcal{T}_w$. If $\mathcal{T}$ is normable, then $\mathcal{T}_w$ is the weak topology defined in ().

XX.4.6.3. Proposition. Let $(X, \mathcal{T})$ be a topological vector space. Then a linear functional is $\mathcal{T}_w$-continuous if and only if it is $\mathcal{T}$-continuous, i.e. the dual space $(X, \mathcal{T}_w)^*$ is the same as $X^* = (X, \mathcal{T})^*$ (as a subset of X^d).

This result is a special case of XX.4.6.11..

The Weak-* Topology

XX.4.6.4. Let $(X, \mathcal{T})$ be a topological vector space, and X^* its topological dual. We define the *weak-* topology* $\mathcal{T}_{w*}$ on X^* to be the topology of pointwise convergence for X, i.e. $\phi_i \to \phi$ if and only if $\phi_i(x) \to \phi(x)$ for all $x \in X$. The sets

$$U_{\{x_1,\ldots,x_n\},\epsilon} = \{\phi \in X^* : |\phi(x_k)| < \epsilon \text{ for } 1 \leq k \leq n\}$$

form a fundamental system of neighborhoods of 0 for $\mathcal{T}_{w*}$.

XX.4.6.5. If $\mathcal{T}$ is an LCS topology, so is $\mathcal{T}_{w*}$. If $\mathcal{T}$ is normable, then $\mathcal{T}_{w*}$ is the weak-* topology defined in (). (In the general case, this is the first topology we have defined on X^*.)

XX.4.6.6. Proposition. Let $(X, \mathcal{T})$ be a topological vector space. Then a linear functional on X^* is $\mathcal{T}_{w*}$-continuous if and only if it is of the form $\hat{x}$ for $x \in X$, where $\hat{x}(\phi) = \phi(x)$, i.e. the dual space $(X^*, \mathcal{T}_{w*})^*$ is X (regarded as a subset of X^{dd}).

This result is also a special case of XX.4.6.11..

Duality

This can all be put in a more general framework.

XX.4.6.7. Definition. Let X and Y be vector spaces over $\mathbb{F}$ (with no assumed topology). A *duality* between X and Y, denoted $\langle X, Y\rangle$ is a bilinear form $\langle \cdot, \cdot \rangle$ from $X \times Y$ to $\mathbb{F}$ which is separating in both X and Y, i.e. for fixed $x \in X$, $\langle x, y\rangle = 0$ for all $y \in Y$ implies $x = 0$ and for fixed $y \in Y$, $\langle x, y\rangle = 0$ for all $x \in X$ implies $y = 0$.

XX.4.6.8. Although the situation is completely symmetric in X and Y, notationally we consider the first space X as the basic space and elements of Y as giving linear functionals on X, i.e. $Y \subseteq X^d$.

XX.4.6.9. Example. If X is a locally convex topological vector space, the bilinear form $\langle x, \phi \rangle = \phi(x)$ gives a duality $\langle X, X^* \rangle$ between X and X^*. This can be written also in reverse order: $\langle \phi, x \rangle = \phi(x)$ gives a duality $\langle X^*, X \rangle$ between X^* and X.

XX.4.6.10. Let X and Y be vector spaces over $\mathbb{F}$, and $\langle X, Y \rangle$ a duality. The duality defines a topology $\sigma(X, Y)$ on X of pointwise convergence: $x_i \to x$ if and only if $\langle x_i, y \rangle \to \langle x, y \rangle$ for all $y \in Y$. The sets

$$U_{\{y_1,\dots,y_n\},\epsilon} = \{x \in X : |\langle x, y_k \rangle| < \epsilon \text{ for } 1 \leq k \leq n\}$$

form a fundamental system of neighborhoods of 0 for $\sigma(X, Y)$. This is an LCS topology on X.

Although the notation does not reflect it, the $\sigma(X, Y)$-topology depends on the bilinear form $\langle \cdot, \cdot \rangle$ and not just on X and Y. In applications there is generally a natural bilinear form which is understood.

For example, if X is a locally convex topological vector space, the $\sigma(X, X^*)$-topology is the weak topology on X, and the $\sigma(X^*, X)$-topology is the weak-* topology on X^*. The notations $\sigma(X, X^*)$ and $\sigma(X^*, X)$ are standard for these topologies.

Here is the general form of (). The proof is essentially identical.

XX.4.6.11. Proposition. Let X and Y be vector spaces over $\mathbb{F}$, and $\langle X, Y \rangle$ a duality. The $\sigma(X, Y)$-continuous linear functionals on X are precisely the linear functionals in Y (regarded as a subset of X^d). Thus $(X, \sigma(X, Y))^* = Y$. The topology $\sigma(X, Y)$ is the weakest LCS topology $\mathcal{T}$ on X for which $(X, \mathcal{T})^* = Y$.

XX.4.6.12. If $\langle X, Y \rangle$ is a duality, an LCS topology $\mathcal{T}$ on X is *consistent* for the duality if $(X, \mathcal{T})^* = Y$ (as a subset of X^d). Thus $\sigma(X, Y)$ is the weakest consistent topology. If $(X, \mathcal{T})$ is a locally convex topological vector space, the topology $\mathcal{T}$ is consistent for $\langle X, X^* \rangle$.

There is always a largest consistent topology $\tau(X, Y)$ for any duality $\langle X, Y \rangle$, called the *Mackey topology* of the duality. This topology is the topology of uniform convergence on all $\sigma(Y, X)$-compact convex balanced subsets of Y. If $(X, \mathcal{T})$ is a locally convex topological vector space, the Mackey topology on X is stronger than $\mathcal{T}$, and coincides with $\mathcal{T}$ in reasonable cases, e.g. if X is barrelled () or bornological (), and in particular if X is metrizable (first countable) [SW99, IV.3.2].

XX.4.7. Uniform Topologies on X^* and the Bidual

We give only a brief survey of the problem of defining a "uniform" topology on the dual of a locally convex topological vector space. For a full treatment see e.g. [SW99].

XX.4.7.1. If X is a normed vector space, besides the weak and weak-* topologies on the dual X^* there is a norm topology defined by the operator norm. It is a delicate and complicated matter to generalize the norm topology to general locally convex topological vector spaces, and there are several reasonable candidates and no obvious one, and none with all the properties desirable. (These topologies are often called *strong topologies* in the literature, but I prefer the term *uniform topologies* for reasons explained in ().)

Even worse, although the dual space X^* of a locally convex topological vector space $(X, \mathcal{T})$ is at least well defined as a set (subset of X^d) by $\mathcal{T}$, the bidual space $X^{**} = (X^*)^*$ is not even determined as a set (subset of $(X^*)^d$) by $\mathcal{T}$, but depends on what topology is chosen for the uniform topology on X^*. This is to be expected: the uniform topology on X^* should not generally be consistent for $\langle X^*, X\rangle$ (this even fails if X is a nonreflexive Banach space).

XX.4.7.2. We work more generally with a duality $\langle X, Y\rangle$. The general idea is that if $\mathcal{B}$ is a collection of subsets of Y with suitable properties (e.g. closed under finite unions), we can define on X the "topology of uniform convergence on sets in $\mathcal{B}$" by $x_i \to x$ if and only if $\langle x_i, y\rangle \to \langle x, y\rangle$ uniformly on B, for every $B \in \mathcal{B}$. The sets

$$U_{B,\epsilon} = \{x \in X : |\langle x, y\rangle| < \epsilon \text{ for all } y \in B\}$$

form a fundamental set of neighborhoods of 0. One can of course interchange X and Y.

For a uniform topology, there are several reasonable choices for $\mathcal{B}$:

All equicontinuous sets in Y.

All $\sigma(Y, X)$-compact subsets of Y.

All $\sigma(Y, X)$-bounded subsets of Y.

Even the weak topology fits into this picture, taking $\mathcal{B}$ to be the collection of finite subsets of Y, and the Mackey topology, taking $\mathcal{B}$ to be the set of $\sigma(Y, X)$-compact convex balanced subsets of Y.

XX.4.7.3. Fortunately, under mild hypotheses, all these topologies coincide, e.g. if X is barrelled () or bornological (), and in particular if X is metrizable (first countable).

XX.4.7.4. There is a standard choice: let $\beta(X, Y)$ be the topology of uniform convergence on all $\sigma(Y, X)$-bounded subsets of Y. We call this the *uniform topology* on X from $\langle X, Y\rangle$ (many references call it the *strong topology*).

In particular, if $(X, \mathcal{T})$ is a locally convex topological vector space, we call $\beta(X, X^*)$ the *uniform topology* on X, and $\beta(X^*, X)$ the *uniform topology* on X^*.

XX.4.7.5. If $(X, \mathcal{T})$ is a locally convex topological vector space, then the uniform topology $\beta(X, X^*)$ is stronger than $\mathcal{T}$. It can be strictly stronger, but the topologies coincide if $(X, \mathcal{T})$ is barrelled ().

The Bidual

XX.4.7.6. Definition. Let $(X, \mathcal{T})$ be a locally convex topological vector space. The *uniform bidual* of $(X, \mathcal{T})$ is the dual space $(X^*, \beta(X^*, X))^*$, denoted X^{**}.

XX.4.7.7. There is a natural embedding of X into X^{**}. One can put the uniform topology $\beta(X^{**}, X^*)$ on X^{**}, but the inclusion of X into X^{**} is not continuous in general for this topology. There is a weaker topology on X^{**}, the *natural topology*, the topology of uniform convergence on equicontinuous subsets of X^*; the inclusion of X into X^{**} is a homeomorphism for this topology. The uniform and natural topologies on X^{**} coincide in reasonable cases, e.g. if X is barrelled or bornological, and in particular if X is metrizable (first countable).

XX.4.8. Exercises

XX.4.8.1. Let $(X, \mathcal{T})$ be a topological vector space. Let $Y = \overline{\{0\}}$, and let Z be a subspace of X complementary to Y (III.11.2.38.; the AC is needed to insure the existence of such a Z). Then $X \cong Y \oplus Z = Y \times Z$ as a vector space (III.11.2.36.).

(a) Show that the restriction to Z of the quotient map from X to X/Y is a linear homeomorphism, where Z has the relative topology and X/Y the quotient topology. [See XX.4.1.12..]

(b) Show that X is linearly homeomorphic to $Y \oplus Z = Y \times Z$ with the product topology (the relative topology on Y is the indiscrete topology). Thus every topological vector space is (uniquely) the direct sum (product) of a Hausdorff topological vector space and a vector space with the indiscrete topology.

(c) Although Y is closed in X by definition, Z is not closed in X if X is not Hausdorff; in fact, Z is dense in X. (If X is Hausdorff, then $Y = \{0\}$ and $Z = X$.)

(d) The result in (a) is not true in general if Y is just a closed subspace of X and Z is a complementary subspace: the restriction to Z of the quotient map from X to X/Y is a continuous linear bijection, but not a linear homeomorphism in general. [Consider the case where X is a Banach space and Z is not closed in X.] The result in (a) (and also (b)) is true if Z is closed in X.

XX.4.8.2. Let $\{(X_i, \mathcal{T}_i) : i \in I\}$ be an inductive system () of topological vector spaces, and $(X, \mathcal{T})$ the inductive limit (). The important case is where X is a vector space, the X_i are subspaces directed by inclusion, each with a topology $\mathcal{T}_i$, such that whenever $i < j$, X_i is a closed subspace of X_j and the relative $\mathcal{T}_j$-topology on X_i is $\mathcal{T}_i$, and $X = \cup_i X_i$; such an inductive limit is called *strict*.

(a) Show that $(X, \mathcal{T})$ has a natural structure as a topological vector space, and that $\mathcal{T}$ is the strongest TVS topology on X making the natural maps from the X_i to X continuous. If each $\mathcal{T}_i$ is Hausdorff and the connecting maps are injective, $\mathcal{T}$ is Hausdorff. [Replace "Hausdorff" by "T_1".]

(b) In the strict case, show that the relative $\mathcal{T}$-topology on X_i is $\mathcal{T}_i$ for all i.

(c) In the strict case with $I = \mathbb{N}$, show that if (x_n) is a sequence in X, then $x_n \to x$ in the $\mathcal{T}$-topology if and only if there is an i such that x and all the x_n are contained in $\mathcal{T}_i$ and $x_n \to x$ in the $\mathcal{T}_i$-topology. (This is true only for *sequences*, and not nets in general. Is it true for sequences for general I?)

(d) Show that $(X, \mathcal{T})$ is not metrizable if I is infinite and the images of the X_i in X are distinct. If $\mathcal{T}$ is Hausdorff, it is not first countable.

(e) If X is any vector space, then X has a strongest TVS topology, and it is the inductive limit topology from its finite-dimensional subspaces (directed by inclusion) with their unique Hausdorff TVS topologies. If X is infinite-dimensional, this topology is Hausdorff and not first countable (it is of course the unique Hausdorff TVS topology if X is finite-dimensional).

XX.4.8.3. [DS88, II.1.12] Let p be a complete weak paranorm (XX.4.3.14.) on a vector space X. Use the Uniform Boundedness Theorem () to show that p is a paranorm, i.e. that scalar multiplication is jointly continuous.

XX.4.8.4. A subset B of a topological vector space $(X, \mathcal{T})$ is a *barrel* if it is closed, radial, balanced, and convex. The space X is *barreled* if every barrel in X is a neighborhood of 0.

(a) Show that every topological vector space which is a Baire space is barreled. In particular, every F-space (e.g. every Fréchet space, hence every Banach space) is barreled.

(b) Give c_{00} () the topology from the supremum norm (i.e. the relative topology from c_0). Show that c_{00} with this norm is not barreled. [Let B be the subset

$$\left\{(x_n) \in c_{00} : |x_n| \leq \frac{1}{n} \text{ for all } n\right\} .$$

Then B is a barrel in c_{00}. Note that the closure of B in c_0 is not radial.] Thus a normed vector space need not be barreled, and a subspace of a barreled space need not be barreled. Even a closed subspace of a barreled space need not be barreled [SW99, Exercise IV.10].

(c) The topological vector space $X = L^q([0,1])$ $(0 < q < 1)$ is barreled by (a), but contains no barrels other than X itself (so is vacuously barreled). [A nontrivial barrel gives rise to at least one nonzero continuous linear functional.]

(d) Show that any barrel B in a topological vector space X absorbs any subset A of X which is balanced, convex, bounded, and complete. [Consider the norm m_A on $\cup_n nA$.]

(e) A barreled vector space need not be bornological (XX.4.8.5.) ([**?**], [**?**]).

XX.4.8.5. A locally convex topological vector space $(X, \mathcal{T})$ is *bornological* if every balanced convex subset of X which absorbs every bounded subset of X is a neighborhood of 0. (This condition is really only interesting if $\mathcal{T}$ is Hausdorff. For example, any vector space with the indiscrete topology is bornological, since every subset is bounded.)

(a) Show that every metrizable locally convex topological vector space is bornological.

(b) Show that if $(X, \mathcal{T})$ is a locally convex topological vector space, there is a strongest locally convex TVS topology $\mathcal{T}_0$ with the same bounded sets as $\mathcal{T}$, and $(X, \mathcal{T}_0)$ is bornological. $\mathcal{T}_0 = \mathcal{T}$ if and only if $(X, \mathcal{T})$ is bornological.

(c) Let $(X, \mathcal{T})$ be a bornological space, Y a locally convex topological vector space, and $T : X \to Y$ a linear transformation. Show that the following are equivalent:

(i) T is continuous.

(ii) T is sequentially continuous, i.e. if $x_n \to x$, then $Tx_n \to Tx$.

(iii) If $x_n \to 0$, then $Tx_n \to 0$.

(iv) T sends bounded sets in X to bounded sets in Y.

(We have (i) $\Rightarrow$ (ii) $\Rightarrow$ (ii) $\Rightarrow$ (iv) if X and Y are any topological vector spaces. The bornological condition is only needed for (iv) $\Rightarrow$ (i).]

(d) Let $(X, \mathcal{T})$ be a locally convex topological vector space. Show that the following are equivalent:

(i) $(X, \mathcal{T})$ is bornological.

(ii) Any seminorm on X which is bounded on bounded sets is continuous.

(iii) For any locally convex topological vector space Y and linear $T : X \to Y$ sending bounded sets in X to bounded sets in Y, T is continuous.

If X is Hausdorff, these are equivalent to

(iv) $(X, \mathcal{T})$ is the inductive limit of a family of normed vector spaces.

XX.5. Bounded Operators

When considering normed vector spaces, or more generally topological vector spaces, the most natural and important functions between them should preserve both the algebraic structure, i.e. be linear, and the topological structure, i.e. be continuous. Thus we usually want to work with continuous linear operators. In the case of normed vector spaces, continuity is equivalent to another property, boundedness, which is usually easier to work with directly.

XX.5.1. Bounded Operators on Normed Vector Spaces

XX.5.1.6. DEFINITION. Let X and Y be vector spaces over $\mathbb{F}$. A *(linear) operator* from X to Y is a linear transformation () from X to Y.

Thus *operator* and *linear transformation* are synonymous. The term *operator* is primarily used in the case $X = Y$, in which case we call it an operator on X. If $T : X \to Y$ is an operator, we usually write Tx instead of $T(x)$ for the image of $x \in X$ under T.

XX.5.1.7. DEFINITION. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces over $\mathbb{F}$, and $T : X \to Y$ an operator (linear transformation). Then T is *bounded*, or a *bounded operator*, if there is a constant M such that

$$\|Tx\|_Y \leq M\|x\|_X$$

for all $x \in X$. The smallest such M is called the *operator norm* of T, written $\|T\|_{op}$ or usually just $\|T\|$ when there is no possibility of confusion.

Although the notation does not reflect it, boundedness of T and its operator norm depend on the choices of norms on X and Y. Thus, although we often write $T : X \to Y$ by slight abuse of notation, there must be understood norms on X and Y.

XX.5.1.8. If T is a bounded operator from X to Y, we have

$$\|T\| = \sup_{x \neq 0} \frac{\|Tx\|_Y}{\|x\|_X} = \sup_{\|x\|_X = 1} \|Tx\|_Y$$

since T is linear and any vector is a scalar multiple of a unit vector. The 0 operator is bounded, and $\|0\| = 0$; if T is bounded, then $\|T\| \geq 0$, and $\|T\| = 0$ if and only if $T = 0$.

The most important feature of boundedness is that it is equivalent to continuity:

XX.5.1.9. PROPOSITION. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces over $\mathbb{F}$, and $T : X \to Y$ a linear transformation. The following are equivalent:

(i) T is continuous at 0.

(ii) T is continuous.

(iii) T is uniformly continuous.

(iv) T is bounded.

PROOF: (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) is trivial.

(iv) $\Rightarrow$ (iii): Suppose T is bounded. We may assume $T \neq 0$, so $\|T\| > 0$. Let $\epsilon > 0$, and set $\delta = \frac{\epsilon}{\|T\|}$. If $x_1, x_2 \in X$ with $\|x_2 - x_1\|_X < \delta$, then

$$\|Tx_2 - Tx_1\|_Y = \|T(x_2 - x_1)\|_Y \leq \|T\|\|x_2 - x_1\|_X < \|T\|\delta = \epsilon \ .$$

Thus T is uniformly continuous.

(i) $\Rightarrow$ (iv): Suppose T is not bounded. Then for every n there is a $v_n \in X$, $v_n \neq 0$, with $\|Tv_n\|_Y > n\|v_n\|_X$. Set $x_n = \frac{1}{n\|v_n\|_X} v_n$. Then $\|x_n\|_X = \frac{1}{n}$ and

$$\|Tx_n\|_Y = \left\| T\left(\frac{1}{n\|v_n\|_X} v_n\right)\right\|_Y = \left\|\frac{1}{n\|v_n\|_X} Tv_n\right\|_Y = \frac{1}{n\|v_n\|_X}\|Tv_n\|_Y > 1$$

for all n. Thus $x_n \to 0$ in X but $Tx_n \not\to 0$ in Y, so T is not continuous at 0. 👍

XX.5.1.10. Let X and Y be vector spaces over $\mathbb{F}$. The set $\mathcal{L}(X, Y)$ of linear transformations from X to Y is a vector space over $\mathbb{F}$ under the usual operations ().

XX.5.1.11. PROPOSITION. Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be normed vector spaces over $\mathbb{F}$, and S and T bounded operators from X to Y. Then

(i) $S + T$ is a bounded operator from X to Y, and $\|S + T\| \leq \|S\| + \|T\|$.

(ii) If $\alpha \in \mathbb{F}$, then αT is bounded and $\|\alpha T\| = |\alpha|\|T\|$.

PROOF: (i): If $x \in X$, we have

$$\|(S + T)x\|_Y = \|Sx + Tx\|_Y \leq \|Sx\|_Y + \|Tx\|_Y \leq \|S\|\|x\|_X + \|T\|\|x\|_X = (\|S\| + \|T\|)\|x\|_X \ .$$

(ii): We have

$$\|\alpha T\| = \sup_{\|x\|_X = 1} \|(\alpha T)x\|_Y = \sup_{\|x\|_X = 1} \|\alpha(Tx)\|_Y = \sup_{\|x\|_X = 1} |\alpha|\|Tx\|_Y = |\alpha| \sup_{\|x\|_X = 1} \|Tx\|_Y = |\alpha|\|T\| \ .$$

👍

XX.5.1.12. COROLLARY. Let X and Y be normed vector spaces over $\mathbb{F}$. Then the set $\mathcal{B}(X, Y)$ of bounded operators from X to Y is a vector subspace of $\mathcal{L}(X, Y)$, and the operator norm is a norm on $\mathcal{B}(X, Y)$.

The operator norm has one more simple but important property:

XX.5.1.13. PROPOSITION. Let X, Y, Z be normed vector spaces over $\mathbb{F}$, $T \in \mathcal{B}(X,Y)$, $S \in \mathcal{B}(Y,Z)$. Then $ST = S \circ T \in \mathcal{B}(X,Z)$ and $\|ST\| \leq \|S\|\|T\|$.

PROOF: If $x \in X$, we have

$$\|(ST)x\|_Z = \|S(Tx)\|_Z \leq \|S\|\|Tx\|_Y \leq \|S\|\|T\|\|x\|_X \ .$$

Simple examples with 2×2 matrices show that the inequality can be strict; if fact, we can have $ST = 0$ even if S and T are both nonzero, if the range of T is contained in the null space of S.

Completeness

XX.5.1.14. THEOREM. Let X and Y be normed vector spaces over $\mathbb{F}$. If Y is complete (a Banach space), then $\mathcal{B}(X,Y)$ is complete (a Banach space) in the operator norm.

PROOF: The proof follows the standard pattern of completeness proofs. Suppose (T_n) is a Cauchy sequence in $\mathcal{B}(X,Y)$. We first find a candidate for the limit T.

Fix $x \in X$. We have, for $n, m \in \mathbb{N}$,

$$\|T_n x - T_m x\|_Y = \|(T_n - T_m)x\|_Y \leq \|T_n - T_m\|\|x\|_X$$

so the sequence $(T_n x)$ is a Cauchy sequence in Y. Since Y is complete, the sequence $(T_n x)$ converges to an element of Y we may call Tx. As x varies over X, this procedure defines a function $T : X \to Y$, and $T_n \to T$ pointwise.

We must show that T is a bounded operator and that $T_n \to T$ in operator norm. If $x_1, x_2 \in X$, then $T_n x_1 \to T x_1$, $T_n x_2 \to T x_2$, and $T_n(x_1 + x_2) \to T(x_1 + x_2)$. But $T_n(x_1 + x_2) = T_n x_1 + T_n x_2$ by linearity, and $T_n x_1 + T_n x_2 \to T x_1 + T x_2$ by continuity of addition in Y. Thus $T(x_1 + x_2) = T x_1 + T x_2$ by uniqueness of limits in Y. Similarly, if $\alpha \in F$, we have $T_n(\alpha x) \to T(\alpha x)$ and $T_n(\alpha x) = \alpha(T_n x) \to \alpha(Tx)$, so $T(\alpha x) = \alpha(Tx)$. Thus T is linear.

To show that T is bounded, note that the sequence (T_n), being a Cauchy sequence, is bounded, say $\|T_n\| \leq M$ for all n. Fix x with $\|x\| = 1$. Then $\|T_n x\|_Y \leq M\|x\|_X = M$, and since $T_n x \to Tx$, we also have $\|Tx\|_Y \leq M$. This is true for every x of norm 1, so T is bounded and $\|T\| \leq M$.

Finally, let $\epsilon > 0$. Since (T_n) is a Cauchy sequence, there is an N such that $\|T_n - T_m\| < \epsilon$ for all $n, m \geq N$. Fix $x \in X$, $\|x\|_X = 1$. Then

$$\|T_n x - T_m x\|_Y = \|(T_n - T_m)x\|_Y \leq \|T_n - T_m\|\|x\|_X < \epsilon$$

for all $n, m \geq N$. Fix $n \geq N$; since $T_m x \to Tx$, we have $\|(T_n - T)x\|_Y = \|T_n x - Tx\|_Y \leq \epsilon$. This is true for all such x, so $\|T_n - T\| \leq \epsilon$. This is true for all $n \geq N$, so $T_n \to T$ in operator norm.

XX.5.1.15. Note that completeness of X is not needed for this result. But if Y is complete, any bounded operator from X to Y extends uniquely to a bounded operator from the completion $\bar{X}$ to Y with the same norm, so $\mathcal{B}(X,Y) \cong \mathcal{B}(\bar{X},Y)$:

XX.5.1.16. PROPOSITION. Let X and Y be normed vector spaces, with Y complete, and let Z be a dense subspace of X. If T is a bounded operator from Z to Y, then T extends uniquely to a bounded operator $\bar{T} : X \to Y$, and $\|\bar{T}\| = \|T\|$.

PROOF: Since T is uniformly continuous (XX.5.1.9.) and Y is complete, T extends uniquely to a continuous function $\bar{T}$ from X to Y (). If $x_1, x_2 \in X$, there are sequences (z_{n1}) and (z_{n2}) in Z converging to x_1 and x_2 respectively. Then $z_{n1} + z_{n2} \to x_1 + x_2$, so $\bar{T}(z_{n1} + z_{n2}) = T(z_{n1} + z_{n2}) \to \bar{T}(x_1 + x_2)$. On the other hand, $\bar{T}z_{n1} = Tz_{n1} \to \bar{T}x_1$ and $\bar{T}z_{n2} = Tz_{n2} \to \bar{T}x_2$, so

$$\bar{T}(z_{n1} + z_{n2}) = T(z_{n1} + z_{n2}) = Tz_{n1} + Tz_{n2} \to \bar{T}x_1 + \bar{T}x_2$$

so $\bar{T}(x_1 + x_2) = \bar{T}x_1 + \bar{T}x_2$ by uniqueness of limits in Y. Similarly, if $\alpha \in \mathbb{F}$, $\bar{T}(\alpha z_{n1})$ converges to both $\bar{T}(\alpha x_1)$ and $\alpha(\bar{T}x_1)$, so $\bar{T}(\alpha x_1) = \alpha(\bar{T}x_1)$ and $\bar{T}$ is linear.

Let $x \in X$, $\|x\|_X = 1$. There is a sequence (z_n) in Z converging to x. Then $\|z_n\|_X \to \|x\|_X = 1$, so if $w_n = \frac{z_n}{\|z_n\|_X}$, then $w_n \in Z$ for all n, $\|w_n\|_X = 1$ for all n, and $w_n \to x$ by joint continuity of scalar multiplication. We have $\|Tw_n\|_Y \leq \|T\|$ for all n, and since $\bar{T}w_n = Tw_n \to \bar{T}x$, we also have $\|\bar{T}x\|_Y \leq \|T\|$. This is true for all such x, so $\bar{T}$ is bounded and $\|\bar{T}\| \leq \|T\|$.

For the opposite inequality, let $\epsilon > 0$. There is a $z \in Z$ with $\|z\|_X = 1$ and $\|Tz\|_Y > \|T\| - \epsilon$. We then have

$$\|T\| - \epsilon < \|Tz\|_Y = \|\bar{T}z\|_Y \leq \|\bar{T}\|\|z\|_X = \|\bar{T}\| \ .$$

The extreme inequality is true for all $\epsilon > 0$, and thus $\|T\| \leq \|\bar{T}\|$. 👍

XX.5.2. Bounded Linear Functionals

XX.5.2.17. DEFINITION. Let X be a vector space over $\mathbb{F}$. A *linear functional* on X is a linear transformation from X to $\mathbb{F}$. If X is a normed vector space, a linear functional ϕ on X is *bounded* if it is bounded as a linear transformation (regarding $\mathbb{F}$ as a normed vector space with norm $|\cdot|$), i.e. if there is a constant M such that $|\phi(x)| \leq M\|x\|$ for all $x \in X$. The smallest such M is the *norm* of ϕ, denoted $\|\phi\|$.

XX.5.2.18. **Notation:** If X is a vector space over $\mathbb{F}$, we will write X' for the set $\mathcal{L}(X, \mathbb{F})$ of linear functionals on X. If X is normed, write X^* for the set $\mathcal{B}(X, \mathbb{F})$ of bounded linear functionals on X. Then X' is a vector space over $\mathbb{F}$, and by XX.5.1.12., X^* is a subspace which is normed by $\|\cdot\|$; X^* is a Banach space by XX.5.1.14.. X' is called the *algebraic dual space* of X, and X^* the *(topological) dual space* of X.

There is no uniformity of notation for dual spaces. Many authors use X' for the topological dual space, and some of these (e.g. [**?**]) even use X^* to denote the algebraic dual. We use what seems to be the most common notation. We usually use ϕ, ψ, etc. to denote linear functionals; some authors use x^* to denote a general element of X^*, leading to poor and confusing expressions like $x^*(x)$ where x denotes a general element of X and x^* a general element of X^*, with no relation between them. ✋

XX.5.2.19. By XX.5.1.9., a linear functional on a normed vector space is continuous if and only if it is bounded. The null space of any bounded operator is closed (). The converse does not hold in general, e.g. there are unbounded one-to-one linear transformations. However:

XX.5.2.20. PROPOSITION. Let X be a normed vector space over $\mathbb{F}$, and $\phi : X \to \mathbb{F}$ a linear functional on X. If the null space N_ϕ is closed in X, then ϕ is bounded.

PROOF: Suppose ϕ is unbounded. Then for each $n \in \mathbb{N}$ there is a $y_n \in X$ with $\|y_n\| = 1$ but $|\phi(y_n)| > n$. Set $x_n = \frac{1}{\phi(y_n)} y_n$. Then $\|x_n\| < 1/n$ and $\phi(x_n) = 1$ for all n. Fix $v \in X$ with $\phi(v) = 1$. We have $x_n \to 0$, so $v - x_n \to v$. But $v - x_n \in N_\phi$ for all n, so $v \in \overline{N_\phi} \setminus N_\phi$, and N_ϕ is not closed. 👍

In fact, this argument shows that if ϕ is unbounded, then N_ϕ is dense in X. (Actually it would have to be either closed or dense since it has codimension 1.)

XX.5.3. Operators on Finite-Dimensional Normed Spaces

XX.5.4. Bounded Operators on Hilbert Spaces

XX.5.5. Algebras of Operators

XX.5.5.1. DEFINITION. A *(concrete) operator algebra* is a *-subalgebra of $\mathcal{B}(\mathcal{H})$ which is topologically closed in a suitable sense (there is also a subject of non-self-adjoint operator algebras, but it is outside the scope of this volume.) A *von Neumann algebra* is a (necessarily unital) *-subalgebra M of $\mathcal{B}(\mathcal{H})$ such that $M = M''$.

A von Neumann algebra is weakly (hence strongly, norm-, $\cdots$) closed, and in particular is a concrete C*-algebra (XX.16.1.3.), albeit of a very special kind.

Commutant and Bicommutant

One of the first, yet still one of the most crucial basic theorems of operator algebra theory is VON NEUMANN'S Bicommutant Theorem [**?**]. We say a *-algebra A of operators on a Hilbert space $\mathcal{H}$ *acts nondegenerately* if $T\xi = 0$ for all $T \in A$ implies $\xi = 0$. Since A is a *-algebra, this is equivalent to the condition that the subspace

$$A\mathcal{H} = span\{T\xi : T \in A, \xi \in \mathcal{H}\}$$

is dense in $\mathcal{H}$. If $I \in A$, then A obviously acts nondegenerately.

XX.5.5.2. THEOREM. [BICOMMUTANT THEOREM] Let A be a *-subalgebra of $\mathcal{B}(\mathcal{H})$ acting nondegenerately. Then A is σ-strongly dense in A''.

PROOF: Using **??**, the proof reduces to showing that if $\xi \in \mathcal{H}$ and $T \in A''$, there is a sequence (T_n) in A with $T_n\xi \to T\xi$. If $\mathcal{X} = \{S\xi : S \in A\}^-$, then $P_\mathcal{X} \in A'$ (since both $\mathcal{X}$ and $\mathcal{X}^\perp$ are invariant under A), so T commutes with $P_\mathcal{X}$, i.e. T leaves $\mathcal{X}$ invariant. It remains to show that $\xi \in \mathcal{X}$, so that $T\xi \in \mathcal{X}$. This is trivial if $I \in A$. In the general nondegenerate case, for each $S \in A$,

$$S[(I - P_\mathcal{X})\xi] = (I - P_\mathcal{X})[S\xi] = 0$$

since $S\xi \in \mathcal{X}$, so $(I - P_\mathcal{X})\xi = 0$. 👍

The Bicommutant Theorem relates a topological closure property with the simple, natural, and purely algebraic property of being the bicommutant (or just commutant) of a *-closed set of operators.

XX.5.5.3. Corollary. Let A be a unital (or just nondegenerate) *-subalgebra of $\mathcal{B}(\mathcal{H})$. The following are equivalent:

(i) A is strongly closed.

(ii) A is σ-strongly closed.

(iii) A is weakly closed.

(iv) A is σ-weakly closed.

(v) A is a von Neumann algebra, i.e. A''.

XX.6. The Hahn-Banach Theorem

The Hahn-Banach Theorem is one of the most fundamental results of functional analysis. In various forms it asserts the existence of "many" linear functionals on real or complex vector spaces. Indeed, this theorem underlies and justifies the whole philosophy of functional analysis: that a space can be completely studied via the scalar-valued functionals on the space.

There are many slightly different forms in which this theorem is commonly expressed. All the general versions require use of some form of the Axiom of Choice, but versions for vector spaces which are "not too large" can be proved directly without using the AC.

XX.6.1. Sublinear Functionals and Version 1

The first version of the Hahn-Banach Theorem concerns existence of linear functionals on a real vector space which are trapped between a sublinear functional and a superlinear functional, and more generally extensions to the whole vector space of such a linear functional already defined on a subspace.

XX.6.1.1. DEFINITION. Let V be a real vector space. A function $p : V \to (-\infty, +\infty]$ is a *sublinear functional* on V if

(i) $p(x + y) \leq p(x) + p(y)$ for all $x, y \in V$.

(ii) $p(\alpha x) = \alpha p(x)$ for all $\alpha > 0$.

A function $q : V \to [-\infty, +\infty)$ is a *superlinear functional* on V if

(i) $q(x + y) \geq q(x) + q(y)$ for all $x, y \in V$.

(ii) $q(\alpha x) = \alpha q(x)$ for all $\alpha > 0$.

A sublinear or superlinear functional taking only real values is called *finite*.

Note that a sublinear functional can take the value $+\infty$ but not $-\infty$, and a superlinear functional can take the value $-\infty$ but not $+\infty$. In most applications, sublinear or superlinear functionals are finite.

XX.6.1.2. EXAMPLES. (i) Any linear functional is both sublinear and superlinear.

(ii) The function $p(x) = +\infty$ for all x is sublinear; the function $q(x) = -\infty$ for all x is superlinear.

(iii) Any seminorm () is a finite sublinear functional taking only nonnegative values.

(iv) Let C be a convex set in V containing 0. For $x \in V$, set

$$m_C(x) = \inf\{\alpha > 0 : \alpha^{-1}x \in C\}$$

(where $\inf(\emptyset) = +\infty$). Then m_C is a sublinear functional on V, called the *Minkowski functional* of C (). It takes only nonnegative values, and is finite if C is radial.

(v) If p is a sublinear functional, then $-p$ is a superlinear functional, and vice versa.

The first version of the Hahn-Banach Theorem is:

XX.6.1.3. Theorem. [Hahn-Banach Theorem, Version 1] Let V be a real vector space, p a sublinear functional on V, and q a superlinear functional on V satisfying $q(x) \leq p(x)$ for all $x \in V$. Assume at least one of p and q is finite. Let W be a subspace of V, and ψ a linear functional on W satisfying $q(y) \leq \psi(y) \leq p(y)$ for all $y \in W$. Then there is a linear functional ϕ on V extending ψ with

$$q(x) \leq \phi(x) \leq p(x)$$

for all $x \in V$.

Taking $q(x) = -\infty$ for all x, we get the following commonly-stated version of the theorem:

XX.6.1.4. Corollary. [Hahn-Banach Theorem, Version 1A] Let V be a real vector space, and p a finite sublinear functional on V. Let W be a subspace of V, and ψ a linear functional on W satisfying $\psi(y) \leq p(y)$ for all $y \in W$. Then there is a linear functional ϕ on V extending ψ with

$$\phi(x) \leq p(x)$$

for all $x \in V$.

There is an analogous result with inequalities reversed for a superlinear functional.

XX.6.1.5. If p is a sublinear functional on V, then $p(0) = \alpha p(0)$ for all $\alpha > 0$. Thus $p(0)$ is either 0 or $+\infty$. Similarly, if q is superlinear, then $q(0)$ is either 0 or $-\infty$. Thus, if p is sublinear and q superlinear, then $q(0) \leq 0 \leq p(0)$, so the hypotheses of XX.6.1.3. and XX.6.1.4. are automatically satisfied if W is the zero subspace and ψ the zero functional. Thus we obtain:

XX.6.1.6. Corollary. Let V be a real vector space, p a sublinear functional on V, and q a superlinear functional on V satisfying $q(x) \leq p(x)$ for all $x \in V$. Assume at least one of p and q is finite. Then there is a linear functional ϕ on V with

$$q(x) \leq \phi(x) \leq p(x)$$

for all $x \in V$.

There are also one-sided versions of this result.

XX.6.1.7. The proof of Theorem XX.6.1.3. consists of two steps:

(1) Show that ψ can be extended to a subspace spanned by W and one additional element.

(2) "Iterate" step 1 to extend ψ all the way to V.

The first step is elementary and straightforward, but takes some work. The second step in general requires some form of the Axiom of Choice.

We begin with step 1.

XX.6.1.8. LEMMA. Let V be a real vector space, p a sublinear functional on V, and q a superlinear functional on V satisfying $q(x) \leq p(x)$ for all $x \in V$. Assume at least one of p and q is finite. Let W be a subspace of V, and ψ a linear functional on W satisfying $q(y) \leq \psi(y) \leq p(y)$ for all $y \in W$. Let $x_0 \in V \setminus W$, and let X be the subspace of V spanned by W and x_0. Then there is a linear functional θ on X extending ψ with

$$q(x) \leq \theta(x) \leq p(x)$$

for all $x \in X$.

PROOF: Every $x \in X$ can be uniquely written as $y + \alpha x_0$, where $y \in W$ and $\alpha \in \mathbb{R}$. Any linear functional on X extending ψ is of the form

$$\theta(y + \alpha x_0) = \psi(y) + \alpha c$$

for some $c \in \mathbb{R}$, so we must choose c so that $q(x) \leq \theta(x) \leq p(x)$ for all $x \in X$, i.e.

$$q(y + \alpha x_0) \leq \psi(y) + \alpha c \leq p(y + \alpha x_0)$$

for all $y \in W$ and $\alpha \in \mathbb{R}$. To show this is possible, we obtain some inequalities.

First, if $y, z \in W$, we have

$$\psi(y) - \psi(z) = \psi(y - z) \leq p(y - z) = p([y + x_0] + [-x_0 - z]) \leq p(y + x_0) + p(-x_0 - z)$$

$$-p(-x_0 - z) - \psi(z) \leq p(y + x_0) - \psi(y) \tag{XX.1}$$

and similarly we obtain

$$q(y + x_0) - \psi(y) \leq -q(-x_0 - z) - \psi(z) \ . \tag{XX.2}$$

Finally, we have

$$\psi(y) - \psi(z) = \psi(y + x_0) - \psi(x_0 + z) \leq p(y + x_0) - q(x_0 + z)$$

$$q(x_0 + z) - \psi(z) \leq p(y + x_0) - \psi(y) \tag{XX.3}$$

and we similarly obtain

$$-p(-x_0 - z) - \psi(z) \leq -q(-x_0 - y) - \psi(y) \ . \tag{XX.4}$$

Now define

$$\sigma_1 = \sup\{-p(-x_0 - y) - \psi(y) : y \in W\}$$

$$\sigma_2 = \sup\{q(x_0 + y) - \psi(y) : y \in W\}$$

$$\tau_1 = \inf\{p(x_0 + y) - \psi(y) : y \in W\}$$

$$\tau_2 = \inf\{-q(-x_0 - y) - \psi(y) : y \in W\} \ .$$

We have $\sigma_1 \leq \tau_1$ by XX.1, $\sigma_2 \leq \tau_2$ by XX.2, $\sigma_2 \leq \tau_1$ by XX.3, and $\sigma_1 \leq \tau_2$ by XX.4. Thus

$$\max(\sigma_1, \sigma_2) \leq \min(\tau_1, \tau_2) \ .$$

If there is a $y \in W$ with $p(x_0 + y) < +\infty$, then τ_1 is not $+\infty$, so σ_1 and σ_2 are not $+\infty$; if there is a $y \in W$ with $p(-x_0 - y) < +\infty$, then σ_1 is not $-\infty$, so τ_1 and τ_2 are not $-\infty$. A similar statement holds for q. Thus, if p or q is finite, we have that $\max(\sigma_1, \sigma_2) \neq +\infty$ and $\min(\tau_1, \tau_2) \neq -\infty$.

Choose any $c \in \mathbb{R}$ such that

$$\max(\sigma_1, \sigma_2) \leq c \leq \min(\tau_1, \tau_2)$$

(to avoid making a choice, just set $c = \min(\tau_1, \tau_2)$). We will show that this c works, i.e. that $\theta(y + \alpha x_0) = \psi(y) + \alpha c$ satisfies

$$q(y + \alpha x_0) \leq \psi(y) + \alpha c \leq p(y + \alpha x_0)$$

for all $y \in W$ and $\alpha \in \mathbb{R}$.

Let $x \in X$. Write $x = y + \alpha x_0$. If $\alpha = 0$, then $x = y \in W$ and $q(x) \leq \theta(x) = \psi(x) \leq p(x)$ by assumption. If $\alpha > 0$, we have $\alpha^{-1} y \in W$, so

$$c \leq \tau_1 \leq p(x_0 + \alpha^{-1} y) - \psi(\alpha^{-1} y)$$

$$\alpha c \leq \alpha p(x_0 + \alpha^{-1} y) - \alpha \psi(\alpha^{-1} y) = p(\alpha x_0 + y) - \psi(y) = p(x) - \psi(y)$$

$$\theta(x) = \psi(y) + \alpha c \leq p(x) \ .$$

Similarly using σ_2 we obtain $q(x) \leq \theta(x)$.

If $\alpha < 0$, we similarly have

$$-p(-x_0 - \alpha^{-1} y) - \psi(\alpha^{-1} y) \leq \sigma_1 \leq c$$

$$-p(x) + \psi(y) = -p(\alpha x_0 + y) + \psi(y) = (-\alpha)[-p(-x_0 - \alpha^{-1} y)] - (-\alpha)\psi(\alpha^{-1} y) \leq -\alpha c$$

$$\theta(x) = \psi(y) + \alpha c \leq p(x) \ .$$

In the same manner using τ_2 we have $q(x) \leq \theta(x)$. 

XX.6.1.9. In the case of XX.6.1.3. where V/W is countably generated, we can iterate XX.6.1.8. inductively to extend ψ to all of V (the Countable AC is technically needed to do this; if V itself is countably generated, the Countable AC can be dispensed with since we can take c to have a definite formula at each step). The same construction can be done transfinitely if a set of generators for V/W can be well ordered; assuming the Well-Ordering Principle (i.e. assuming AC: cf. ()), the proof of the general version of XX.6.1.3. can be completed this way.

XX.6.1.10. We will give an alternate argument using Zorn's Lemma to finish the proof of XX.6.1.3. in the general case. Let $\mathcal{E}$ be the set of pairs (X, θ), where X is a subspace of V containing W and θ is a linear functional on X extending ψ, with $q(x) \leq \theta(x) \leq p(x)$ for all $x \in X$. Put a partial ordering $\precsim$ on $\mathcal{E}$ by setting $(X_1, \theta_1) \precsim (X_2, \theta_2)$ if and only if $X_1 \subseteq X_2$ and θ_2 extends θ_1.

If $\mathcal{C} = \{(X_i, \theta_i : i \in I\}$ is a chain in $\mathcal{E}$, set $X = \cup_i X_i$. If $x, y \in X$, then $x \in X_i$ and $y \in X_j$ for some i, j; but either $X_i \subseteq X_j$ or $X_j \subseteq X_i$ since $\mathcal{C}$ is a chain, so either $x, y \in X_i$ or $x, y \in X_j$; in either case $x + y \in X$, so X is closed under addition. X is obviously closed under scalar multiplication, so X is a subspace of V. Define $\theta : X \to \mathbb{R}$ by $\theta(x) = \theta_i(x)$ if $x \in X_i$; this function is well defined since the θ_i all agree wherever they are defined. It is routine to check that θ is linear. Thus (X, θ) is an upper bound for $\mathcal{C}$ in $\mathcal{E}$.

Thus Zorn's lemma applies to give a maximal element (X, θ) in $\mathcal{E}$. But we must have $X = V$, for otherwise if $x_0 \in V \setminus X$ Lemma XX.6.1.8. gives an extension of θ to the span of X and x_0 with the right properties, contradicting maximality of (X, θ). Then $\phi = \theta$ is the desired extension.

This completes the proof of Theorem XX.6.1.3.. 

XX.6.1.11. The proof we have given of the Hahn-Banach Theorem uses the AC. But the full AC is not necessary; one can get away with weaker assumptions such as the BPI (). Even this assumption is not necessary: the Hahn-Banach Theorem is logically equivalent (over ZF) to the Measure Extension Property (), which is strictly weaker than the BPI []. The Hahn-Banach Theorem is not a theorem of ZF, i.e. some form of Choice is necessary.

XX.6.2. Extension of Bounded Linear Functionals

One of the most common special cases of XX.6.1.4. is extension of bounded linear functionals from subspaces of normed vector spaces without increasing the norm:

XX.6.2.1. THEOREM. [HAHN-BANACH THEOREM, VERSION 2] Let V be a real vector space, and $\|\cdot\|$ a norm on V. Let W be a subspace of V, and ψ a bounded linear functional on W. Then there is a bounded linear functional ϕ on V extending ψ with $\|\phi\| = \|\psi\|$.

PROOF: Apply XX.6.1.4. with $p(x) = \|\psi\|\|x\|$. 👍

XX.6.2.2. Note that no completeness is assumed in this result, and the subspace W need not be closed in V. (This is not a significant observation, though, since a bounded linear functional can automatically be extended to the completion or closure without increase in norm by uniform continuity ().)

XX.6.2.3. COROLLARY. Let $V \neq \{0\}$ be a normed real vector space, and $x \in V$. Then there is a bounded linear functional ϕ of norm 1 on V with $\phi(x) = \|x\|$.

PROOF: Suppose $x \neq 0$. Let W be the one-dimensional subspace of V spanned by x. Define $\psi : W \to \mathbb{R}$ by $\psi(\alpha x) = \alpha\|x\|$. Then ψ is a bounded linear functional of norm 1 on W with $\psi(x) = \|x\|$. Extend ψ to a bounded linear functional of norm 1 on V.

If $x = 0$, let ϕ be any bounded linear functional on V of norm 1 (such a linear functional exists by the first part of the proof applied to any nonzero vector in V). 👍

XX.6.2.4. This result shows that if V is a normed vector space, then the dual space V^* () is "large" (large enough to determine the norm on V). Without this result it is not obvious that a normed vector space has any nonzero bounded linear functionals.

XX.6.2.5. Theorem XX.6.2.1. requires some version of the AC (XX.6.1.11.). However, if V is separable, or more generally contains a well-ordered total subset, no version of the AC is necessary. Thus this version of the Hahn-Banach Theorem holds for separable Banach spaces with no Choice needed.

The Canonical Embedding into the Second Dual

XX.6.2.6. If V is a normed (real) vector space, then V^* is a Banach space (). V^* itself has a dual space $V^{**} = (V^*)^*$. If $x \in V$, then there is a canonically associated $\hat{x} \in V^{**}$ defined by

$$\hat{x}(\phi) = \phi(x)$$

for $\phi \in V^*$. It is easily checked that $\hat{x}$ is a linear functional on V^*, and that $x \mapsto \hat{x}$ is linear in x. Since $|\hat{x}(\phi)| = |\phi(x)| \leq \|x\|\|\phi\|$ for any $\phi \in V^*$, $\hat{x}$ is bounded and $\|\hat{x}\| \leq \|x\|$. By XX.6.2.3., we actually have $\|\hat{x}\| = \|x\|$. So:

XX.6.2.7. Corollary. Let V be a normed real vector space. The map $x \mapsto \hat{x}$ is an isometric linear embedding of V into V^{**}, called the *canonical embedding*.

XX.6.2.8. V is said to be *reflexive* if the canonical map from V to V^{**} is surjective. Since V^{**} is a Banach space, if V is reflexive it must be complete (a Banach space). Some Banach spaces are reflexive (), () while others are not ().

XX.6.3. The Complex Hahn-Banach Theorem

The complex case of the Hahn=Banach Theorem is slightly different: since a nonzero complex-linear functional takes all values in $\mathbb{C}$, the statement of XX.6.1.3. does not make sense. But there is a simple modification, based on the following observation:

XX.6.3.1. Proposition. Let V be a vector space over $\mathbb{C}$, and ϕ a complex-linear functional on X. Let $\theta = Re(\phi)$. Then θ is a real-linear functional on V, and

$$\phi(x) = \theta(x) - i\theta(ix)$$

for all $x \in V$. Conversely, if θ is any real-linear functional on V, define ϕ by this formula; then ϕ is a complex-linear functional on V. Thus every real linear functional on V is the real part of a unique complex-linear functional.

Proof: if ϕ is complex-linear and θ and λ are the real and imaginary parts, so $\phi(x) = \theta(x) + i\lambda(x)$, we have

$$\theta(ix) + i\lambda(ix) = \phi(ix) = i\phi(x) = i\theta(x) + i^2\lambda(x) = -\lambda(x) + i\theta(x)$$

for $x \in V$, so $\lambda(x) = -\theta(ix)$. Conversely, if θ is a real-linear functional on V, and $\phi(x) = \theta(x) - i\theta(ix)$, then ϕ is real-linear, and for $x \in V$ we have

$$\phi(ix) = \theta(ix) - i\theta(i^2x) = \theta(ix) - i\theta(-x) = \theta(ix) + i\theta(x)$$

$$i\phi(x) = i\theta(x) - i^2\theta(x) = i\theta(x) + \theta(ix)$$

so ϕ is complex-linear. 👍

XX.6.3.2. Theorem. [Complex Hahn-Banach Theorem, Version 1] Let V be a complex vector space, p a sublinear functional on V, and q a superlinear functional on V satisfying $q(x) \leq p(x)$ for all $x \in V$. Assume at least one of p and q is finite. Let W be a subspace of V, and ψ a linear functional on W satisfying $q(y) \leq Re(\psi(y)) \leq p(y)$ for all $y \in W$. Then there is a linear functional ϕ on V extending ψ with

$$q(x) \leq Re(\phi(x)) \leq p(x)$$

for all $x \in V$.

Proof: Let $\lambda = Re(\psi)$, and extend λ to a real linear functional θ on V satisfying $q(x) \leq \theta(x) \leq p(x)$ for all $x \in V$ by XX.6.1.3.. Set $\phi(x) = \theta(x) - i\theta(ix)$. 👍

Taking $q(x) = -\infty$ for all x, we get the following commonly-stated version of the theorem:

XX.6.3.3. Corollary. [Complex Hahn-Banach Theorem, Version 1A] Let V be a complex vector space, and p a finite sublinear functional on V. Let W be a subspace of V, and ψ a linear functional on W satisfying $Re(\psi(y)) \leq p(y)$ for all $y \in W$. Then there is a linear functional ϕ on V extending ψ with

$$Re(\phi(x)) \leq p(x)$$

for all $x \in V$.

There is an analogous result with inequalities reversed for a superlinear functional.

XX.6.3.4. Corollary. Let V be a complex vector space, p a sublinear functional on V, and q a superlinear functional on V satisfying $q(x) \leq p(x)$ for all $x \in V$. Assume at least one of p and q is finite. Then there is a linear functional ϕ on V with

$$q(x) \leq Re(\phi(x)) \leq p(x)$$

for all $x \in V$.

There are also one-sided versions of this result.

XX.6.3.5. Theorem. [Complex Hahn-Banach Theorem, Version 2] Let V be a complex vector space, and $\|\cdot\|$ a norm on V. Let W be a subspace of V, and ψ a bounded linear functional on W. Then there is a bounded linear functional ϕ on V extending ψ with $\|\phi\| = \|\psi\|$.

Proof: Apply XX.6.3.3. with $p(x) = \|\psi\|\|x\|$. We have $Re(\phi(x)) \leq \|\psi\|\|x\|$ for all $x \in V$. For any $x \in V$, there is an $\alpha \in \mathbb{C}$, $|\alpha| = 1$, such that $\phi(\alpha x) \geq 0$. Thus

$$|\phi(x)| = |\alpha||\phi(x)| = |\phi(\alpha x)| = Re(\phi(\alpha x)) \leq \|\psi\|\|\alpha x\| = \|\psi\|\|x\|$$

so $\|\phi\| \leq \|\psi\|$, and the opposite inequality is obvious. 👍

XX.6.3.6. Note that no completeness is assumed in this result, and the subspace W need not be closed in V. (This is not a significant observation, though, since a bounded linear functional can automatically be extended to the completion or closure without increase in norm by uniform continuity ().)

XX.6.3.7. COROLLARY. Let $V \neq \{0\}$ be a normed complex vector space, and $x \in V$. Then there is a bounded linear functional ϕ of norm 1 on V with $\phi(x) = \|x\|$.

PROOF: Suppose $x \neq 0$. Let W be the one-dimensional subspace of V spanned by x. Define $\psi : W \to \mathbb{C}$ by $\psi(\alpha x) = \alpha\|x\|$. Then ψ is a bounded linear functional of norm 1 on W with $\psi(x) = \|x\|$. Extend ψ to a bounded linear functional of norm 1 on V.

If $x = 0$, let ϕ be any bounded linear functional on V of norm 1 (such a linear functional exists by the first part of the proof applied to any nonzero vector in V). 👍

XX.6.3.8. This result shows that if V is a normed vector space, then the dual space V^* () is "large" (large enough to determine the norm on V). Without this result it is not obvious that a normed vector space has any nonzero bounded linear functionals.

XX.6.3.9. Theorem XX.6.3.5. requires some version of the AC (XX.6.1.11.). However, if V is separable, or more generally contains a well-ordered total subset, no version of the AC is necessary. Thus this version of the Hahn-Banach Theorem holds for separable Banach spaces with no Choice needed.

XX.6.3.10. Actually the statements of the various versions of the complex Hahn-Banach Theorem apply verbatim to real vector spaces, since $\phi = Re(\phi)$ for a real linear functional. But it seems clearer to state the real and complex cases separately, especially since the real case is used to prove the complex case.

The Canonical Embedding into the Second Dual

XX.6.3.11. If V is a normed complex vector space, then V^* is a Banach space (). V^* itself has a dual space $V^{**} = (V^*)^*$. If $x \in V$, then there is a canonically associated $\hat{x} \in V^{**}$ defined by

$$\hat{x}(\phi) = \phi(x)$$

for $\phi \in V^*$. It is easily checked that $\hat{x}$ is a linear functional on V^*, and that $x \mapsto \hat{x}$ is linear in x. Since $|\hat{x}(\phi)| = |\phi(x)| \leq \|x\|\|\phi\|$ for any $\phi \in V^*$, $\hat{x}$ is bounded and $\|\hat{x}\| \leq \|x\|$. By XX.6.2.3., we actually have $\|\hat{x}\| = \|x\|$. So:

XX.6.3.12. COROLLARY. Let V be a normed complex vector space. The map $x \mapsto \hat{x}$ is an isometric linear embedding of V into V^{**}, called the *canonical embedding*.

XX.6.4. Separation of Convex Sets and Version 3

The Hahn-Banach Theorem can be rephrased as a statement that disjoint convex sets can be separated by linear functionals or hyperplanes under mild additional hypotheses.

We will first work in real vector spaces, and then give the (slight) modifications necessary in the complex case.

XX.6.4.1. DEFINITION. Let X be a vector space over $\mathbb{R}$, and $A, B \subseteq X$. Then A and B are *separated* (in X) if there is a nonzero linear functional ϕ on X with $\phi(x) \leq \phi(y)$ for all $x \in A$, $y \in B$, i.e.

$$\sup_{x \in A} \phi(x) \leq \inf_{y \in B} \phi(y) \ .$$

We say ϕ *separates* A and B. The sets A and B are *strictly separated* if there is a (necessarily nonzero) linear functional ϕ on X with

$$\sup_{x \in A} \phi(x) < \inf_{y \in B} \phi(y) \ .$$

We say ϕ *strictly separates* A and B.

The definition is symmetric in A and B: if ϕ (strictly) separates A and B, then $-\phi$ (strictly) separates B and A.

XX.6.4.2. Strictly separated sets must be disjoint. Separated sets (by this definition) are not necessarily disjoint, but must be "almost disjoint." We will usually only consider separation for disjoint sets. Indeed, some references require separated sets to be disjoint.

Separation can be rephrased "geometrically" in terms of hyperplanes:

XX.6.4.3. DEFINITION. Let X be a vector space over $\mathbb{R}$. A *hyperplane* in X is an affine subspace of X of codimension 1.

XX.6.4.4. If H is a hyperplane in X, then there is a (nonunique, necessarily nonzero) linear functional ϕ and a constant c such that $H = \{x \in X : \phi(x) = c\}$: fix $x_0 \in H$, and let $Y = H - x_0$; then Y is a subspace of X of codimension 1. If $z \in X \setminus Y$, then X is spanned by Y and z; for $x = y + \alpha z \in X$, where $y \in Y$, set $\phi(x) = \alpha$ and $c = \phi(x_0)$. (Conversely, if ϕ is a nonzero linear functional, and $c \in \mathbb{R}$, then $\{x \in X : \phi(x) = c\}$ is a hyperplane in X.)

As a result, X splits into two *closed half-spaces* $H^+ = \{x \in X : \phi(x) \geq c\}$ and $H^- = \{x \in X : \phi(x) \leq c\}$, whose intersection is H, and two disjoint *open half-spaces* $H_0^+ = \{x \in X : \phi(x) > c\}$ and $H_0^- = \{x \in X : \phi(x) < c\}$ with H^+ the disjoint union of H_0^+ and H, and H^- the disjoint union of H_0^- and H. These half-spaces are independent of the choice of ϕ except that $+$ and $-$ can be interchanged if ϕ is changed.

Thus separation can be rephrased:

XX.6.4.5. PROPOSITION. Let X be a vector space over $\mathbb{R}$, and $A, B \subseteq X$. Then A and B are separated if and only if there is a hyperplane H in X such that $A \subseteq H^-$ and $B \subseteq H^+$. (We say A and B are *separated by* H.)

Strict separation cannot be quite so cleanly stated; see Exercise XX.6.5.1..

Hahn-Banach Theorem Version 3

There is an important rephrasing of the Hahn-Banach Theorem in terms of separation of convex sets, called Version 3 or the *Hahn-Banach Separation Theorem*:

XX.6.4.6. THEOREM. [HAHN-BANACH SEPARATION THEOREM] Let X be a vector space over $\mathbb{R}$, and A, B disjoint nonempty convex sets in X. If one of A, B has an internal point, then A and B are separated (by a linear functional or hyperplane).

PROOF: Suppose without loss of generality that A has an internal point z. Separating A and B is equivalent to separating $A - z$ and $B - z$, so we may assume $z = 0$. And separating A and B is the same as separating $A - B$ (which is convex, and $0 \notin A - B$) and $\{0\}$.

Fix $y \in B$. Then $-y$ is an internal point of $A - B$, so 0 is an internal point of the convex set $C = A - B + y$, i.e. C is radial. We have $y \notin C$ since $0 \notin A - B$.

Let m_C be the Minkowski functional of C. Then m_C is subadditive since C is convex; $m_C \leq 1$ on C, and $m_C(y) \geq 1$ since $y \notin C$. Define a linear functional ψ on the one-dimensional subspace spanned by y by $\psi(\alpha y) = \alpha$. Then $\psi(\alpha y) \leq m_C(\alpha y)$ for all $\alpha \in \mathbb{R}$. By the Hahn-Banach Theorem Version 1A (), there is a linear functional ϕ on X extending ψ with $\phi(x) \leq m_C(x)$ for all $x \in X$. Thus $\phi(x) \leq 1$ for $x \in C$ and $\phi(y) \geq 1$, so ϕ separates C and y and therefore separates $A - B$ and 0, i.e. separates A and B. 👍

XX.6.4.7. EXAMPLE. Here is an interesting "folklore" example of a pair of convex sets which cannot be separated by a linear functional, showing that the hypothesis of an internal point cannot be eliminated.

Let $X = c_{00}$ (). Let A be the set of nonzero sequences in X whose last nonzero term is positive. It is obvious that A is convex. Similarly, let $B = -A$ be the set of nonzero sequences whose last nonzero term is negative. Then B is convex, and X is the disjoint union of A, B, and $\{0\}$.

If X is given the uniform norm () (or any vector space topology), then A and B are dense in X: if $x \in X$, fix a k larger than the index of the last nonzero term in x, and then $x + \epsilon \mathbf{e}_k \in A$ for any $\epsilon > 0$, where $\mathbf{e}_k$ is the k'th standard basis vector (). A similar argument shows that A and B have no interior (or internal) points.

Since A and B are dense in X, it is obvious they cannot be separated by a bounded linear functional. But in fact they cannot be separated by *any* nonzero linear functional, bounded or not, i.e. there is no nonzero linear functional ϕ on X such that $\phi(x) \geq 0$ for all $x \in A$ (a separating linear functional or its negative would have to have this property). For if ϕ is such a linear functional, and $x \in X$ is arbitrary, fix k larger than the index of the last nonzero term in x. Then $x + \epsilon \mathbf{e}_k \in A$ for any $\epsilon > 0$, so

$$0 \leq \phi(x + \epsilon \mathbf{e}_k) = \phi(x) + \epsilon \phi(\mathbf{e}_k)$$

for any $\epsilon > 0$. Letting $\epsilon \to 0$, we obtain $\phi(x) \geq 0$. Since $x \in X$ is arbitrary, we also have $\phi(-x) \geq 0$, so $\phi(x) = 0$ for any x and ϕ is the zero linear functional.

Note that we have referred to a norm topology on X, but actually this example is purely algebraic and can be written (except for the density of A and B) without any reference to a topology on X.

The Complex Case

Just as with the previous versions of the Hahn-Banach Theorem, there is a similar statement for complex vector spaces. Separation must be defined slightly differently:

XX.6.4.8. Definition. Let X be a vector space over $\mathbb{C}$, and $A, B \subseteq X$. Then A and B are *separated* (in X) if there is a nonzero linear functional ϕ on X with $Re(\phi(x)) \leq Re(\phi(y))$ for all $x \in A$, $y \in B$, i.e.

$$\sup_{x \in A} Re(\phi(x)) \leq \inf_{y \in B} Re(\phi(y)) \ .$$

We say ϕ *separates* A and B. The sets A and B are *strictly separated* if there is a (necessarily nonzero) linear functional ϕ on X with

$$\sup_{x \in A} Re(\phi(x)) < \inf_{y \in B} Re(\phi(y)) \ .$$

We say ϕ *strictly separates* A and B.

There is not a good rephrasing in terms of hyperplanes.

XX.6.4.9. Theorem. [Complex Hahn-Banach Separation Theorem] Let X be a vector space over $\mathbb{C}$, and A, B disjoint nonempty convex sets in X. If one of A, B has an internal point, then A and B are separated.

XX.6.5. Exercises

XX.6.5.1. Here are two natural separation properties between separation and strict separation.

(a) If X is a topological vector space and A and B disjoint convex sets in X, then A and B are *completely separated* if there is a hyperplane H such that A and B are contained in distinct open half-spaces for H.

(b) If X is a topological vector space and A and B disjoint convex sets in X, then A and B are *fully separated* if there is a hyperplane H such that one of A and B is contained in an open half-space for H and the other in the complementary closed half-space.

(c) Observe that strict separation $\Rightarrow$ complete separation $\Rightarrow$ full separation $\Rightarrow$ separation.

(d) Show that A and B are fully separated if and only if there is a continuous linear functional ϕ on X such that $\phi(a) < \phi(b)$ for all $a \in A$, $b \in B$, and that A and B are completely separated if and only if there is a continuous linear functional ϕ and a real number m such that $f(a) < m < f(b)$ for all $a \in A$, $b \in B$.

(e) Show that if A and B are separated convex sets in a topological vector space, and if one is open, then A and B are fully separated. If both are open, then they are completely separated. If A and $\bar{B}$ are not disjoint, then A and B cannot be completely separated, and if $\bar{A}$ and B are also not disjoint, then A and B cannot be fully separated.

(f) Let $X = \mathbb{R}^2$, A the open upper half-plane along with the positive x-axis, and B the lower half-plane along with the negative x-axis. Show that A and B are disjoint, convex, separated, but not fully separated.

(g) Let $X = \mathbb{R}^2$, A the x-axis, and $B = \left\{(x, y) : x > 0, y \geq \frac{1}{x}\right\}$. Show that A and B are disjoint closed convex sets which are fully separated but not completely separated.

(h) Show that if A and B are disjoint closed convex subsets of $\mathbb{R}^2$, then A and B are fully separated.

(i) [KN76, 3G] Let $X = \mathbb{R}^3$,

$$A = \{(x, y, z) : x \geq 0, y \geq 0, z \geq 0, z^2 \leq xy\}$$

$$B = \{(0, y, 1) : y \in \mathbb{R}\} \ .$$

Show that A and B are disjoint, closed, convex, and separated, but not fully separated.

Thus full or complete separation is not always possible, even in finite-dimensional spaces.

XX.7. Dual Spaces and Weak Topologies

One of the core tenets of functional analysis is that spaces can be studied via the scalar-valued continuous functions on them. This section describes the linear version of this philosophy.

XX.7.1. Bounded Linear Functionals

XX.7.1.1. Definition. Let X be a vector space over $\mathbb{F}$. A *linear functional* on X is a linear transformation from X to $\mathbb{F}$. If X is a normed vector space, a linear functional ϕ on X is *bounded* if it is bounded as a linear transformation, i.e. if there is a constant M such that $|\phi(x)| \leq M\|x\|$ for all $x \in X$. The smallest such M is the *norm* of ϕ, denoted $\|\phi\|$.

XX.7.1.2. **Notation:** If X is a vector space over $\mathbb{F}$, we will write X^d for the set of linear functionals on X. If X is normed, write X^* for the set $\mathcal{B}(X, \mathbb{F})$ of bounded linear functionals on X. We have that X^d is a vector space over $\mathbb{F}$ in the obvious way, and by (), X^* is a subspace which is normed by $\|\cdot\|$; X^* is a Banach space by (). X^d is called the *algebraic dual space* of X, and X^* the *(topological) dual space* of X.

There is no uniformity of notation for dual spaces. Many authors use X' for the topological dual space, and some use X^* or X' to denote the algebraic dual. We use what seems to be the most common notation for the topological dual, although the notation X^d is not standard for the algebraic dual. We usually use ϕ, ψ, etc. to denote linear functionals; some authors use x^* to denote a general element of X^*, leading to poor and confusing expressions like $x^*(x)$ where x denotes a general element of X and x^* a general element of X^*, with no relation between them.

XX.7.1.3. By (), a linear functional on a normed vector space is continuous if and only if it is bounded. The null space of any bounded operator is closed (). The converse does not hold in general, e.g. there are unbounded one-to-one linear transformations. However:

XX.7.1.4. Proposition. Let X be a normed vector space over $\mathbb{F}$, and $\phi : X \to \mathbb{F}$ a linear functional on X. If the null space N_ϕ is closed in X, then ϕ is bounded.

Proof: Suppose ϕ is unbounded. Then for each $n \in \mathbb{N}$ there is a $y_n \in X$ with $\|y_n\| = 1$ but $|\phi(y_n)| > n$. Set $x_n = \frac{1}{\phi(y_n)} y_n$. Then $\|x_n\| < 1/n$ and $\phi(x_n) = 1$ for all n. Fix $v \in X$ with $\phi(v) = 1$. We have $x_n \to 0$, so $v - x_n \to v$. But $v - x_n \in N_\phi$ for all n, so $v \in \overline{N_\phi} \setminus N_\phi$, and N_ϕ is not closed.

In fact, this argument shows that if ϕ is unbounded, then N_ϕ is dense in X. (Actually it would have to be either closed or dense since it has codimension 1.)

Without the Hahn-Banach Theorem, it is not obvious that a normed vector space has any nonzero bounded linear functionals. But by XX.6.2.3. and (), we have:

XX.7.1.5. Corollary. Let $X \neq \{0\}$ be a normed vector space, and $x \in X$. Then

$$\|x\| = \max\{|\phi(x)| : \phi \in X^*, \|\phi\| = 1\} = \sup\{|\phi(x)| : \phi \in X^*, \|\phi\| = 1\} .$$

Here is a simple but useful consequence:

XX.7.1.6. PROPOSITION. Let X be a normed vector space and Y a proper closed subspace. Then there is a bounded linear functional ϕ on X with $\|\phi\| = 1$ and $Y \subseteq \mathcal{N}(\phi)$ (i.e. $\phi(y) = 0$ for all $y \in Y$).

PROOF: The normed vector space X/Y has a nonzero bounded linear functional ψ. Let $\pi : X \to X/Y$ be the quotient map; then $\|\pi\| \leq 1$ (in fact $\|\pi\| = 1$ by ()). Then $\psi \circ \pi$ is a nonzero bounded linear functional on X whose null space contains Y, and can be normalized to ϕ. 👍

Subspaces and Quotients

We can be more specific. Let X be a normed space and Y a closed subspace. What is the relation between X^*, Y^*, and $(X/Y)^*$?

XX.7.1.7. We define the *annihilator* $Y^\perp$ to be

$$Y^\perp = \{\phi \in X^* : \phi(y) = 0 \ \forall y \in Y\} .$$

Then $Y^\perp$ is a closed subspace of X^*. By XX.7.1.6., $Y^\perp = \{0\}$ if and only if $Y = X$. At the other extreme, $\{0\}^\perp = X^*$.

XX.7.1.8. Let ρ be the restriction map from X^* to Y^*. Then ρ is surjective by the Hahn-Banach Extension Theorem (), and its null space is exactly $Y^\perp$. Thus ρ drops to a vector space isomorphism $\bar{\rho}$ from $X^*/Y^\perp$ to Y^*.

XX.7.1.9. PROPOSITION. The map $\bar{\rho} : X^*/Y^\perp \to Y^*$ is an isometric isomorphism.

PROOF: It is convenient to show that $\eta = \bar{\rho}^{-1} : Y^* \to X^*/Y^\perp$ is an isometry. If $\psi \in Y^*$, then $\eta(\psi)$ is the equivalence class of all extensions of ψ to X, and $\|\eta(\phi)\|$ is by definition the infimum of the norms of all extensions. Each extension ϕ has $\|\phi\| \geq \|\psi\|$, so $\|\eta(\psi)\| \geq \|\psi\|$, and by the Hahn-Banach Extension Theorem there is a ϕ with $\|\phi\| = \|\psi\|$, so $\|\eta(\psi)\| \leq \|\psi\|$. 👍

XX.7.1.10. Let $\pi : X \to X/Y$ be the quotient map. Since π is bounded, every $\psi \in (X/Y)^*$ defines a bounded linear functional $\psi \circ \pi$ on X, which vanishes on Y, i.e. $\psi \circ \pi \in Y^\perp$. The map $\psi \mapsto \psi \circ \pi$ is a linear map σ from $(X/Y)^*$ to $Y^\perp$.

XX.7.1.11. PROPOSITION. The map $\sigma : (X/Y)^* \to Y^\perp$ is an isometric isomorphism.

PROOF: The map σ is clearly injective. Any $\phi \in Y^\perp$ has null space containing Y and thus drops to a linear functional $\bar{\phi}$ on X/Y. If $z \in X/Y$, $\|z\| = 1$, then for any $\alpha > 1$ there is an $x \in X$ with $\pi(x) = z$ and $\|x\| < \alpha$. Thus

$$|\bar{\phi}(z)| = |\phi(x)| \leq \|\phi\|\|x\| < \alpha\|\phi\|$$

and since this holds for all z and all $\alpha > 1$, $\|\bar{\phi}\| \leq \|\phi\|$. In particular, $\bar{\phi}$ is bounded, hence in $(X/Y)^*$; thus σ is surjective. The map $\omega : \phi \mapsto \bar{\phi}$ is exactly σ^{-1}, and this argument shows that $\|\omega\| \leq 1$. Conversely, if $\psi \in (X/Y)^*$, let $\phi = \sigma(\psi)$. If $x \in X$, we have

$$|\phi(x)| = |\psi(\pi(x))| \leq \|\psi\|\|\pi(x)\| \leq \|\psi\|\|x\|$$

so $\|\phi\| \leq \|\psi\|$. Therefore $\|\sigma\| \leq 1$. 👍

XX.7.1.12. Thus the dual of a subspace of X is a quotient of X^*, and the dual of a quotient of X is a subspace of X^*. The correspondence between X and X^* is an important example of a *duality* (). In a duality, subobjects correspond to quotients and vice versa.

XX.7.1.13. The duality between X and X^* is not complete, even if we restrict to Banach spaces (i.e. assume X is complete): not every closed subspace of X^* is $Y^\perp$ for a closed subspace Y of X in general. See XX.7.4.13. for further discussion of this.

XX.7.2. The Weak Topology

If X is a normed vector space, then X of course has a norm topology. But it also has another natural and important topology, the weak topology.

XX.7.2.1. Let X be a normed vector space. If $\phi_1, \ldots, \phi_n \in X^*$ and $\epsilon > 0$, set

$$U_{\{\phi_1,\ldots,\phi_n\},\epsilon} = \{x \in X : |\phi_1(x)| < \epsilon, \ldots, |\phi_n(x)| < \epsilon\} .$$

Note that $U_{\{\phi_1,\ldots,\phi_n\},\epsilon}$ is an open neighborhood of 0 (in the norm topology) and is convex and circled (). These sets form a local base at 0 for a topology making X into a locally convex topological vector space. In fact, by scaling the ϕ_k, it suffices to take basic open sets with $\epsilon = 1$; and it suffices to take basic open sets with the ϕ_k linearly independent (Exercise XX.7.5.7.).

XX.7.2.2. Definition. This topology is called the *weak topology* on X (from X^*).

Since the basic open sets for the weak topology are norm-open, the weak topology is weaker than the norm topology. It is almost always strictly weaker (XX.7.2.6.).

XX.7.2.3. Proposition. The weak topology on X is Hausdorff.

Proof: It suffices to show that the intersection of the basic open sets is $\{0\}$. Let $x \in X$, $x \neq 0$, and let ϕ be a bounded linear functional on X with $\phi(x) = \|x\|$. Set $\epsilon = \|x\|/2$. Then $U_{\{\phi\},\epsilon}$ is a basic open neighborhood of 0 not containing x. 👍

The weak topology is the "topology of pointwise convergence":

XX.7.2.4. PROPOSITION. Let X be a normed vector space, (x_i) a net in X, and $x \in X$. Then $x_i \to x$ in the weak topology if and only if $\phi(x_i) \to \phi(x)$ for every $\phi \in X^*$.

PROOF: Replacing x_i by $x_i - x$, we may assume $x = 0$. Suppose $x_i \to 0$ in the weak topology. Let $\phi \in X^*$ and $\epsilon > 0$. Then $x_i \in U_{\{\phi\},\epsilon}$ for all sufficiently large i, so $\phi(x_i) \to 0$. Conversely, if $\phi(x_i) \to 0$ for all $\phi \in X^*$, let $\phi_1, \ldots, \phi_n \in X^*$ and $\epsilon > 0$. Then for each k, $1 \leq k \leq n$, there is an i_k such that $|\phi(x_i)| < \epsilon$ for all $i > i_k$. Since the index set of the net is directed, there is an i_0 with $i_0 > i_k$ for $1 \leq k \leq n$. If $i > i_0$, then $x_i \in U_{\{\phi_1,\ldots,\phi_n\},\epsilon}$. Thus $x_i \to 0$ in the weak topology. 👍

XX.7.2.5. In light of XX.7.1.5., the norm topology can be called "the topology of uniform convergence." In many references, the norm topology is called the *strong topology*, but this is too easily confused with the strong operator topology () when X is a space of operators, so we will avoid this terminology.

XX.7.2.6. PROPOSITION. Let X be a normed vector space. Then the weak topology on X coincides with the norm topology if and only if X is finite-dimensional.

PROOF: If X is finite-dimensional, then all Hausdorff TVS topologies on X coincide (). Conversely, if the norm and weak topologies on X coincide, the open unit ball of X is a weak neighborhood of 0, hence contains a basic neighborhood $U_{\{\phi_1,\ldots,\phi_n\},1}$, i.e. there are $\phi_1, \ldots, \phi_n \in X^*$ such that $|\phi_k(x)| < 1$ for all k implies $\|x\| < 1$. Then for any $\epsilon > 0$, $|\phi_k(x)| < \epsilon$ for all k implies $\|x\| < \epsilon$, so $\phi_k(x) = 0$ for all k implies $x = 0$. Thus the linear map $T : X \to \mathbb{F}^n$, $T(x) = (\phi_1(x), \ldots, \phi_n(x))$, has $\mathcal{N}(T) = \{0\}$ and is therefore injective, so $dim(X) \leq n$. 👍

XX.7.2.7. Existence of the weak topology distinct from the norm topology reinforces the warning in (): spaces in analysis can have more than one natural topology, and it is crucial to always keep track of what topology is being considered. We sometimes need to even consider delicate interactions of two or more topologies (e.g. ()).

Since the norm and weak topologies are distinct in general, a norm-closed subset of a normed vector space is not generally weakly closed. But there is an important case where this holds:

XX.7.2.8. PROPOSITION. Let X be a normed vector space. Every norm-closed convex subset of X is weakly closed.

PROOF: Let C be a norm-closed convex subset of X. If $x \in X$, $x \notin C$, then by the Hahn-Banach Separation Theorem () there is a bounded linear functional ϕ on X with

$$|\phi(x)| > \sup\{|\phi(y)| : y \in C\} .$$

By XX.7.2.4., x is not in the weak closure of C. Since this is true for every $x \in X \setminus C$, C is weakly closed. 👍

XX.7.2.9. Corollary. Every norm-closed subspace Y of X is weakly closed. The weak topology on Y is the relative topology from the weak topology on X.

Proof: For the last statement, every basic open neighborhood for the weak topology of Y is the intersection with Y of a basic open neighborhood for the weak topology of X by the Hahn-Banach Extension Theorem. 👍

Here is a nice consequence:

XX.7.2.10. Corollary. Let X be a normed vector space. If $S \subseteq X$ and $C = Conv(S)$, then any x in the weak closure of S is in the norm-closure of C, and is thus a norm-limit of a sequence in C. In particular, if (x_i) is a net in X converging weakly to $x \in X$, then there is a sequence in the convex hull of the x_i which converges to x in norm.

Proof: The norm-closed convex hull $\bar{C}$ of S is norm-closed and convex, hence weakly closed. Since $S \subseteq C$, $x \in \bar{C}$. Since X is a metric space, there is a sequence in C converging to x in norm. 👍

XX.7.2.11. Note that the result of Proposition XX.7.2.8. does not hold with "closed" replaced by "open." Indeed, open balls in X are norm-open convex subsets which are not generally weakly open (if they were, the topologies would coincide!) ✋

Similarly, we have:

XX.7.2.12. Proposition. Let X be a normed vector space. Then a linear functional on X is weakly continuous if and only if it is norm-continuous (i.e. bounded).

Proof: Since the weak topology is weaker than the norm topology, a weakly continuous linear functional is norm-continuous. Conversely, let ϕ be norm-continuous, i.e. $\phi \in X^*$. Then $U_{\{\phi\},\epsilon} = \{x \in X : |\phi(x)| < \epsilon\}$ is weakly open for every $\epsilon > 0$, so ϕ is weakly continuous at 0, hence everywhere. 👍

XX.7.2.13. If X is a normed vector space with completion $\bar{X}$, then X^* can be naturally identified with $(\bar{X})^*$, and it follows that the weak topology on X is just the relative topology from the weak topology on $\bar{X}$.

XX.7.2.14. In the infinite-dimensional case, weak (and weak-*) topologies cannot be entirely described by convergence of sequences (XX.7.2.17.; cf. Exercise XX.7.5.4.). But in most applications, one considers only convergence of sequences, so it is important to have results about weak (or weak-*) convergence of sequences. Thus we will have various results about sequences, such as the following.

XX.7.2.15. PROPOSITION. Let X be a normed vector space. Then every weakly convergent sequence in X is norm-bounded.

PROOF: Let (x_k) be a weakly convergent sequence in X. Then, for every ϕ in X^*, the sequence $(\phi(x_k))$ is convergent in $\mathbb{F}$, hence bounded (). For each k, define $T_k : X^* \to \mathbb{F}$ by $T_k(\phi) = \phi(x_k)$. The T_k are pointwise bounded. Thus by the Uniform Boundedness Theorem () there is an M such that $\|T_k\| \leq M$ for all k, i.e. $|\phi(x_k)| \leq M$ for all k and all $\phi \in X^*$, $\|\phi\| \leq 1$. This implies that $\|x_k\| \leq M$ for all k (XX.7.1.5.). 👍

XX.7.2.16. Note that this result only applies to *sequences*, not nets. In fact, if X is infinite-dimensional, weak neighborhoods of 0 are always unbounded, and it is easy to construct an unbounded net (indexed by basic open neighborhoods) converging to 0. ✋

When the weak topology does not coincide with the norm topology, it is somewhat pathological:

XX.7.2.17. PROPOSITION. Let X be an infinite-dimensional normed vector space. Then the weak topology on X is not first countable (hence not metrizable).

PROOF: Suppose there is a decreasing sequence (U_k) of weakly open neighborhoods of 0 forming a local base. Since each weak neighborhood of 0 is unbounded, for each k there is an $x_k \in U_k$ with $\|x_k\| \geq k$. Then $x_k \to 0$ weakly, but is unbounded, contradicting XX.7.2.15.. 👍

However, in good cases the restriction of the weak topology to norm-bounded sets is metrizable (note that no weak neighborhood of 0 is norm-bounded if X is infinite-dimensional):

XX.7.2.18. PROPOSITION. Let X be a normed vector space. If X^* is separable (in the norm topology), then for any norm-bounded subset A of X, the relative weak topology on A is metrizable.

PROOF: It suffices to prove the statement when A is the unit ball of X. Let (ψ_k) be a countable norm-dense subset of X^*. For each k let ρ_k be the pseudometric on A defined by the seminorm $\|x\|_k = |\psi_k(x)|$. If $\phi_1, \ldots, \phi_n \in X^*$ and $\epsilon > 0$, for each k, $1 \leq k \leq n$, choose ψ_k with $\|\phi_k - \psi_k\| < \frac{\epsilon}{2}$. Then the basic open set $U_{\{\psi_1,\ldots,\psi_n\},\epsilon/2} \cap A$ is contained in $U_{\{\phi_1,\ldots,\phi_n\},\epsilon} \cap A$. It follows that the relative weak topology on A is defined by the pseudometrics ρ_k. Thus the topology is metrizable (). 👍

XX.7.2.19. Note that in this proof, $U_{\{\psi_1,\ldots,\psi_n\},\epsilon/2}$ is not generally contained in $U_{\{\phi_1,\ldots,\phi_n\},\epsilon}$; only the intersection with the unit ball has the containment. Thus there is no contradiction with XX.7.2.17.. ✋

In connection with XX.7.2.18., we have:

XX.7.2.20. PROPOSITION. Let X be a normed vector space. If X^* is (norm-)separable, then X is (norm-)separable.

PROOF: Let (ψ_k) be a norm-dense sequence in X^*. For each k let $x_k \in X$ with $\|x_k\| = 1$ and $|\psi_k(x_k)| > \frac{1}{2}\|\psi_k\|$. Let Y be the closed span of the x_k. If Y is a proper subspace of X, then there is a $\phi \in X^*$ with $\|\phi\| = 1$ and $Y \subseteq \mathcal{N}(\phi)$ (XX.7.1.6.), so $\phi(x_k) = 0$ for all k. Fix k with $\|\psi_k - \phi\| < \frac{1}{3}$; then $\|\psi_k\| > \frac{2}{3}$, so $|\psi_k(x_k)| > \frac{1}{3}$. But

$$|\psi_k(x_k)| = |\psi_k(x_k) - \phi(x_k)| \leq \|\psi_k - \phi\|\|x_k\| < \frac{1}{3}$$

which is a contradiction. Thus $Y = X$, so finite linear combinations of the x_k are (norm-)dense in X, hence finite linear combinations with coefficients in $\mathbb{Q}$ ($\mathbb{Q} + \mathbb{Q}i$ if $\mathbb{F} = \mathbb{C}$) form a countable norm-dense set in X. 👍

XX.7.2.21. The converse of this result is false: many (perhaps "most") separable Banach spaces, e.g. ℓ^1 (), have nonseparable duals. ✋

XX.7.3. The Second Dual

XX.7.3.1. Let X be a normed vector space. Then X^* is also a normed vector space, even a Banach space, so has a dual space $(X^*)^*$, denoted X^{**} and called the *second dual* of X.

The Canonical or Natural Embedding

XX.7.3.2. If $x \in X$, there is a natural linear functional $\hat{x}$ on X^* given by evaluation:

$$\hat{x}(\phi) = \phi(x) \ .$$

Since $|\hat{x}(\phi)| = |\phi(x)| \leq \|\phi\|\|x\|$ for any ϕ, we have that $\hat{x}$ is bounded, i.e. $\hat{x} \in X^{**}$, and $\|\hat{x}\| \leq \|x\|$. On the other hand, by () there is a ϕ with $\|\phi\| = 1$ and $|\hat{x}(\phi)| = \|x\|$, so $\|\hat{x}\| \geq \|x\|$; thus $\|\hat{x}\| = \|x\|$.

XX.7.3.3. Thus the map $\iota : x \mapsto \hat{x}$ is an isometric embedding of X into X^{**}, called the *canonical embedding* or *natural embedding*. So we can, and often will, regard X as a subspace of X^{**}.

XX.7.3.4. The canonical embedding is not surjective in general. If X is not complete, it cannot be surjective since X^{**} is complete. But even if X is complete, ι is not surjective in general; indeed, X^{**} is often *much* larger than X (it is often nonseparable even if X is separable).

XX.7.3.5. EXAMPLE. Let $X = c_0$ (), the Banach space of sequences (in $\mathbb{F}$) converging to 0, with the supremum norm. Then X^* can be naturally identified with ℓ^1, the absolutely summable sequences. Then X^{**} can be identified with the dual of ℓ^1, which is ℓ^∞, the Banach space of bounded sequences. The canonical embedding of X into X^{**} is just the natural embedding of c_0 as a subspace of ℓ^∞. Note that in this case c_0 is separable but $\ell^\infty = c_0^{**}$ is nonseparable. In this case $c_0^* = \ell^1$ is separable. We also have that $(\ell^1)^{**} = (\ell^\infty)^*$ is nonseparable by XX.7.2.20., so the canonical map from ℓ^1 to $(\ell^1)^{**}$ is far from surjective.

Reflexive Banach Spaces

XX.7.3.6. There are Banach spaces for which the canonical embedding is surjective, hence an isometric isomorphism. Such a Banach space is called *reflexive*. Reflexive Banach spaces are exactly the ones for which the duality between X and X^* is complete ().

XX.7.3.7. Examples. (i) Any finite-dimensional normed vector space is reflexive (by dimension counting).

(ii) Any Hilbert space is reflexive ().

(iii) If $(X, \mathcal{A}, \mu)$ is a measure space, then $L^p(X, \mathcal{A}, \mu)$ is reflexive for $1 < p < \infty$ (). In particular, ℓ^p is reflexive for $1 < p < \infty$.

(iv) c_0, ℓ^1, ℓ^∞ are not reflexive. More generally, $L^1(X, \mathcal{A}, \mu)$ and $\ell^\infty(X, \mathcal{A}, \mu)$ are not reflexive for most measure spaces $(X, \mathcal{A}, \mu)$ (the exact condition is a little tricky).

(iv) If X is a compact Hausdorff space, then $C(X)$ is not reflexive unless X is a finite discrete space (). More generally, no infinite-dimensional C*-algebra is reflexive ().

Reflexivity will be discussed in more detail in ().

Higher Duals

XX.7.3.8. Why stop at two? X^{**} has a dual space, denoted X^{***}. We can go on *ad infinitum*.

XX.7.3.9. There is a canonical embedding of X^* into X^{***}, and of X^{**} into X^{****}, etc. The canonical embedding of X^* into X^{***} is surjective (i.e. X^* is reflexive) if and only if X is reflexive; the same holds for higher duals (cf. XX.7.1.6.). In general, the duals get much larger at every step.

The duals form two towers: $X \subseteq X^{**} \subseteq X^{****} \subseteq \cdots$ and $X^* \subseteq X^{***} \subseteq X^{*****} \subseteq \cdots$. The two towers are in duality.

XX.7.3.10. If the n'th dual of X is separable, then X and the k'th dual of X for $k < n$ are also separable by repeated application of XX.7.2.20.. But they can become nonseparable at some step (cf. XX.7.3.5.).

XX.7.4. The Weak-* Topology

If X is a normed vector space, the dual space X^* is itself a Banach space, so has a weak topology (and of course also a norm topology). But it also has a more important topology, the weak-* topology.

XX.7.4.1. If $x_1, \ldots, x_n \in X$ and $\epsilon > 0$, we let

$$U_{\{x_1,\ldots,x_n\},\epsilon} = \{\phi \in X^* : |\phi(x_1)| < \epsilon, \ldots, |\phi(x_n)| < \epsilon\} .$$

This is a norm-open convex circled neighborhood of 0 in X^*, and the set of all these basic open sets forms a base for a TVS topology on X^*.

XX.7.4.2. Definition. This topology is called the *weak-* topology* on X^* (from X).

XX.7.4.3. The basic open neighborhoods and the weak-* topology are defined without reference to the second dual X^{**}. But the definition can be rephrased to look more like the definition of the weak topology: $U_{\{x_1,\dots,x_n\},\epsilon} = U_{\{\hat{x}_1,\dots,\hat{x}_n\},\epsilon}$, where $\hat{x}_k$ is the linear functional on X^* corresponding to x_k. Thus the basic open neighborhoods for the weak-* topology are basic open neighborhoods for the weak topology, so it follows that the weak-* topology is weaker than the weak topology on X^*. If X is reflexive, the basic open neighborhoods are the same, so the weak and weak-* topologies coincide (the converse also holds (XX.7.4.21.), i.e. the weak-* topology is strictly weaker than the weak topology unless X is a reflexive Banach space).

XX.7.4.4. The weak-* topology on X^* has a very natural interpretation: it is exactly the topology of pointwise convergence on X, i.e. a net (ϕ_i) in X^* converges weak-* to ϕ if and only if $\phi_i(x) \to \phi(x)$ for all $x \in X$. (The proof is almost identical to the proof of XX.7.2.4. and is left to the reader.)

XX.7.4.5. If $\phi \in X^*$, $\phi \neq 0$, let $x \in X$ with $\phi(x) \neq 0$. Then $U_{\{x\},|\phi(x)|}$ is an open neighborhood of 0 not containing ϕ. It follows that the weak-* topology is Hausdorff.

The most important fact about the weak-* topology, the main reason it is so useful, is Alaoglu's Theorem:

XX.7.4.6. Theorem. [Alaoglu's Theorem] Let X be a normed vector space. Then the norm-closed unit ball B of X^* is compact in the relative weak-* topology.

Proof: If $\alpha > 0$, let D_α be the closed ball in $\mathbb{F}$ of radius α around 0 (either $[-\alpha, \alpha]$ if $\mathbb{F} = \mathbb{R}$ or the closed disk of radius α around 0 if $\mathbb{F} = \mathbb{C}$). Then D_α is compact. Each $\phi \in B$ defines a function on X whose value at x is in $D_{\|x\|}$, which can be thought of as a point in the product

$$P := \prod_{x \in X} D_{\|x\|}$$

so we can regard B as a subset of P. The relative weak-* topology is the topology of coordinatewise convergence, i.e. the product topology. P is compact in the product topology by Tikhonov's Theorem (). So it suffices to show that B is a closed subset of P.

Suppose (ϕ_i) is a net in B converging to $\phi \in P$. Then ϕ can be regarded as a function from X to $\mathbb{F}$: $\phi(x) = \lim \phi_i(x)$ for all $x \in X$. If $x, y \in X$, we have

$$\phi(x+y) = \lim \phi_i(x+y) = \lim(\phi_i(x) + \phi_i(y)) = \phi(x) + \phi(y)$$

by continuity. Similarly, $\phi(\alpha x) = \alpha\phi(x)$ for $\alpha \in \mathbb{F}$, $x \in X$. So ϕ is a linear functional. For any x, $\phi(x) = \lim \phi_i(x) \in D_{\|x\|}$ since $D_{\|x\|}$ is closed in $\mathbb{F}$; thus ϕ is bounded and $\|\phi\| \leq 1$, i.e. $\phi \in B$. 👍

XX.7.4.7. Proposition. Let X be a separable normed vector space. Then norm-bounded subsets of X^* are metrizable in the relative weak-* topology from X.

The proof is almost identical to the proof of XX.7.2.18..

Combining the last two results and XIII.11.2.2., we obtain:

XX.7.4.8. Corollary. Let X be a separable normed vector space. Then the closed unit ball of X^* is sequentially compact in the weak-* topology, i.e. any norm-bounded sequence in X^* has a weak-* convergent subsequence.

This result is sometimes called *Helly's Theorem* since it was obtained in [].

The weak-* analogs of XX.7.2.15. and XX.7.2.8. are not unrestrictedly true. We have:

XX.7.4.9. Proposition. Let X be a Banach space. Then every weak-* convergent sequence in X is norm-bounded.

XX.7.4.10. The proof is almost identical to the proof of XX.7.2.15., except that the T_n are defined on X, not X^*, so completeness of X is needed to apply the Uniform Boundedness Theorem. In fact, the result can fail if X is not complete (Exercise XX.7.5.1.).

Not every norm-closed convex set in X^*, not even every closed subspace of X^*, is weak-* closed in general (XX.7.4.14.). In fact:

XX.7.4.11. Proposition. Let X be a normed vector space. Then a linear functional on X^* is weak-* continuous if and only if it is of the form $\hat{x}$ for some $x \in X$ (i.e. if and only if it is in the canonical image of X in X^{**}).

We will prove a more general result in XX.7.4.20..

XX.7.4.12. Let X be a normed vector space, and Z a subspace of X^*. Set

$$Z_\perp = \{x \in X : \phi(x) = 0 \ \forall \phi \in Y\} \ .$$

$Z_\perp$ is the *annihilator* of Z in X, a norm-closed subspace of X. Then

$$(Z_\perp)^\perp = \{\phi \in X^* : \phi(x) = 0 \ \forall x \in Z_\perp\}$$

is a norm-closed subspace of X^* containing Z. We have $[(Z_\perp)^\perp]_\perp \subseteq Z_\perp$ since $Z \subseteq (Z_\perp)^\perp$; on the other hand, if $x \in Z_\perp$, then $\phi(z) = 0$ for all $\phi \in (Z_\perp)^\perp$, so $x \in [(Z_\perp)^\perp]_\perp$, i.e. $[(Z_\perp)^\perp]_\perp = Z_\perp$.

XX.7.4.13. Proposition. The subspace $(Z_\perp)^\perp$ is the weak-* closure of the subspace Z. In particular, Z is weak-* closed if and only if $Z = (Z_\perp)^\perp$, if and only if $Z = Y^\perp$ for some (closed) subspace Y of X.

Proof: If ϕ is in the weak-* closure of $(Z_\perp)^\perp$, i.e. there is a net (ϕ_i) in $(Z_\perp)^\perp$ with $\phi_i \to \phi$ in the weak-* topology, then if $x \in Z_\perp$, $0 = \phi_i(x) \to \phi(x)$, so $\phi(x) = 0$; thus $\phi \in (Z_\perp)^\perp$. Therefore $(Z_\perp)^\perp$ is weak-* closed.

Conversely, if Z is weak-* closed and $\phi \in X^* \setminus Z$, then since the weak-* topology is locally convex there is a weak-* continuous linear functional f on X^* with $Z \subseteq \mathcal{N}(f)$ and $f(\phi) = 1$. By XX.7.4.11., $f = \hat{x}$ for some $x \in X$, i.e. $\phi(x) = 1$ and $\psi(x) = 0$ for all $\psi \in Z$. Then $x \in Z_\perp$, so $\phi \notin (Z_\perp)^\perp$. Therefore $(Z_\perp)^\perp = Z$. 👍

XX.7.4.14. Corollary. Let X be a normed vector space, canonically embedded in X^{**}. Then X is weak-* dense in X^{**}.

Proof: $X_\perp = \{0\}$ (in X^*). 👍

See Exercise XX.7.5.5. for a direct proof.

XX.7.4.15. The weak topology on a normed vector space X is intrinsic to X. But the weak-* topology on a Banach space Y is not intrinsic to Y: it depends on whether, and how, Y is realized as the dual of a normed space X. Not every Banach space is the dual of another Banach space (see ()). And there can be nonisomorphic Banach spaces X_1 and X_2 with X_1^* and X_2^* isometrically isomorphic, and in this case there may not be an isomorphism which is a homeomorphism for the weak-* topologies (see ()). As a related matter, if X is not complete, we have $\bar{X}^*$ naturally isomorphic to X^*, but the weak-* topologies from X and $\bar{X}$ are subtly different (XX.7.4.21.; see XX.7.5.1. for an example). The difference is indeed quite subtle: the relative topologies on norm-bounded sets agree (XX.7.5.2.).

Unified Notation

There is a nice alternate notation which describes both the weak and weak-* topologies, and carries more information including the distinctions of XX.7.4.15..

XX.7.4.16. Definition. Let X be a normed vector space and Y a subspace of X^* (not necessarily closed). The $\sigma(X,Y)$-*topology* on X is the topology of pointwise convergence of functionals in Y, i.e. $x_i \to x$ in the $\sigma(X,Y)$-topology if and only if $\phi(x_i) \to \phi(x)$ for all $\phi \in Y$.

XX.7.4.17. A system of basic neighborhoods of 0 for the $\sigma(X,Y)$-topology is the set of $U_{\{\phi_1,\ldots,\phi_n\},\epsilon}$ as in XX.7.2.1. with the $\phi_k \in Y$.

XX.7.4.18. We have that the weak topology on X is the $\sigma(X,X^*)$-topology and the weak-* topology on X^* from X is the $\sigma(X^*,X)$-topology, where X is regarded as the canonically embedded subspace of X^{**}.

XX.7.4.19. Proposition. (i) Let X be a normed vector space and Y a subspace of X^*. Then the $\sigma(X,Y)$-topology is Hausdorff if and only if Y separates the points of X, i.e. for any $x \in X$, $x \neq 0$, there is a $\phi \in Y$ with $\phi(x) \neq 0$. Such a subspace is called *total* in some references since Y separates points of X if and only if it is weak-*dense in X^* (XX.7.4.13.), although it need not be norm-dense.

(ii) If X is a normed vector space and $Y_1 \subseteq Y_2 \subseteq X^*$, then the $\sigma(X,Y_1)$-topology is weaker than the $\sigma(X,Y_2)$-topology.

(iii) If $X_1 \subseteq X_2$ are normed vector spaces, Y a subspace of X_2^*, and Z the set of restrictions of Y to X_1 (so $Z \cong Y/X_1^\perp$), then the $\sigma(X_1,Z)$-topology on X_1 is the relative $\sigma(X_2,Y)$-topology.

Continuity of Linear Functionals

XX.7.4.20. THEOREM. Let X be a normed vector space, and Y a subspace of X^*. Then the linear functionals on X which are $\sigma(X,Y)$-continuous are precisely the linear functionals in Y.

PROOF: Since the $\sigma(X,Y)$-topology is weaker than the norm topology, any $\sigma(X,Y)$-continuous linear functional is norm-continuous, i.e. in X^*.

If $\phi \in Y$, then $U_{\{\phi\},\epsilon}$ is a $\sigma(X,Y)$-open neighborhood of 0 for any $\epsilon > 0$; thus ϕ is $\sigma(X,Y)$-continuous at 0, hence everywhere.

Conversely, suppose ϕ is $\sigma(X,Y)$-continuous. Then $V = \{x \in X : |\phi(x)| < 1\}$ is a $\sigma(X,Y)$-open neighborhood of 0, so there are $\psi_1, \dots, \psi_n \in Y$ such that $U_{\{\psi_1,\dots,\psi_n\},1} \subseteq V$, i.e. $|\psi_k(x)| < 1$ for $1 \le k \le n$ implies $|\phi(x)| < 1$. Then, for any $\epsilon > 0$, $|\psi_k(x)| < \epsilon$ for $1 \le k \le n$ implies $|\phi(x)| < \epsilon$. Thus $\psi_k(x) = 0$ for $1 \le k \le n$ implies $\phi(x) = 0$. So ϕ is a linear combination of $\{\psi_1, \dots, \psi_n\}$ by III.11.5.16. and therefore $\phi \in Y$. 👍

XX.7.4.21. COROLLARY. Let X be a normed linear space, and Y_1 and Y_2 distinct subspaces of X^*. Then the $\sigma(X,Y_1)$-topology and $\sigma(X,Y_2)$-topology on X are different. If $Y_1 \subseteq Y_2$, then the $\sigma(X,Y_1)$-topology is strictly weaker than the $\sigma(X,Y_2)$-topology. In particular:

(i) If X is not reflexive, the weak-* topology on X^* from X is strictly weaker than the weak topology.

(ii) If X is not complete, the $\sigma(X^*,X)$-topology is strictly weaker than the $\sigma(X^*,\bar{X})$-topology.

PROOF: The topologies must be distinct since different linear functionals are continuous. 👍

This shows that there are an enormous number of weak-type topologies on any infinite-dimensional normed vector space (one for each weak-* dense subspace of X^*), although only a few are useful in practice.

XX.7.5. Exercises

XX.7.5.1. Let c_0 be the Banach space of sequences in $\mathbb{F}$ converging to 0 with supremum norm, and c_{00} the dense subspace of sequences eventually 0 (). Then $c_{00}^* = c_0^* \cong \ell^1$. Let $(\boldsymbol{e}_k)$ denote the standard basis vectors of ℓ^1.

(a) Show that the $\sigma(\ell^1, c_{00})$-topology on ℓ^1 is the topology of coordinatewise convergence. In particular, if (α_k) is any sequence of scalars, then the sequence $(\alpha_k \boldsymbol{e}_k)$ converges to 0 in the $\sigma(\ell^1, c_{00})$-topology. Since (α_k) need not be bounded, the conclusion of XX.7.4.9. does not hold for the weak-* topology from c_{00}.

(b) Show that a sequence $(\boldsymbol{x}_k)$ in ℓ^1 converges to $\boldsymbol{x} \in \ell^1$ in the $\sigma(\ell^1, c_0)$-topology if and only if the sequence is uniformly bounded and converges coordinatewise.

(c) The $\sigma(\ell^1, c_{00})$-topology and the $\sigma(\ell^1, c_0)$-topology are distinct but agree on norm-bounded sets.

(d) A sequence $(\boldsymbol{x}_k)$ in ℓ^1 converges weakly to $\boldsymbol{x} \in \ell^1$ if and only if it converges in norm to $\boldsymbol{x}$ [**?**]. (This does not mean the norm and weak topologies are the same!) ✋

XX.7.5.2. Show that if X is a normed vector space and $\bar{X}$ its completion, then the $\sigma(X^*,X)$- and $\sigma(X^*,\bar{X})$-topologies agree on norm-bounded subsets. [Use (), XX.7.4.19.(ii), and Alaoglu's Theorem.]

XX.7.5.3. (a) Let X be a Banach space. The closed unit ball B of X^* is weak-* compact by Alaoglu's Theorem, thus is the weak-* closed convex hull of its extreme points by the Krein-Milman Theorem (). In particular, B has many extreme points.

(b) The closed unit ball of the Banach space c_0 has no extreme points. Thus there is no Banach space X with $X^* \cong c_0$.

(c) Show that if X is reflexive, then the closed unit ball of X is the norm-closed convex hull of its extreme points. [X is the dual of X^*. Use (a), XX.7.2.8., and that the weak and weak-* topologies coincide.]

(d) The closed unit ball of ℓ^1 is the norm-closed convex hull of its extreme points, although ℓ^1 is not reflexive. But this is false for most dual spaces, e.g. $C([0,1])^*$.

(e) The closed unit ball of the Banach space c of convergent sequences is also the norm-closed convex hull of its extreme points, although c is not the dual of any Banach space. (Any unital C*-algebra has this convex hull property (), but is rarely a dual space ()).

XX.7.5.4. Let X be an infinite-dimensional normed space, and let (X_m) be an increasing sequence of subspaces of X with $dim(X_m) = m$. Set $Y = \overline{\cup_n X_n}$ (norm-closure). For each m let S_m be a finite set of points x_{mj} in X_m of norm m such that every $x \in X_m$, $\|x\| = m$, is within $\frac{1}{m}$ of some x_{mj}. Set $S = \cup_m S_m \subseteq Y$.

(a) Show that S is weakly closed (resp. weakly sequentially closed) in X if and only if it is weakly closed (resp. weakly sequentially closed) in Y. [Use XX.7.2.9.,]

(b) Fix a basic weakly open neighborhood $U := U_{\{\phi_1,\dots,\phi_n\},\epsilon}$ of 0 in Y, with $\|\phi_k\| = 1$ for all k. For sufficiently large $m > n$, there is an $x_m \in X_m$ with $\|x_m\| = m$ and $\phi_k(x_m) = 0$ for all k. [Use III.11.5.17.(iii).]

(c) Fix j with $\|x_m - x_{mj}\| < \frac{1}{m}$. Then $|\phi_k(x_{mj})| < \frac{1}{m}$ for all k. Thus for sufficiently large $m > \frac{1}{\epsilon}$, $x_{mj} \in U$.

(d) Conclude from (c) that 0 is in the weak closure of S. Since $0 \notin S$, S is not weakly closed.

(e) For any M, S contains only finitely many points of norm $\leq M$. Thus the only weakly convergent sequences in S are sequences which are eventually constant. In particular, S is weakly sequentially closed.

This construction is in [Lax02], but is not quite correct there.

XX.7.5.5. Let X be a normed vector space.

(a) Use III.11.5.17. to show that if $f \in X^{**}$, then for any $\phi_1, \dots, \phi_n \in X^*$ there is an $x \in X$ such that $\phi_k(x) = f(\phi_k)$ for all k, $1 \leq k \leq n$. [The values of f on the ϕ_k are determined by the values on a linearly independent subset with the same span.]

(b) Use (a) to give a direct proof that X is weak-* dense in X^{**}.

XX.7.5.6. (a) Show that if X is separable, then the closed unit ball of X^* is sequentially compact in the weak-* topology.

(b) The closed unit ball of $(\ell^\infty)^*$ is not sequentially compact. [Let $\phi_k((x_n)) = x_k$. The sequence (ϕ_k) has no weak-* convergent subsequence.]

(c) If X is reflexive, then the closed unit ball of X^* is weak-* (weakly) sequentially compact. [Every sequence in X^* is contained in a separable closed subspace.] Applying this to X^*, the closed unit ball of X is weakly sequentially compact. Conversely, if the closed unit ball of a Banach space Y is weakly sequentially compact, then Y is reflexive.

XX.7.5.7. Show that in the definition of the $\sigma(X, Y)$-topology (and, in particular, for the weak and weak-* topologies), it suffices to consider basic open sets $U_{\{\phi_1,\ldots,\phi_n\},\epsilon}$ (even with $\epsilon = 1$) for which $\{\phi_1, \ldots, \phi_m\}$ is linearly independent.

XX.7.5.8. Let X be a normed vector space, and Y a subspace of X^*. Give a characterization of the subspaces of X which are $\sigma(X, Y)$-closed generalizing XX.7.4.13. and XX.7.2.9..

XX.7.5.9. Let X be a vector space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, and let Y be a subspace of X^d. Define a $\sigma(X, Y)$-topology on X in an analogous manner to the definition for normed vector spaces. To what extent do the results of this section hold in this situation? Generalize to vector spaces over other topological fields.

XX.8. The Open Mapping Theorem, Closed Graph Theorem, and Uniform Boundedness Theorem

These are three of the most basic theorems of functional analysis. The first two are essentially one theorem, in the sense that either of them can be derived relatively easily from the other. The Uniform Boundedness Theorem is a consequence of the other two, or can be proved independently (it is apparently a "lower-level theorem" since the other two theorems cannot be obviously deduced from Uniform Boundedness). All of them are a consequence of the Baire Category Theorem, although they are not immediate corollaries and take some considerable additional work to prove. We may work over either $\mathbb{R}$ or $\mathbb{C}$ (denoted $\mathbb{F}$, which will usually just be understood); in fact, it suffices to work over $\mathbb{R}$.

XX.8.1. Open Mappings

XX.8.1.1. Recall () that if f is a function from a topological space X to a topological space Y, then f is an *open mapping* if $f(U)$ is open in Y for every open set U in X. An open mapping does not have to be continuous. It is unusual for a continuous function, even a surjective one, to be an open mapping; however, we will show that a surjective bounded linear transformation between Banach spaces is an open mapping. In this subsection, we give alternate characterizations of when linear transformations are open mappings. The arguments make repeated use of the following principle which is an immediate result of linearity:

XX.8.1.2. [Basic Principle of Open Balls] Let X and Y be normed vector spaces, and $T : X \to Y$ a linear transformation. Then the images under T of all open balls in X look essentially the same, up to scaling and translation.

XX.8.1.3. Proposition. Let X and Y be normed vector spaces, and $T : X \to Y$ a linear transformation (not necessarily bounded). The following are equivalent:

(i) T is an open mapping.

(ii) The image under T of some open ball in X contains an open ball in Y.

(iii) The image under T of every open ball in X contains an open ball in Y.

(iv) The image under T of some open ball in X centered at 0 contains an open ball in Y centered at 0.

(v) The image under T of every open ball in X centered at 0 contains an open ball in Y centered at 0.

Proof: (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii) and (i) $\Rightarrow$ (v) $\Rightarrow$ (iv) $\Rightarrow$ (ii) are trivial (since $T(0) = 0$). (ii) $\Rightarrow$ (iii) is an instance of the Basic Principle of Open Balls: if $T(B_r(x_0))$ contains $B_\epsilon(y_0)$, then for any x_1 we have that $T(B_r(x_1))$ contains $B_\epsilon(y_1)$, where $y_1 = y_0 + T(x_1 - x_0)$. In particular, $T(B_r(0))$ contains $B_\epsilon(y_2)$, where $y_2 = y_0 - T(x_0)$. Then, for any s, we have $T(B_s(0))$ contains $B_\delta(\frac{s}{r}y_2)$, where $\delta = \frac{s}{r}\epsilon$, and, for any x_1, $T(B_s(x_1))$ contains $B_\delta(T(x_1) + \frac{s}{r}y_2)$. The proof of (iv) $\Rightarrow$ (v) is similar but easier.

To prove (iii) $\Rightarrow$ (iv), suppose $T(B_r(0))$ contains $B_\epsilon(y_0)$. Then there is an $x_0 \in B_r(0)$ with $T(x_0) = y_0$. Then $T(B_r(-x_0))$ contains $B_\epsilon(0)$. If $s = r + \|x_0\|$, then $B_r(-x_0) \subseteq B_s(0)$, so $T(B_s(0))$ contains $B_\epsilon(0)$.

To finish the proof, it suffices to show (v) $\Rightarrow$ (i). Let U be an open set in X, and $y_0 \in T(U)$. Let $x_0 \in U$ with $T(x_0) = y_0$, and $\delta > 0$ for which $B_\delta(x_0) \subseteq U$. There is an $\epsilon > 0$ such that $T(B_\delta(0))$ contains $B_\epsilon(0)$ by

(v). Then $T(B_\delta(x_0))$ contains $B_\epsilon(y_0)$, so $T(U)$ contains $B_\epsilon(y_0)$. Since $y_0 \in T(U)$ is arbitrary, $T(U)$ is open. 👍

A slight modification of a subset of the proof of XX.8.1.3. (adding the words "the closure of" in a few places) gives the following:

XX.8.1.4. Proposition. Let X and Y be normed vector spaces, and $T : X \to Y$ a linear transformation (not necessarily bounded). The following are equivalent:

(ii) The closure of the image under T of some open ball in X contains an open ball in Y.

(iii) The closure of the image under T of every open ball in X contains an open ball in Y.

(iv) The closure of the image under T of some open ball in X centered at 0 contains an open ball in Y centered at 0.

(v) The closure of the image under T of every open ball in X centered at 0 contains an open ball in Y centered at 0.

XX.8.2. The Open Mapping Theorem

The main result of this subsection is:

XX.8.2.1. Theorem. [Open Mapping Theorem] Let X and Y be Banach spaces, and $T : X \to Y$ a surjective bounded operator. Then T is an open mapping.

A crucial special case is when T is injective as well as surjective:

XX.8.2.2. Corollary. [Bounded Inverse Theorem] Let X and Y be Banach spaces, and $T : X \to Y$ a bijective bounded operator. Then $T^{-1} : Y \to X$ is also bounded. Equivalently, T is bounded below, i.e. a homeomorphism.

Proof: T^{-1} is continuous, hence bounded by (), if and only if T is an open mapping, which follows from the Open Mapping Theorem. 👍

XX.8.2.3. We can derive the Open Mapping Theorem from the Bounded Inverse Theorem, as follows. Let X and Y be Banach spaces and $T : X \to Y$ a surjective bounded operator. Let N be the null space of T; then N is a closed subspace of X, so we may form the quotient space X/N which is a Banach space under the quotient norm (). Let $P : X \to X/N$ be the projection map. Then there is an injective bounded operator $\tilde{T} : X/N \to Y$ with $T = \tilde{T} \circ P$ (). $\tilde{T}$ is an open mapping by the Bounded Inverse Theorem. P is also an open mapping: this is a general fact about quotient spaces (), but in this instance it is also an immediate corollary of XX.8.1.3., since it follows from the definition of the quotient norm () that the image of the open unit ball in X is exactly the open unit ball in X/N. Thus T is a composition of two open mappings, hence also an open mapping. Hence the Open Mapping Theorem and the Bounded Inverse Theorem are equivalent.

Although the Open Mapping Theorem can be easily deduced from the Bounded Inverse Theorem, there does not appear to be a proof of the Bounded Inverse Theorem which is simpler than the proof of the Open Mapping Theorem.

XX.8.2.4. Examples. All the hypotheses of the Open Mapping Theorem are necessary in general, as the following examples show. These examples illustrate that the statement of the Open Mapping Theorem is a nontrivial assertion which may even be considered somewhat surprising.

(i) Let $X = Y = \ell^2$, and define $T : X \to Y$ by

$$T((\xi_1, \xi_2, \dots)) = \left(\xi_1, \frac{1}{2}\xi_2, \dots, \frac{1}{n}\xi_n, \dots\right) .$$

Then T is injective and has dense range, but is not surjective. T is bounded ($\|T\| = 1$). T is not an open map since no positive scalar multiple of $(1, \frac{1}{2}, \dots, \frac{1}{n}, \dots)$ is in the range. Thus the hypothesis that T is surjective is necessary. (In fact, a linear map which is not surjective can never be an open map, since the range is a subspace and a proper subspace can never be an open set.)

(ii) Take the same example as in (i) but let Y be the range of T. Y is a dense subspace of ℓ^2 and inherits a norm from ℓ^2, but is not complete. T is now a bijection from $X = \ell^2$ to Y, but T^{-1} is not bounded since T is not bounded below. Thus the hypothesis that Y be complete is necessary.

(iii) It is trickier to give an example showing that the hypothesis that X be a Banach space is necessary. Let Z be a Banach space and N a closed subspace of Z. Set $Y = Z/N$ with the quotient norm (), and $P : Z \to Z/N$ the quotient map (). Suppose X is a subspace of Z complementary to N which is not closed (e.g. X is the null space of a discontinuous linear functional on Z and N a complementary one-dimensional subspace; cf. also ()). X is a normed vector space with norm inherited from Z, and $T = P|_X$ is an injective bounded operator from X onto Y. But $T^{-1} : Y \to X$ cannot be bounded, since if it were X would be complete, hence closed in Z.

(iv) The hypothesis that T be bounded is also necessary. Expand the standard orthonormal basis of ℓ^2 to a Hamel basis, and take a permutation of the Hamel basis interchanging the orthonormal basis with a subset of the remaining Hamel basis vectors. This permutation extends to a bijective but unbounded operator T from ℓ^2 to ℓ^2. If T were an open mapping, then T^{-1} would be continuous, hence bounded, so T would be bounded by the Bounded Inverse Theorem, a contradiction.

Note that some form of the Axiom of Choice is needed in (iii) and (iv). (Actually, some form of Choice is even needed for the Open Mapping Theorem, since the Baire Category Theorem depends on DC ().)

XX.8.2.5. We now give the proof of the Open Mapping Theorem.

PROOF: Write B_r^X and B_r^Y for the open balls of radius r around 0 in X and Y respectively. We have that

$$\cup_{n=1}^{\infty} T(B_n^X) = T(X) = Y$$

since T is surjective. Thus, since Y is a Banach space, by the Baire Category Theorem $T(B_n^X)$ is not nowhere dense in Y for some n, i.e. $\overline{T(B_n^X)}$ contains an open ball in Y. Thus by XX.8.1.4., $\overline{T(B_r^X)}$ contains an open ball around 0 in Y for any r. In particular, there is an $\epsilon > 0$ such that $\overline{T(B_1^X)}$ contains B_ϵ^Y. We will show that $T(B_2^X)$ contains B_ϵ^Y; it will then follow from XX.8.1.3. that T is an open mapping.

Let $y \in B_\epsilon^Y$; we will construct an $x \in X$ with $\|x\| < 2$ and $Tx = y$. A key observation is that $\overline{T(B_r^X)}$ contains $B_{\epsilon r}^Y$ for every $r > 0$. Since y is in the closure of $T(B_1^X)$, for any $\delta > 0$ we can find an $x_0 \in X$, $\|x_0\| < 1$, with $\|y - Tx_0\| < \delta$. Let $x_1 \in X$, $\|x_1\| < 1$, with

$$\|y - Tx_1\| < \frac{\epsilon}{2} .$$

Since $y - Tx_1$ is in the closure of $T(B_{1/2}^X)$, there is an $x_2 \in X$, $\|x_2\| < \frac{1}{2}$, with

$$\|(y - Tx_1) - Tx_2)\| < \frac{\epsilon}{4} .$$

Continuing inductively, we obtain for each n an $x_n \in X$ with $\|x_n\| < 2^{-n+1}$ and

$$\|y - Tx_1 - Tx_2 - \cdots - Tx_n\| < 2^{-n}\epsilon .$$

The infinite series

$$\sum_{k=1}^{\infty} x_k$$

converges absolutely, hence converges to some $x \in X$ since X is a Banach space (). We have

$$\|x\| < \sum_{k=1}^{\infty} 2^{-k+1} = 2 .$$

Since T is continuous, we have

$$Tx = \sum_{k=1}^{\infty} Tx_k$$

in Y (). But we have

$$\left\| y - \sum_{k=1}^{n} Tx_k \right\| < 2^{-n}\epsilon$$

for all n, and thus the infinite series

$$\sum_{k=1}^{\infty} Tx_k$$

converges to y, i.e. $Tx = y$, completing the proof. 👍

XX.8.3. The Closed Graph Theorem

XX.8.3.1. Recall () that if X and Y are sets and $f : X \to Y$ is a function, the set

$$\Gamma(f) = \{(x, y) \in X \times Y : y = f(x)\}$$

is the *graph* of f. (Technically, using the ordered pair definition of function, a function *is* its graph.) If X and Y are topological spaces, then f has *closed graph* if $\Gamma(f)$ is closed in $X \times Y$.

XX.8.3.2. PROPOSITION. If X and Y are topological spaces and $f : X \to Y$ is a function, then f has closed graph if and only if, whenever (x_i) is a net in X with $x_i \to x$ and $f(x_i) \to y$ for some $x \in X$ and $y \in Y$, we have $y = f(x)$.

PROOF: By definition, f has closed graph if and only if, whenever $((x_i, y_i))$ is a net in $\Gamma(f)$ converging to $(x, y) \in X \times Y$, we have $(x, y) \in \Gamma(f)$. The statement of the proposition is a simple rephrasing. 👍

A continuous function between Hausdorff spaces has closed graph (), but a function between Hausdorff spaces which has closed graph need not be continuous (). However, we have the following, which is the main result of this subsection:

XX.8.3.3. THEOREM. [CLOSED GRAPH THEOREM] Let X and Y be Banach spaces, and $T : X \to Y$ be a linear transformation. Then T is bounded (continuous) if and only if T has closed graph.

PROOF: If T is continuous, it has closed graph by (). Conversely, suppose T has closed graph. We have that $X \times Y$ is a Banach space under a number of equivalent norms (). Fix one of these norms, say

$$\|(x, y)\| = \max(\|x\|, \|y\|) \ .$$

Since T is linear, $\Gamma(T)$ is a subspace of $X \times Y$; by assumption, it is closed, hence a Banach space under the induced norm. The map $S : \Gamma(T) \to X$ given by $S((x, Tx)) = x$ is bijective and linear, and bounded ($\|S\| \leq 1$). Thus by the Bounded Inverse Theorem $S^{-1} : X \to \Gamma(T)$ is bounded. If P is the projection from $X \times Y$ to Y, then P is bounded, and hence $T = P \circ S^{-1}$ is also bounded. 👍

XX.8.3.4. The proof shows that the Closed Graph Theorem is a simple corollary of the Bounded Inverse Theorem. The converse is true too: if X and Y are Banach spaces and $T : X \to Y$ is a continuous (bounded) bijection, then $\Gamma(T)$ is closed. Thus $\Gamma(T^{-1})$ is also closed since it is essentially the same as $\Gamma(T)$ with coordinates reversed. So T^{-1} is bounded by the Closed Graph Theorem.

We thus obtain that the Open Mapping Theorem, the Bounded Inverse Theorem, and the Closed Graph Theorem are equivalent statements by simple arguments.

XX.8.3.5. The importance of the Closed Graph Theorem is that it is often much easier to prove that a map $T : X \to Y$ has closed graph than to directly prove that it is continuous or bounded. To prove directly that T is continuous, it must be shown that if (x_n) is a sequence in X with $x_n \to x$, then the sequence (Tx_n) converges to Tx. The sequence (Tx_n) cannot be assumed to converge or even to have a convergent subsequence. But to show that T has closed graph, one only needs to show that if (x_n) is a sequence in X with $x_n \to x$, *and* $Tx_n \to y$ *for some* $y \in Y$, then $y = Tx$, i.e. the sequence (Tx_n) can be assumed to converge to something. The next result is a good illustration of the use of the Closed Graph Theorem.

XX.8.3.6. PROPOSITION. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces, and $T : \mathcal{H}_1 \to \mathcal{H}_2$ a linear operator. If T has an adjoint $T^* : \mathcal{H}_2 \to \mathcal{H}_1$, i.e. T^* is a function (which is automatically a linear operator) satisfying

$$\langle x, T^* y\rangle_{\mathcal{H}_1} = \langle Tx, y\rangle_{\mathcal{H}_2}$$

for all $x \in \mathcal{H}_1$, $y \in \mathcal{H}_2$, then T is bounded (and hence T^* is also bounded).

PROOF: We need only show that T has closed graph. Suppose (x_n) is a sequence in $\mathcal{H}_1$ with $x_n \to x$ and $Tx_n \to y$ for some $y \in \mathcal{H}_2$. Then, for any $z \in \mathcal{H}_2$, we have

$$\langle Tx_n, z\rangle_{\mathcal{H}_2} \to \langle y, z\rangle_{\mathcal{H}_2}$$

by continuity of the inner product in $\mathcal{H}_2$ (). On the other hand, we have

$$\langle Tx_n, z\rangle_{\mathcal{H}_2} = \langle x_n, T^* z\rangle_{\mathcal{H}_1} \to \langle x, T^* z\rangle_{\mathcal{H}_1} = \langle Tx, z\rangle_{\mathcal{H}_2}$$

by continuity of the inner product in $\mathcal{H}_1$. Thus by uniqueness of limits we have

$$\langle y, z\rangle_{\mathcal{H}_2} = \langle Tx, z\rangle_{\mathcal{H}_2}$$

for every $z \in \mathcal{H}_2$, and so $y = Tx$ by ().

XX.8.3.7. **Caution:** Our convention is that "T is a linear map from X to Y" means that the domain of T is all of X. In some references, this phrase means only that the domain of T is a subspace (usually dense) of X. The Closed Graph Theorem does *not* apply in this situation; in fact, such an operator can be unbounded but have closed graph. Densely defined unbounded operators with closed graph, called *closed operators*, are very important; see section (). Proposition XX.8.3.6. is also false in general for operators which are just densely defined (with densely defined adjoint); cf. ().

XX.8.3.8. Since in the statement of the Closed Graph Theorem T is not required to be surjective, one might think the hypothesis that Y be complete is unnecessary: Y could just be replaced by its completion $\bar{Y}$. But this is incorrect. The graph of T can be a closed subset of $X \times Y$ but not closed in $X \times \bar{Y}$.

For a counterexample, let X, Y, T be as in XX.8.2.4.(iii). Then the graph of T is closed in $X \times Y$ since T is continuous, so the graph of $T^{-1} : Y \to X$ is closed in $Y \times X$. But T^{-1} is not bounded.

XX.8.4. The Uniform Boundedness Theorem

The Uniform Boundedness Theorem is a theorem which states that a family of bounded operators on a Banach space which is "pointwise bounded" is actually uniformly bounded. This result, like the Open Mapping Theorem and Closed Graph Theorem, has many uses. It can be obtained as a simple consequence of the Closed Graph Theorem or proved directly from scratch; we will give three arguments.

XX.8.4.1. THEOREM. [UNIFORM BOUNDEDNESS THEOREM] Let X be a Banach space, $\{Y_i : i \in I\}$ a collection of normed vector spaces, and $T_i : X \to Y_i$ a bounded operator for each i. Suppose, for each $x \in X$, the set

$$\{\|T_i(x)\| : i \in I\}$$

is bounded (the bound may depend on x). Then the set

$$\{\|T_i\| : i \in I\}$$

is bounded.

XX.8.4.2. In other words, if for each $x \in X$ there is a constant M_x such that $\|T_i x\| \leq M_x$ for all $i \in I$, then there is a single constant M such that $\|T_i\| \leq M$ for all $i \in I$.

The theorem is commonly applied in the case where all the Y_i are the same, and/or when $I = \mathbb{N}$, but these restrictions do not make the proof any easier.

The theorem is sometimes stated in the contrapositive, where it is called the *Resonance Theorem*:

XX.8.4.3. THEOREM. [RESONANCE THEOREM] Let X be a Banach space, $\{Y_i : i \in I\}$ a collection of normed vector spaces, and $T_i : X \to Y_i$ a bounded operator for each i. If $\sup_i \|T_i\| = +\infty$, then there is an $x \in X$ such that $\sup_i \|T_i(x)\| = +\infty$.

The idea, and origin of the name, is that if the T_i "blow up" globally, then there is at least one specific point ("point of resonance") where the T_i blow up.

XX.8.4.4. PROOF: Our first proof deduces the theorem from the Closed Graph Theorem.

First note that since the T_i are not assumed to be surjective, there is no harm in replacing Y_i by its completion. Thus we may assume without loss of generality that each Y_i is a Banach space.

Let Y be the ℓ^∞ product of the Y_i (). Then Y is a Banach space. Define $T : X \to Y$ by

$$T(x) = (\cdots T_i(x) \cdots)$$

(the pointwise boundedness hypothesis exactly means that $T(x) \in Y$ for all $x \in X$). Then T is a linear operator; T is bounded if and only if the T_i are uniformly bounded (in fact, $\|T\| = \sup_i \|T_i\|$).

To show T is bounded, it suffices to show that T has closed graph. But this is easy. Suppose (x_n) is a sequence in X and $x_n \to x$ and $Tx_n \to y$ for some $x \in X$, $y \in Y$. Set $y = (\cdots y_i \cdots)$. Since convergence in Y implies coordinatewise convergence, we have $T_i x_n \to y_i$ for all i. But since T_i is bounded, we have $T_i x_n \to T_i x$, i.e. $y_i = T_i x$ for all i, and thus $y = Tx$. 👍

XX.8.4.5. PROOF: We give a second direct proof from the Baire Category Theorem.

For $n \in \mathbb{N}$, define

$$A_n = \{x \in X : \|T_i x\| \leq n \text{ for all } i \in I\}$$

and note that A_n is a closed subset of X ($A_n = \cap_{i \in I} T_i^{-1}(\{y \in Y_i : \|y\| \leq n\})$).

By the pointwise boundedness assumption, $\cup_n A_n = X$, so by the Baire Category Theorem some A_n has nonempty interior. Thus there is an $x_0 \in X$ and $\epsilon > 0$ such that $\|T_i x\| \leq n$ whenever $\|x - x_0\| < \epsilon$. Let m be a number such that $\|T_i(x_0)\| \leq m$ for all $i \in I$. Then, if $\|x\| < \epsilon$, we have $\|[x + x_0] - x_0\| < \epsilon$, so, for all i,

$$\|T_i(x + x_0)\| \leq n$$

$$\|T_i x\| = \|T_i([x + x_0] - x_0)\| = \|T_i(x + x_0) - T_i(x_0)\| \leq \|T_i(x + x_0)\| + \|T_i(x_0)\| \leq n + m$$

which implies that

$$\|T_i\| \leq \frac{n + m}{\epsilon}$$

for all i. 👍

XX.8.4.6. PROOF: Our third proof is the most elementary, using a "gliding hump" argument. This proof is due to HAUSDORFF [Hau32] (cf. [Hen80]).

Suppose $\{\|T_i\| : i \in I\}$ is unbounded. Inductively define sequences (x_n) in X and $i_n \in I$ as follows. Suppose x_k and i_k have been defined for $k < n$. Set $M_0 = 0$, and if $n > 1$ set

$$M_{n-1} = \sup_i \{T_i(x_1 + \cdots + x_{n-1})\} .$$

Now choose i_n such that

$$\|T_{i_n}\| > 3 \cdot 4^n [n + M_{n-1}]$$

and x_n such that

$$\|x_n\| = 4^{-n}$$

$$\|T_{i_n} x_n\| > \frac{2}{3} \|T_{i_n}\| \|x_n\| .$$

Set

$$x = \sum_{n=1}^{\infty} x_n$$

(note that the series converges absolutely, hence converges since X is complete). For each n set

$$z_n = x_1 + \cdots + x_{n-1}$$

(set $z_1 = 0$)

$$y_n = \sum_{k=n+1}^{\infty} x_k$$

and note that $\|y_n\| \leq \frac{4^{-n}}{3} = \frac{1}{3}\|x_n\|$. We then have

$$\|T_{i_n} x_n\| > 2[n + M_{n-1}]$$

$$\|T_{i_n} y_n\| \leq \|T_{i_n}\| \|y_n\| \leq \frac{1}{3}\|T_{i_n}\| \|x_n\| < \frac{1}{2}\|T_{i_n} x_n\|$$

for each n. Thus for each n we have

$$\|T_{i_n} x\| = \|T_{i_n}(x_n + y_n + z_n)\| \geq [\|T_{i_n} x_n\| - \|T_{i_n} y_n\|] - \|T_{i_n} z_n\| > [n + M_n] - M_n = n$$

which contradicts that $\{\|T_i x\| : i \in I\}$ is bounded. 👍

Here is a typical application of the Uniform Boundedness Theorem:

XX.8.4.7. PROPOSITION. Let X be a Banach space. If (ϕ_n) is a weak-* convergent sequence in X^*, then (ϕ_n) is uniformly bounded, i.e. there is a K such that $\|\phi_n\| \leq K$ for all n.

PROOF: We have $X^* = \mathcal{B}(X, \mathbb{F})$. For each $x \in X$, the sequence $(\phi_n(x))$ is a convergent sequence in $\mathbb{F}$, hence bounded (). 👍

This is only true for *sequences*, not nets! In fact, even a convergent net of real numbers is not necessarily bounded. ✋

XX.8.5. Exercises

XX.8.5.1. Let X and Y be normed vector spaces, with X complete, and $T : X \to Y$ a linear operator.
(a) Show that T is bounded if and only if $T^{-1}(B_Y)$ is closed, where B_Y is the closed unit ball of Y. [One direction is trivial from continuity. Conversely, suppose $T^{-1}(B_Y)$ is closed. Then $T^{-1}(B)$ is closed for every closed ball B in Y. Apply the Baire Category Theorem to conclude that $T^{-1}(B_Y)$ contains an open ball.]
(b) Let $T : c_{00} \to c_{00}$ () be defined by

$$T(\xi_1, \xi_2, \dots) = (\xi_1, 2\xi_2, \dots, n\xi_n, \dots) \ .$$

Show that $T^{-1}(B_{c_{00}})$ is closed. Thus the result does not necessarily hold if X is not complete.
(c) What is the relation between this property of T and having closed graph?

XX.9. Compact Convex Sets

A fairly detailed analysis can be given of the structure of compact convex sets in locally convex topological vector spaces. We describe the basics of this theory in this section; for more detailed analyses, see [Alf71], [**?**],

We will work throughout this section with *real* topological vector spaces. The results will immediately apply also to complex topological vector spaces, since a complex topological vector space has an underlying structure as a real topological vector space and none of the statements or proofs of the results make use of complex scalar multiplication. To avoid unnecessary subscripts, if K is a compact Hausdorff space, in this section write $C(K)$ for the set $C_{\mathbb{R}}(K)$ of real-valued continuous functions on K.

XX.9.1. Convex Sets in Topological Vector Spaces

The notion of convexity makes sense in any vector space over the real numbers (or field containing the real numbers such as $\mathbb{C}$), but the primary applications in analysis concern convex sets in topological vector spaces, usually with additional topological properties.

The first simple observation is a consequence of the continuity of addition and scalar multiplication:

XX.9.1.1. Proposition. Let C be a convex set in a topological vector space X. Then the closure $\bar{C}$ and the interior C° of C are also convex (C° may be empty in general).

XX.9.1.2. A convex set with empty (topological) interior (e.g. $[x, y]$ for $x \neq y$ in $\mathbb{R}^2$) can have inner points in the sense of XX.1.3.7.. However:

XX.9.1.3. Proposition. Let C be a convex set in a topological vector space X. Any internal point of C (in particular, any interior point of C) is an inner point.

We modify the definition of isomorphism of convex sets (XX.1.2.9.) to one more appropriate in topological vector spaces:

XX.9.1.4. Definition. Let C and D be convex sets in topological vector spaces X and Y respectively. An *isomorphism* from C to D is an affine homeomorphism from C to D.

Thus an isomorphism from C to D is a bijection which preserves both the convex structure and the topology.

XX.9.2. Faces and Extreme Points

The crucial notion underlying the whole analysis of compact convex sets is that of a face (XX.1.3.1.).

Closed Faces

Let X be a topological vector space, and K a convex subset of X. The faces of K which are relatively closed in K are particularly important, and are called the *closed faces* of K. This situation is especially important if K is closed in X, although we do not make this assumption in the definition. Note that (if X is Hausdorff) every extreme point is a closed face.

XX.9.2.1. It looks at first from the definition that a face should be closed; in fact, if F is a face in K and Y is a one-dimensional (more generally, finite-dimensional) affine subspace of X, then $F \cap Y$ is always relatively closed in $K \cap Y$. However, a face of a convex set, even a compact convex set (Choquet simplex), need not be closed in general, and the closure of a face is not necessarily a face (XX.9.8.2.). ✋

XX.9.2.2. As in XX.1.3.6.(i), an arbitrary intersection of closed faces of a convex set K, if nonempty, is a closed face (the other parts of XX.1.3.6. have straightforward analogs for closed faces too, with "face" replaced by "closed face" throughout). Thus if S is an arbitrary nonempty subset of K, there is a smallest closed face $\bar{F}(S)$ of K containing S, the intersection of all closed faces containing S.

XX.9.2.3. Similarly, if K is a convex set in a topological vector space X, and ϕ is a continuous linear functional on X, a support face for K from ϕ (XX.1.3.11.) is a closed face of K.

Convex Hulls and Extreme Points

XX.9.2.4. PROPOSITION. Let X be a topological vector space, and $K_1, \ldots, K_n$ a finite collection of compact convex sets in X. Let K be the convex hull of $\cup_{j=1}^n K_j$. Then K is compact.

PROOF: Since each K_j is convex, it is easy to check that

$$\left\{ \sum_{j=1}^n \alpha_j x_j : x_j \in K_j, \alpha_j \geq 0, \sum_{j=1}^n \alpha_j = 1 \right\}$$

is convex; since it is contained in K and contains each K_j, it equals K. Set

$$S = \left\{ (\alpha_1, \ldots, \alpha_n) : \alpha_j \geq 0, \sum_{j=1}^n \alpha_j = 1 \right\} .$$

Then S is a compact subset of $\mathbb{R}^n$ (it is an $(n-1)$-simplex), so the map

$$(\alpha_1, \ldots, \alpha_n, x_1, \ldots, x_n) \mapsto \sum_{j=1}^n \alpha_j x_j$$

is a continuous function from the compact set $S \times K_1 \times \cdots \times K_n$ onto K. Thus K is compact by (). 👍

The next result is known as *Milman's (partial) converse to the Krein-Milman Theorem*. Our proof is adapted from [SW99].

XX.9.2.5. THEOREM. Let S be a subset of a locally convex (Hausdorff) topological vector space X such that the closed convex hull K of S is compact. Then all extreme points of K lie in $\bar{S}$.

PROOF: Since $\bar{S}$ is a subset of K, it is compact. Let U be an open convex neighborhood of 0 in X. For $x \in \bar{S}$, set $U_x = x + U$. The U_x form an open cover of $\bar{S}$, so there is a finite subcover $\{U_{x_1}, \ldots, U_{x_n}\}$. Let K_j be the closed convex hull of $S \cap \bar{U}_{x_j}$. Then $K_j \subseteq K$ for all j. The convex hull K' of $\cup_{j=1}^n K_j$ contains S, $K' \subseteq K$, and K' is compact by XX.9.2.4.; thus $K' = K$. So if p is an extreme point of K, we have $p = \sum_{j=1}^n \alpha_j y_j$ for some α_j's in S and $y_j \in K_j$; since p is extreme, we must have $p = y_j$ for some j, i.e. $p \in K_j$. Since $\bar{U}_{x_j}$ is closed and convex, $K_j \subseteq \bar{U}_{x_j} \subseteq \bar{S} + \bar{U}$. Thus $p \in \bar{S} + \bar{U}$. This is true for every U, so $p \in \bar{S}$. 👍

XX.9.3. Compact Convex Sets in Finite-Dimensional Spaces

We now look at compact convex sets in finite-dimensional topological vector spaces. A finite-dimensional real vector space has a unique topology making it a Hausdorff topological vector space, which can be identified with $\mathbb{R}^n$ for some n. So without loss of generality we will work in Euclidean space. We can thus use geometric arguments in our analysis.

XX.9.3.1. If K is a convex set in $\mathbb{R}^n$, then K spans an m-dimensional affine subspace of $\mathbb{R}^n$ for some $m \leq n$; we say K is *m-dimensional* (this terminology is justified in **??**). By replacing $\mathbb{R}^n$ with this subspace and identifying it with $\mathbb{R}^m$, we may if desired for analysis of K assume it spans all of $\mathbb{R}^n$ as an affine space (i.e. is n-dimensional), although we will not routinely make this restriction. Dimension in this sense is an isomorphism invariant of K.

XX.9.3.2. PROPOSITION. If K is an n-dimensional convex set in $\mathbb{R}^n$, then K has nonempty (topological) interior K°; in fact K° is an open convex set in $\mathbb{R}^n$ which is dense in K.

PROOF: K contains $n+1$ points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_n$ which span $\mathbb{R}^n$ as an affine space (any set which spans $\mathbb{R}^n$ as an affine space contains $n+1$ points which also span it). Set $\boldsymbol{x} = \frac{1}{n+1}\sum_{k=0}^n \boldsymbol{x}_k$. Every point $\boldsymbol{y}$ of $\mathbb{R}^n$ can be uniquely written as $\sum_{k=0}^n \alpha_k \boldsymbol{x}_k$ for $\alpha_k \in \mathbb{R}$, $\sum_{k=0}^n \alpha_k = 1$; if $\boldsymbol{y}$ is sufficiently close to $\boldsymbol{x}$, each α_k is close to $\frac{1}{n+1}$ and is hence positive. Thus $\boldsymbol{y} \in K$ if $\boldsymbol{y}$ is sufficiently close to $\boldsymbol{x}$, i.e. K contains an open ball around $\boldsymbol{x}$. Thus $\boldsymbol{x}$ is in K°, so $K^\circ \neq \emptyset$. The density and convexity of K° follow from (). 👍

This result fails for infinite-dimensional compact convex sets, since it implies that the spanned affine subspace contains a nonempty open set with compact closure.

Combining this with the geometric version of the Hahn-Banach Theorem in () (which is elementary in a finite-dimensional vector space, not requiring any form of Choice), we obtain:

XX.9.3.3. PROPOSITION. Let K be an n-dimensional convex set in $\mathbb{R}^n$. If $\boldsymbol{x} \in K$ is in the topological boundary of K, then $\boldsymbol{x}$ is contained in a proper face of K, and hence is not an interior point in the sense of XX.1.3.7..

PROOF: By () applied to K° and $\boldsymbol{x}$, there is a linear functional ϕ on $\mathbb{R}^n$, which is automatically continuous, and a number r such that $\phi(\boldsymbol{x}) \geq r$ and $\phi(\boldsymbol{y}) < r$ for all $\boldsymbol{y} \in K^\circ$. Since ϕ is continuous, $\phi(\boldsymbol{y}) \leq r$ for

all $\boldsymbol{y} \in \overline{K^\circ}$, and in particular $\phi(\boldsymbol{x}) = r$ and $\phi(\boldsymbol{y}) \leq r$ for all $\boldsymbol{y} \in K$ since K° is dense in K. The set $F = \{\boldsymbol{y} \in K : \phi(\boldsymbol{y}) = r\}$ is thus a proper face of K containing $\boldsymbol{x}$. 👍

This result fails in the infinite-dimensional case, and even in the finite-dimensional case if the dimension of the convex set is strictly smaller than the dimension of the space.

Extreme Points

Our next goal is to show that a compact convex set K in a finite-dimensional vector space has "many" extreme points, enough that K is the convex hull of its extreme points. This result is a special case of the Krein-Milman Theorem, although we will obtain a sharper result with a more elementary proof in the finite-dimensional case.

XX.9.3.4. Example. It can be shown that if K is a compact convex set in $\mathbb{R}^2$, then the set of extreme points of K is closed in K, hence also compact. But this fails in $\mathbb{R}^3$. Let K be the convex hull of the points of $C = \{(x, y, 0) : (x-1)^2 + y^2 = 1\}$ and the points $(0, 0, 1)$ and $(0, 0, -1)$; see Fig. (). The extreme points of K are the points of C except the origin, and the points $(0, 0, 1)$ and $(0, 0, -1)$. The origin is in the closure of the set of extreme points but is not itself an extreme point.

XX.9.3.5. Proposition. An n-dimensional compact convex set K has at least $n + 1$ extreme points. The extreme points of K span the same affine subspace as K.

XX.9.3.6. Lemma. A nonempty compact convex set in $\mathbb{R}^n$ has an extreme point.

Proof: Let K be a compact convex set in $\mathbb{R}^n$, and $\boldsymbol{x} \in K$. There is a point $\boldsymbol{y} \in K$ of maximum distance d from $\boldsymbol{x}$ by compactness ($\boldsymbol{y}$ is not necessarily unique). Then $\boldsymbol{y}$ is an extreme point of the ball $\bar{B}_d(\boldsymbol{x})$, which contains K, so $\boldsymbol{y}$ is an extreme point of K. 👍

In fact, if K contains more than one point, it contains a second extreme point too, any point of maximum distance from $\boldsymbol{y}$. This result and proof hold more generally in any strictly convex Banach space. The result is true in much greater generality (XX.9.4.1.), but requires a different proof.

Proof: Proof of XX.9.3.5.: let K be an n-dimensional compact convex subset of $\mathbb{R}^n$. Let A be the affine subspace spanned by the extreme points of K (there is at least one extreme point by XX.9.3.6.). If A does not contain K, then there is a point $\boldsymbol{x}$ of K of maximum distance $d > 0$ from A. If S is the set of points of K of distance d from A, then the face $F(\{\boldsymbol{x}\})$ of K generated by $\boldsymbol{x}$ is contained in S. $F(\{\boldsymbol{x}\})$ is a compact convex set and thus has an extreme point $\boldsymbol{y}$ (XX.9.3.6.), which must be an extreme point of K also. But $\boldsymbol{y} \in S$, so $\boldsymbol{y} \notin A$, a contradiction. The first statement then follows since the affine subspace spanned by m points has dimension $\leq m - 1$. 👍

The next result is the main theorem of this subsection, due first to Minkowski without the $n + 1$, and in the sharp form with $n + 1$ by Carathéodory.

XX.9.3.7. THEOREM. Let K be a compact convex set in $\mathbb{R}^n$. Then every point of K is a convex combination of at most $n+1$ extreme points of K.

PROOF: By complete induction on the dimension m of K. If $m = 1$, K is isomorphic to a closed bounded interval and is the convex hull of its two endpoints (and the case $m = 0$ is trivial). For a general K of dimension m, let $\boldsymbol{x} \in K$. If $\boldsymbol{x}$ is in the topological boundary of K, then $\boldsymbol{x}$ lies in a proper face F of K (XX.9.3.3.), and thus $\boldsymbol{x}$ is a convex combination of at most m extreme points of F by the inductive hypothesis, all of which are extreme points of K. If $\boldsymbol{x}$ is an interior point of K, let $\boldsymbol{y}$ be an extreme point of K. The ray from $\boldsymbol{y}$ through $\boldsymbol{x}$ intersects the boundary of K in a (unique) point $\boldsymbol{z}$, which is in K since K is compact. Then as before $\boldsymbol{z}$ is a convex combination of at most m extreme points in a proper face of K. Since $\boldsymbol{x}$ is a convex combination of $\boldsymbol{z}$ and $\boldsymbol{y}$, $\boldsymbol{x}$ is a convex combination of $\boldsymbol{y}$ and these at most m other extreme points of K. 👍

Combining this result with XX.9.2.5., we get:

XX.9.3.8. COROLLARY. Let E be a compact subset of $\mathbb{R}^n$. Then the convex hull K of E is also compact.

PROOF: Let

$$S = \left\{ (\alpha_0, \ldots, \alpha_n) : \alpha_j \geq 0, \sum_{j=0}^{n} \alpha_j = 1 \right\} \subseteq \mathbb{R}^{n+1}$$

and consider the map $f : S \times E^{n+1} \to K$ defined by

$$f(\alpha_0, \ldots, \alpha_n, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_n) = \sum_{j=0}^{n} \alpha_j \boldsymbol{x}_j \ .$$

Then f is continuous. Each extreme point of K is in E by XX.9.2.5., and every element of K is a convex combination of at most $n+1$ extreme points by XX.9.3.7., so f is surjective. $S \times E^{n+1}$ is compact, so K is compact by (). 👍

XX.9.4. The Krein-Milman Theorem

XX.9.4.1. THEOREM. [KREIN-MILMAN THEOREM] Let K be a nonempty compact convex set in a locally convex (Hausdorff) topological vector space X. Then K is the closed convex hull of its extreme points (and in particular has extreme points).

PROOF: Recall that we regard X as a *real* topological vector space by ignoring the complex structure if necessary, since all relevant notions (convexity, compactness, etc.) depend only on the real topological vector space structure. Let $\mathcal{F}$ be the collection of all closed faces of K. Partially order $\mathcal{F}$ by reverse inclusion: if $E, F \in \mathcal{F}$, then $E \leq F$ if $F \subseteq E$. If $\mathcal{C} = \{F_i : i \in I\}$ is a chain in $\mathcal{F}$, then $\mathcal{C}$ has an upper bound, namely $F = \cap_i F_i$, which is a closed face (XX.9.2.2.) since it is nonempty because $\mathcal{C}$ is a nested family of compact subsets of K (this is the first of several places compactness of K is used).

Thus, by Zorn's Lemma, $\mathcal{F}$ has a maximal element, i.e. K has a minimal closed face E. We must have that E is a singleton: if $x, y \in E$, $x \neq y$, there is a continuous linear functional ϕ on X with $\phi(x) \neq \phi(y)$ (this is the first place where we need that X is locally convex and Hausdorff). Then $\{\phi(z) : z \in E\}$ is a compact subset of $\mathbb{R}$ (actually a closed bounded interval) since E is compact. If λ is the largest number in this set, then $F = \{z \in E : \phi(z) = \lambda\}$ is a closed face of E and hence of K, and is a proper subset of E since x and y cannot both be in F, contradicting that E is a minimal closed face of K. Thus $E = \{p\}$ for an extreme point p of K, and K has an extreme point. A stronger conclusion also holds: if F is any closed face of K, then F is itself a nonempty compact convex set in X, and so has a point which is an extreme point of F and thus an extreme point of K. Therefore every closed face of K contains an extreme point of K.

Now let L be the closed convex hull of the extreme points of K. If $L \neq K$, let $x \in K \setminus L$. Then, since L is compact and convex, there is a continuous linear functional ψ on X with

$$\psi(x) > \sup\{\psi(y) : y \in L\}$$

by the geometrical Hahn-Banach Theorem () (we again use that X is locally convex and Hausdorff). The set $\{\psi(z) : z \in K\}$ is a compact subset of $\mathbb{R}$ (actually a closed bounded interval), so has a largest number μ. Then $F = \{z \in K : \psi(z) = \mu\}$ is a closed face in K disjoint from L. But F contains an extreme point of K by the previous proof, contradicting that all extreme points of K are in L. Thus $L = K$. 👍

XX.9.4.2. The local convexity hypothesis in the Krein-Milman Theorem is to insure that there are sufficiently many continuous linear functionals, or equivalently enough closed supporting hyperplanes. There exists a compact convex set in a topological vector space which is complete and metrizable, but not locally convex, which has no extreme points [Rob77]. (In a non-Hausdorff topological vector space, there is a trivial counterexample: the closure of $\{0\}$.)

XX.9.4.3. The hypotheses of the Krein-Milman Theorem are not necessary. If A is a unital C*-algebra (), then every element of the open unit ball of A is a convex combination of unitaries, which are extreme points of the closed unit ball [Bla06]. Thus the closed unit ball of a unital C*-algebra is the closed convex hull of its extreme points, even though there is no (Hausdorff) topological vector space topology on A in general under which the closed unit ball is compact. (The closed unit ball of a nonunital C*-algebra has no extreme points.) A closed bounded set in a Banach space with the Radon-Nikodym Property () is the closed convex hull of its extreme points ([Phe74], [Edg75]).

XX.9.5. Choquet's Theorem

There is an important sharpening of the Krein-Milman Theorem for metrizable compact convex sets, due to G. Choquet [Cho56]. Our exposition is adapted from [Phe66].

Representing Measures

XX.9.5.1. Definition. Let X be a topological vector space, and E a compact subset of X. Let μ be a Borel probability measure on E, and $x \in X$. Then μ *represents* x, or x is the *barycenter*, *resultant*, or *moment* of μ, if for every continuous linear functional ϕ on X we have

$$\phi(x) = \mu(\phi) := \int_E \phi \, d\mu$$

(note that ϕ is Borel measurable and bounded on E, so the integral is defined; the notation $\mu(\phi)$ comes from regarding μ as a bounded linear functional on $C(E)$ by the Riesz Representation Theorem ()).

The point x, if it exists, is the weak integral $\int_E \iota \, d\mu$ (), where ι is the identity map on E, regarded as a continuous function from E to X.

XX.9.5.2. Examples. (i) If $x \in X$ and E is any compact subset of X containing x, then the point mass δ_x (regarded as a Borel probability measure on E) represents x. The measure δ_x is called ϵ_x in some references.

(ii) Let K be the convex hull of n extreme points $\{p_1, \ldots, p_n\}$ in a Hausdorff topological vector space X. Then K is compact. If $x \in K$, then $x = \sum_{j=1}^n \alpha_j p_j$ for some $\alpha_1, \ldots, \alpha_n \geq 0$ with $\sum_{j=1}^n \alpha_j = 1$. If μ is the purely atomic Borel probability measure on K with $\mu(p_j) = \alpha_j$ (i.e. $\mu = \sum_{j=1}^n \alpha_j \delta_{p_j}$), then μ represents x (x is the barycenter of μ). Thus every point of K is the barycenter of a Borel probability measure on K supported on the extreme points of K. (Since the convex combination giving x is not necessarily unique, the representing measure for x is not unique in general.)

Choquet's Theorem extends (ii) to an arbitrary metrizable compact convex set K, giving a representing measure (not unique in general) for any point of K which is supported on the extreme points of K.

XX.9.5.3. Proposition. If X is locally convex (Hausdorff), and μ is a Borel probability measure on a compact subset E of X, then μ represents at most one point of X.

Proof: If μ represents x and y, then $\phi(x) = \mu(\phi) = \phi(y)$ for all $\phi \in X^*$. Thus $x = y$ since X^* separates the points of X. 👍

XX.9.5.4. Proposition. If K is a compact convex set in a topological vector space X, and μ is a Borel probability measure on K, of X, then μ represents at least one point of X (exactly one point x_μ if X is locally convex), and the barycenter is in K. If X is locally convex, the function $\mu \mapsto x_\mu$ is weak-* continuous, i.e. if $\mu_i \to \mu$ weak-* in $C(K)^*$, then $x_{\mu_i} \to x_\mu$ in K.

Proof: If $x \in X$ is represented by μ, then for any $\phi \in X^*$,

$$\phi(x) = \mu(\phi) \leq \sup\{\phi(y) : y \in K\}$$

and thus by the geometrical Hahn-Banach Theorem () $x \in K$ since it is closed and convex.

For existence, for each $\phi \in X^*$ let $H_\phi = \{x \in K : \phi(x) = \mu(\phi)\}$. Then H_ϕ is a closed hyperplane in X, and we wish to show that

$$\left[\bigcap_\phi H_\phi\right] \cap K \neq \emptyset \ .$$

Since these sets are compact, it suffices to show that for any $\phi_1, \ldots, \phi_n \in X^*$,

$$H_{\phi_1} \cap \cdots \cap H_{\phi_n} \cap K \neq \emptyset \ .$$

Define a bounded linear map $T : X \to \mathbb{R}^n$ by

$$T(x) = (\phi_1(x), \ldots, \phi_n(x)) \ .$$

Then $T(K)$ is a compact convex subset of $\mathbb{R}^n$. Set $\boldsymbol{y} = (\mu(\phi_1), \ldots, \mu(\phi_n)) \in \mathbb{R}^n$. It suffices to show $\boldsymbol{y} \in T(K)$. If $\boldsymbol{y} \notin T(K)$, there is a linear functional ψ on $\mathbb{R}^n$ with $\psi(\boldsymbol{y}) > \sup\{\psi(\boldsymbol{z}) : \boldsymbol{z} \in T(K)\}$. There is a vector $\boldsymbol{a} \in \mathbb{R}^n$ such that $\psi(\boldsymbol{z}) = \langle \boldsymbol{a}, \boldsymbol{z} \rangle$ for all $\boldsymbol{z} \in \mathbb{R}^n$ (). Set $\boldsymbol{a} = (a_1, \ldots, a_n)$. We have

$$\psi(\boldsymbol{y}) = \langle \boldsymbol{a}, \boldsymbol{y} \rangle > \sup\{\langle \boldsymbol{a}, T(x) \rangle : x \in K\} \ .$$

If $\theta \in X^*$ is defined by $\theta = \sum_{j=1}^n a_j \phi_j$, we have

$$\int_K \theta \, d\mu = \psi(\boldsymbol{y}) > \sup\{\theta(x) : x \in K\} \ .$$

This is impossible since μ is a probability measure. Thus $\boldsymbol{y} \in T(K)$.

Now suppose X is locally convex (Hausdorff), and (μ_i) is a net of Borel probability measures on K with $\mu_i \to \mu$ weak-*. Set $x_i = x_{\mu_i}$ and $x = x_\mu$. To show $x_i \to x$, we use the hard-to-state criterion XIII.3.2.13.. Fix a subnet (x_j) of (x_i). Then (x_j) has a convergent subnet by compactness of K. Fix a convergent subnet (x_r), say $x_r \to y \in K$. Then, for any $\phi \in X^*$, $\phi(x_r) \to \phi(y)$ by continuity of ϕ. On the other hand,

$$\phi(x_i) = \mu_i(\phi) \to \mu(\phi) = \phi(x)$$

and so the subnet $(\phi(x_r))$ also converges to $\phi(x)$. Thus $\phi(y) = \phi(x)$. This is true for all $\phi \in X^*$, so $y = x$ since X^* separates the points of X. So (x_j) has a subnet converging to x, and so $x_i \to x$ by XIII.3.2.13.. 🖒

XX.9.5.5. Corollary. If E is a compact set in a locally convex topological vector space X, and the closed convex hull K of E is also compact (this is automatic if X is topologically complete, e.g. if X is a Banach space in the norm topology, or the dual of a Banach space in the weak-* topology, cf. ()), then every Borel probability measure μ on E has a unique barycenter x_μ in X, and $x_\mu \in K$. The map $\mu \mapsto x_\mu$ is weak-* continuous.

XX.9.5.6. Proposition. If E is a compact set in a locally convex topological vector space X, and $x \in X$, then x is in the closed convex hull $\overline{co}(E)$ if and only if x is represented by a Borel probability measure on E.

Proof: One direction generalizes part of XX.9.5.5. (since it is not assumed that $\overline{co}(E)$ is compact), and has the same proof: if x is represented by μ, then for any $\phi \in X^*$,

$$\phi(x) = \mu(\phi) \leq \sup\{\phi(y) : y \in E\} = \sup\{\phi(y) : y \in \overline{co}(E)\}$$

and thus $x \in \overline{co}(E)$ since it is closed and convex.

For the converse, suppose $x \in \overline{co}(E)$. Then there is a net (x_i) in $co(E)$ with $x_i \to x$. Each x_i can be represented by a Borel probability measure μ_i on E which is a finite convex combination of point masses as in XX.9.5.2.(ii). Since the set of Borel probability measures on E is a weak-* compact subset of $C(E)^*$, there

is a subnet (μ_j) of the net (μ_i) which converges weak-* to a Borel probability measure μ. Then $x_j \to x$, and for any $\phi \in X^*$ we have

$$\phi(x_j) \to \phi(x)$$

$$\phi(x_j) = \mu_j(\phi) \to \mu(\phi)$$

and thus $\phi(x) = \mu(\phi)$. Since this holds for all $\phi \in X^*$, μ represents x. 👍

The Krein-Milman Theorem Revisited

The Krein-Milman Theorem can thus be rephrased using XX.9.5.6.:

XX.9.5.7. Theorem. [Krein-Milman Theorem, Representing Measure Form] Let K be a compact convex set in a locally compact (Hausdorff) topological vector space X, and E the closure of the set $\partial_e(K)$ of extreme points of K. Then every point of K is the barycenter of a Borel probability measure on E.

For the proof, just note that $\overline{co}(E) = \overline{co}(\partial_e(K))$, and apply XX.9.4.1. and XX.9.5.6.. Conversely, XX.9.4.1. follows from XX.9.5.7. and XX.9.5.6., so XX.9.4.1. and XX.9.5.7. are equivalent via XX.9.5.6..

XX.9.5.8. This form of the Krein-Milman Theorem is not optimally useful, since E can be a rather large subset of K; in fact, E is "generically" the whole set K (i.e. the set of extreme points of K is generically dense in K)! [Kle59].

It would be far better to have representing measures supported on the set $\partial_e(K)$ of extreme points of K itself, not just on the closure.

XX.9.5.9. In order for the desired statement to even make sense, we need the set $\partial_e(K)$ to be a Borel set. Although it is not closed in general, even if K is a metrizable Choquet simplex (), in the metrizable case we do have:

XX.9.5.10. Proposition. Let K be a metrizable compact convex set in a topological vector space X. Then the set $\partial_e(K)$ of extreme points of K is a G_δ, and in particular a Borel set.

Proof: Fix a metric ρ giving the topology of K. For $n \in \mathbb{N}$, let A_n be the set of all $x \in K$ such that there exist $y, z \in K$ with $x = \frac{1}{2}(y+z)$ and $\rho(y,z) \geq \frac{1}{n}$. Then $\cup_{n=1}^{\infty} A_n$ is the complement of $\partial_e(K)$, so it suffices to show that each A_n is a closed set.

Fix $n \in \mathbb{N}$, and let (x_j) be a sequence in A_n with $x_j \to x \in K$. For each j, there are $y_j, z_j \in K$ with $x_j = \frac{1}{2}(y_j + z_j)$ and $\rho(y_j, z_j) \geq \frac{1}{n}$. By compactness of K, there are convergent subsequences (y_{r_j}) and (z_{r_j}) of (y_j) and (z_j), which can be taken to have the same indices (), say $y_{r_j} \to y$, $z_{r_j} \to z$. Then

$$\frac{1}{2}(y_{r_j} + z_{r_j}) \to \frac{1}{2}(y+z) \ .$$

But

$$\frac{1}{2}(y_{r_j} + z_{r_j}) = x_{r_j} \to x$$

and thus $x = \frac{1}{2}(y+z)$ since K is Hausdorff. Since $\rho(y,z) \geq \frac{1}{n}$ by continuity of ρ, we have $x \in A_n$ and A_n is closed. 👍

We then have the following strengthening of XX.9.5.7. in the metrizable case:

XX.9.5.11. THEOREM. [CHOQUET'S THEOREM] Let K be a metrizable compact convex set in a locally convex (Hausdorff) topological vector space X. Then every $x \in K$ is represented by a measure which is supported on the set $\partial_e(K)$ of extreme points of K.

XX.9.5.12. We give a nice proof due to BONSALL [Bon63] (cf. [Phe66]) based on the Hahn-Banach Theorem. The argument will be a series of lemmas. To begin, fix a point $x_0 \in K$; we will show that x_0 is represented by a measure supported on $\partial_e(K)$. We will work with the subspace $\mathrm{Aff}(K)$ of $C(K)$ of real-valued continuous affine functions on K. Since K is metrizable, $\mathrm{Aff}(K)$ is separable; fix a sequence (h_n) which is dense in the unit sphere of $\mathrm{Aff}(K)$, and set

$$f = \sum_{n=1}^{\infty} 2^{-n} h_n^2 \in C(K) \ .$$

XX.9.5.13. LEMMA. The function f is strictly convex on K, i.e. $f(tx+(1-t)y) < tf(x)+(1-t)f(y)$ for $x, y \in K$, $x \neq y$, $0 < t < 1$.

PROOF: Each h_n^2 is clearly convex, so f is convex. If $x \neq y$, there is an h_n with $h_n(x) \neq h_n(y)$, so h_n is affine and not constant on the segment between x and y, and therefore h_n^2 is strictly convex on this segment. Hence f is too. 👍

XX.9.5.14. DEFINITION. If $g \in C(K)$, write

$$\bar{g} = \inf\{h : h \in \mathrm{Aff}(K), g \leq h\} \ .$$

The function $\bar{g}$ is called the *upper envelope* of g.

XX.9.5.15. LEMMA. Let $g, g_1, g_2 \in C(K)$. Then

(i) $\bar{g}$ is concave, i.e. $\bar{g}(tx+(1-t)y) \geq t\bar{g}(x)+(1-t)\bar{g}(y)$ for $x, y \in K$, $0 < t < 1$.

(ii) $\bar{g}$ is bounded and upper semicontinuous, hence Borel measurable.

(iii) $g \leq \bar{g}$, and $g = \bar{g}$ if and only if g is concave.

(iv) $\overline{g_1+g_2} \leq \bar{g}_1 + \bar{g}_2$, and $\overline{g_1+g_2} = \bar{g}_1 + \bar{g}_2 = \bar{g}_1 + g_2$ if $g_2 \in \mathrm{Aff}(K)$.

(v) $|\bar{g}_1 - \bar{g}_2| \leq \|g_1 - g_2\|$.

(vi) If $\alpha \geq 0$, then $\overline{\alpha g} = \alpha \bar{g}$.

The simple proof is straightforward and left to the reader.

XX.9.5.16. Let Y be the subspace of $C(K)$ spanned by $\mathrm{Aff}(K)$ and f (defined above). Define a linear functional ψ on Y by $\psi(h+\alpha f) = h(x_0) + \alpha \bar{f}(x_0)$.

XX.9.5.17. Lemma. For $g \in C(K)$, define $p(g) = \bar{g}(x_0)$. Then

(i) p is subadditive and positive-homogeneous.

(ii) If $h + \alpha f \in Y$, then $\psi(h+\alpha f) \leq p(h+\alpha f)$.

Proof: (i) follows immediately from XX.9.5.15.(iv) and (vi).
(ii): If $\alpha \geq 0$, then $\overline{h+\alpha f} = h + \alpha \bar{f}$ by XX.9.5.15.(iv) and (vi), and in particular

$$\psi(h+\alpha f) = h(x_0) + \alpha \bar{f}(x_0) = \overline{h+\alpha f}(x_0) = p(h+\alpha f)$$

and if $\alpha < 0$, then $h + \alpha f$ is concave, so $\overline{h+\alpha f} = h + \alpha f \geq h + \alpha \bar{f}$ and in particular

$$\psi(h+\alpha f) = h(x_0) + \alpha \bar{f}(x_0) \leq h(x_0) + \alpha f(x_0) = \overline{h+\alpha f}(x_0) = p(h+\alpha f)\ .$$



XX.9.5.18. Thus by the Hahn-Banach Theorem there is a linear functional ϕ on $C(K)$ with

$$\phi(g) \leq p(g) = \bar{g}(x_0)$$

for all $g \in C(K)$, which extends ψ.

XX.9.5.19. Lemma. The linear functional ϕ is a state (positive linear functional of norm 1) on $C(K)$.

Proof: If $g \in C(K)$, $g \leq 0$, then since the constant function 0 is in $\mathrm{Aff}(K)$, $\bar{g} \leq 0$, and in particular $\phi(g) \leq p(g) = \bar{g}(x_0) \leq 0$. Taking negatives, ϕ is a positive linear functional. Since the constant function 1 is in $\mathrm{Aff}(K)$, we have $\phi(1) = \psi(1) = 1(x_0) = 1$.



XX.9.5.20. Thus, by the Riesz Representation Theorem, there is a Borel probability measure μ on K with

$$\phi(g) = \mu(g) := \int_K g\, d\mu$$

for all $g \in C(K)$. We have $\mu(f) = \bar{f}(x_0)$. Since $f \leq \bar{f}$, $\mu(f) \leq \mu(\bar{f})$ (note that $\mu(\bar{f}) = \int_K \bar{f}\, d\mu$ is defined even though $\bar{f}$ may not be continuous, since it is bounded and Borel measurable); but if $h \in \mathrm{Aff}(K)$, $h \geq f$, we have $h \geq \bar{f}$, so

$$h(x_0) = \psi(h) = \phi(h) = \mu(h) \geq \mu(\bar{f})\ .$$

Taking the infimum over all such h, we get

$$\bar{f}(x_0) \geq \mu(\bar{f}) \geq \mu(f) = \bar{f}(x_0)$$

and thus equality holds, i.e.

$$\int_K \bar{f}\, d\mu = \mu(\bar{f}) = \mu(f) = \int_K f\, d\mu$$

and so

$$\int_K (\bar{f} - f)\, d\mu = 0 \ .$$

Thus μ vanishes on the complement of the set

$$S = \{x \in K : f(x) = \bar{f}(x)\} \ .$$

XX.9.5.21. LEMMA. The set S is contained in $\partial_e(K)$.

PROOF: If x, y, z are distinct points of K with $x = \frac{1}{2}(y + z)$, then by the strict convexity of f we have

$$f(x) < \frac{1}{2}f(y) + \frac{1}{2}f(z) \leq \frac{1}{2}\bar{f}(y) + \frac{1}{2}\bar{f}(z) \leq \bar{f}(x)$$

since $\bar{f}$ is concave. 

This completes the proof of XX.9.5.11.. 

XX.9.5.22. In fact, the set S is exactly $\partial_e(K)$; cf. [Phe66, 3.1].

The Nonmetrizable Case

XX.9.5.23. There is a generalization of Choquet's Theorem to nonmetrizable compact convex sets, the Bishop-de Leeuw Theorem [BdL59]. This theorem is considerably more difficult to prove, and is even somewhat tricky to state, because (among other things) the set of extreme points in a nonmetrizable compact convex set, even a nonmetrizable Choquet simplex, need not be a Borel set. See e.g. [Phe66] or [Alf71] for details.

XX.9.5.24. The Bishop of Bishop-de Leeuw is the Errett Bishop of constructivism fame (I.7.1.20.). Karel de Leeuw is known (besides for his work) for a darker reason: he was murdered in his office by a graduate student.

XX.9.6. Regular Embeddings, Cones, and Affine Functions

XX.9.7. Choquet Simplexes

XX.9.7.1. A compact convex set can be thought of as being defined by "generators" and "relations." The "generators" are the set of extreme points, and the "relations" are pairs of measures which represent the same point. For example, a square (or parallelogram, cf. XX.9.8.1.(b)) has four generators v_1, v_2, v_3, v_4 and one relation

$$\frac{1}{2}(v_1 + v_3) = \frac{1}{2}(v_2 + v_4) \ .$$

From this (somewhat imprecise) point of view, a Choquet simplex is a "free" compact convex set, i.e. one with no relations, although this should not be taken too literally unless the set of extreme points is closed (cf. XX.9.8.6.).

Bauer Simplexes

XX.9.7.2. Definition. A Choquet simplex Δ in which the set of extreme points is closed in Δ is a *Bauer simplex*.

XX.9.7.3. Since a Bauer simplex is compact, its set of extreme points is a compact Hausdorff space. Every compact Hausdorff space X is the set of extreme points of a Bauer simplex Δ_X, which is unique up to isomorphism. Δ_X can be concretely described as the state space of $C(X)$ (), which by the Riesz Representation Theorem can be identified with the set of Radon probability measures on X. The extreme points are just the point masses, which can be identified with X (). The representing measure of a point is just the point itself; thus uniqueness is essentially tautological, so Δ_X is a Choquet simplex. Δ_X is metrizable if and only if X is metrizable, i.e. $C(X)$ is separable (). All finite-dimensional simplexes are, of course, Bauer simplexes.

Other Simplexes

XX.9.7.4. "Most" Choquet simplexes are not Bauer simplexes. The simplest example of a Choquet simplex which is not a Bauer simplex is given in Exercise XX.9.8.2.. The construction can be modified to give far more complicated examples.

In fact, if X is any Polish space, there is a metrizable Choquet simplex Δ such that $\partial_e(\Delta)$ is homeomorphic to X (); the set of extreme points of a metrizable Choquet simplex is always a G_δ, hence a Polish space

in the relative topology). The simplex Δ is highly nonunique in general; for example, the set of extreme points for each of the Δ_c ($0 < c \leq 1$) of XX.9.8.2.(f) is a countable discrete space (and there are much more complicated simplexes with the same space of extreme points).

XX.9.7.5. The most extreme case is the Poulsen simplex. There is a metrizable Choquet simplex Δ, unique up to isomorphism, called the *Poulsen simplex*, for which the set of extreme points of Δ is dense in Δ [Pou61]. Every metrizable Choquet simplex is isomorphic to a face of the Poulsen simplex.

The set of extreme points of the Poulsen simplex is thus a Polish space in which every Polish space can be embedded as a closed subset. It can be shown that $\partial_e(\Delta)$ is homeomorphic to $\mathbb{R}^\infty$ (), the underlying topological space of every separable infinite-dimensional Banach space ().

The Facial Topology

XX.9.7.6. Let Δ be a Choquet simplex. The collection

$$\{\emptyset\} \cup \{F \cap \partial_e(\Delta) : F \text{ a closed face of } \Delta\}$$

of subsets of $\partial_e(\Delta)$ contains $\emptyset$ and $\partial_e(\Delta)$ and is closed under finite unions and arbitrary intersections (this is nontrivial to prove and uses the notion of a *split face*; every closed face in a Choquet simplex is split, cf. [Alf71]), and is thus the collection of closed sets for a topology on $\partial_e(\Delta)$ called the *facial topology*. Each of these sets is closed in the relative topology from Δ, and thus the facial topology is weaker than the relative topology. Since extreme points of Δ are singleton closed faces, the facial topology is T_1. But it is not necessarily Hausdorff. See XX.9.8.2.–XX.9.8.4..

The facial topology was defined and first studied in [Eff67]. A facial topology can be defined more generally on the set of extreme points of any compact convex set, using the notion of split face, although it will often be the indiscrete topology on sets which are not simplexes.

XX.9.8. Exercises

XX.9.8.1. (a) Show that all (solid) triangles are isomorphic as compact convex sets.

(b) Show that all parallelograms are isomorphic as convex sets. Show that a convex quadrilateral which is not a parallelogram is not isomorphic to a parallelogram. Describe the isomorphism classes of convex quadrilaterals. [Consider how the diagonals intersect.]

XX.9.8.2. Give ℓ^1 the weak-* topology as c_0^* (). Then the closed unit ball is compact (), metrizable (), and convex. Let $(\boldsymbol{e}_n)$ be the standard Schauder basis. Let Δ be the set of sequences $\boldsymbol{x} = (\lambda_1, \lambda_2, \dots)$ such that $\|\boldsymbol{x}\| \leq 1$, $-1 \leq \lambda_1 \leq 1$, and $\lambda_n \geq 0$ for all $n > 1$.

(a) Show that Δ is the closed convex hull of $E = \{\boldsymbol{e}_1, -\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \dots\}$, and that E is the set of extreme points of Δ. Thus Δ is compact (and metrizable).

(b) Show that Δ is a Choquet simplex. [Use ().]

(c) Show that E is not closed in Δ: $\bar{E} = E \cup \{\mathbf{0}\}$. Note that $\mathbf{0}$ is the midpoint of the segment between $\boldsymbol{e}_1$ and $-\boldsymbol{e}_1$.

(d) Let F be the set of sequences in Δ for which $\lambda_1 = 0$ and $\sum_{k=2}^\infty \lambda_k = 1$. Show that F is a face of Δ.

(e) Show that $\bar{F}$ is the set of sequences in Δ with $\lambda_1 = 0$, which is not a face of Δ.

(f) More generally, fix c, $0 \leq c \leq 1$, and let Δ_c be the closed convex hull of $\{\boldsymbol{e}_1, -c\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3, \dots\}$. Then each Δ_c is a metrizable Choquet simplex with a sequence of extreme points converging to $\mathbf{0}$. No two of the Δ_c are isomorphic. Δ_0 is a Bauer simplex ($\mathbf{0}$ is an extreme point).

(g) In all these examples except Δ_0, the relative topology on $\partial_e(\Delta_c)$ is the discrete topology. The facial topology on $\partial_e(\Delta_c)$ is the T_1 topology consisting of a sequence of isolated points ($\boldsymbol{e}_n$ for $n \geq 2$) converging simultaneously to $\boldsymbol{e}_1$ and $-c\boldsymbol{e}_1$. The relative and facial topologies on $\partial_e(\Delta_0)$ coincide and consist of a sequence of isolated points (the $\boldsymbol{e}_n$) converging to $\mathbf{0}$.

XX.9.8.3. Let Δ_0 be the simplex of XX.9.8.2.(f), and let π be the quotient map from ℓ^1 onto ℓ^1/X, where X is the one-dimensional subspace of ℓ^1 spanned by $\boldsymbol{x} = (2^{-1}, 2^{-2}, 2^{-3}, \dots)$. Set $\Delta = \pi(\Delta_0)$.

(a) Show that Δ is a Choquet simplex, and that $\partial_e(\Delta)$ consists of $\{\boldsymbol{f}_n : n \in \mathbb{N}\}$, where $\boldsymbol{f}_n = \pi(\boldsymbol{e}_n)$.

(b) The relative topology on $\partial_e(\Delta)$ is the discrete topology, while the facial topology is the finite complement topology.

This Δ is a *prime simplex* (cf. XX.9.8.4.(b)).

XX.9.8.4. Let Δ be a Choquet simplex.

(a) Prove that the following are equivalent:

(i) The facial topology on $\partial_e(\Delta)$ is the same as the relative topology from Δ.

(ii) The facial topology on $\partial_e(\Delta)$ is Hausdorff.

(iii) $\mathrm{Aff}(\Delta)$ is a lattice.

(iv) Every bounded uniformly continuous function from $\partial_e(\Delta)$ to $\mathbb{R}$ can be extended to an element of $\mathrm{Aff}(\Delta)$.

(iv′) Every continuous function from $\partial_e(\Delta)$ to $\mathbb{R}$ can be extended to an element of $\mathrm{Aff}(\Delta)$.

(v) Whenever F is a face of Δ, then $\bar{F}$ is also a face.

(vi) Δ is a Bauer simplex.

(b) Prove that the following are equivalent:

(vii) Whenever E, F are closed faces of Δ with $co(E \cup F) = \Delta$, then $E = \Delta$ or $F = \Delta$.

(viii) Any two nonempty facially open subsets of $\partial_e(\Delta)$ have nonempty intersection.

(ix) $\mathrm{Aff}(\Delta)$ is an antilattice, i.e. if $f, g \in \mathrm{Aff}(\Delta)$, then $\{f, g\}$ has an infimum if and only if $f \leq g$ or $g \leq f$.

A Choquet simplex satisfying conditions (vii)-(ix) is called a *prime simplex*. Prime simplexes are the "opposite extreme" to Bauer simplexes. See [Alf71].

XX.9.8.5. ([Bau61]; cf. [Phe66, 1.4], [Alf71, I.2.4]) Let K be a nonempty compact convex set in a locally convex topological vector space.

(a) If $x \in K$, show that x is an extreme point of K if and only if δ_x is the only Borel probability measure on K representing x.

(b) Use (a) and XX.9.5.6. to give an alternate proof of Milman's partial converse XX.9.2.5..

XX.9.8.6. If K is a compact convex set in a locally convex topological vector space, say a Choquet simplex Δ *covers* K if there is a continuous affine map $\phi : \Delta \to K$ mapping $\partial_e(\Delta)$ homeomorphically to $\partial_e(K)$. The map ϕ is a *covering*.

(a) Show that if $\partial_e(K)$ is closed, there is a covering of K by the Bauer simplex $\Delta_{\partial_e(K)}$, and this covering of K is the unique covering of K by a Choquet simplex up to isomorphism over K (in the obvious sense).

(b) If $\partial_e(K)$ is not closed, is there always a covering of K by a Choquet simplex?

(c) If $\partial_e(K)$ is not closed, a covering of K by a Choquet simplex need not be unique. Let K be a compact convex set with a sequence (e_n) of extreme points, such that the convex hull of $\{e_1, e_2, e_3, e_4\}$ is a square with these points as corners in cyclic order, and the sequence (e_n) converges to the midpoint of the square (such a K exists in $\mathbb{R}^4$, but probably not in $\mathbb{R}^3$). The Choquet simplex Δ_{13} constructed as in XX.9.8.2.(a) with extreme points converging to $a = \frac{1}{2}(e_1 + e_3)$ covers K, as does the simplex Δ_{24} with extreme points converging to $b = \frac{1}{2}(e_2 + e_4)$, and though Δ_{13} and Δ_{24} are isomorphic, they are not isomorphic over K. There are many other Choquet simplexes covering K: if T is the tetrahedron (3-simplex) with extreme points $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_4\}$ and Y is a nonempty closed subset of the line segment between $\boldsymbol{a} = \frac{1}{2}(\boldsymbol{e}_1 + \boldsymbol{e}_3)$ and $\boldsymbol{b} = \frac{1}{2}(\boldsymbol{e}_2 + \boldsymbol{e}_4)$ in T, there is a Choquet simplex covering K with a sequence of extreme points whose set of limit points is Y. There are $2^{\aleph_0}$ isomorphism classes of such simplexes.

XX.9.8.7. Let K be a convex set in a topological vector space X. A point p of K is an *exposed point* of K if there is a closed hyperplane H such that $H \cap K = \{p\}$, i.e. if there is a continuous linear functional ϕ on X with $\phi(x) < \phi(p)$ for all $x \in K \setminus \{p\}$.

(a) Show that every exposed point is an extreme point.

(b) [Alf71, II.5.21] Show that every extreme point of a metrizable Choquet simplex is an exposed point.

(c) Let K be the union in $\mathbb{R}^2$ of the closed unit disk and the square

$$\{(x, y) : -1 \leq x \leq 1, -2 \leq y \leq 0\} .$$

Then K is compact and convex. Show that $(1, 0)$ and $(-1, 0)$ are extreme points of K but not exposed points. Find all extreme points and exposed points of K.

There are other related notions: strongly exposed points, dentable points, etc. These notions are tied in with the structure theory of Banach spaces.

XX.9.8.8. Let K be a compact convex set in a locally convex topological vector space.

(a) Prove the following theorem (cf. [Alf71,], [Goo86a,]):

THEOREM. The following are equivalent:

(i) K is a Choquet simplex.

(ii) $(\mathrm{Aff}(K)^*, \leq)$ is a lattice.

(iii) $(\mathrm{Aff}(K), \leq)$ has the Riesz Interpolation Property ().

(iv) $(\mathrm{Aff}(K), \ll)$ has the Riesz Interpolation Property.

(v) Whenever $x_1, \dots, x_n, y_1, \dots, y_m \in K$, $\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_m \in (0,1)$, $\sum_{j=1}^n \alpha_j = \sum_{k=1}^m \beta_k = 1$, and

$$\sum_{j=1}^n \alpha_j x_j = \sum_{k=1}^m \beta_k y_k$$

there are $z_{jk} \in K$ for $1 \le j \le n$, $1 \le k \le m$ and $\gamma_{jk} \in [0,1]$ such that $\sum_{k=1}^m \gamma_{jk} = \alpha_j$, $\sum_{j=1}^n \gamma_{ij} = \beta_k$, and

$$x_j = \sum_{k=1}^m \alpha_j^{-1} \gamma_{jk} z_{jk} \text{ and } y_k = \sum_{j=1}^n \beta_k^{-1} \gamma_{jk} z_{jk}$$

for all j, k (the (z_{jk}) and (γ_{jk}) are called a *Riesz decomposition* of $\sum_j \alpha_j x_j$ and $\sum_k \beta_k y_k$.) We then have

$$\sum_{j=1}^n \alpha_j x_j = \sum_{k=1}^m \beta_k y_k = \sum_{j=1}^n \sum_{k=1}^m \gamma_{jk} z_{jk} \ .$$

(b) Prove the following theorem (cf. [Alf71,], [Goo86a,]):

THEOREM. The following are equivalent:

(i) K is a Bauer simplex.

(ii) $(\mathrm{Aff}(K), \le)$ is a lattice.

(iii) $(\mathrm{Aff}(K), \le, 1)$ is isometrically linearly isomorphic to $(C(X), \le, 1)$ for some compact Hausdorff space X.

(iv) If $X = \partial_e K$, then X is closed in K and the restriction from K to X is an isometric linear isomorphism from $\mathrm{Aff}(K)$ onto $C(X)$.

In (iii) and (iv), K is isomorphic to Δ_X.

XX.9.8.9. Let $(K_i, \phi_{j,i})$ be an inverse system where each K_i is a compact convex set in a locally convex topological vector space and the connecting maps $\phi_{j,i}$ are continuous and affine (not necessarily surjective). Let K be the inverse limit, and $\phi_i : K \to K_i$ the induced map for each i.

(a) Show that K has a natural structure as a compact convex set in a locally convex topological vector space, making the ϕ_i continuous and affine. [If K_i is in a LCS space V_i, construct K as a subset of $\prod_i V_i$ as in ().]

(b) For each i let ψ_i be the induced map from $\mathrm{Aff}(K_i)$ to $\mathrm{Aff}(K)$ given by $\psi_i(f) = f \circ \phi_i$. Show that $\cup_i \psi_i(\mathrm{Aff}(K_i))$ is uniformly dense in $\mathrm{Aff}(K)$. [Use XIII.14.5.19..]

(c) Use the equivalence of (i) and (iv) in XX.9.8.8.(a) to conclude that if each K_i is a Choquet simplex, then K is also a Choquet simplex. Use the equivalence of (i) and (v) in XX.9.8.8.(a) to give an alternate proof.

(d) As a special case, if (K_n) is a decreasing sequence of Choquet simplexes, then $\cap_n K_n$ is also a Choquet simplex.

(e) [Goo86a,] Conversely, show that every Choquet simplex is an inverse limit of a system of finite-dimensional simplexes with surjective affine connecting maps. (In particular, an inverse limit of Bauer simplexes is not necessarily a Bauer simplex.)

XX.10. The Stone-Weierstrass Theorem

The Stone-Weierstrass Theorem is one of the important theorems of functional analysis, giving a useful criterion for uniformly approximating general continuous functions by "nice" functions.

XX.10.1. The Weierstrass Approximation Theorem

The original version proved by Weierstrass in 1885 was:

XX.10.1.1. Theorem. Let I be a closed bounded interval in $\mathbb{R}$. Then every continuous function from I to $\mathbb{R}$ can be uniformly approximated arbitrarily closely on I by a polynomial.

Bernstein's Proof

In 1912 Sergei Bernstein gave a better proof of the Weierstrass Approximation Theorem in which he gave an explicit sequence of polynomials converging uniformly to a given continuous function on an interval:

XX.10.1.2. Theorem. If f is a continuous real-valued function on $[0,1]$, define, for $0 \leq x \leq 1$,

$$B_n(f)(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}$$

(This is the n'th *Bernstein polynomial* of f.) Then $(B_n(f))$ converges uniformly to f on $[0,1]$.

Other intervals can be handled via an affine transformation.

XX.10.1.3. Bernstein, a probabilist, was apparently motivated by the following probabilistic argument (cf. [**?**]), which suggests that $B_n(f)$ should at least converge to f pointwise.

Fix f and $x \in [0,1]$. We repeatedly toss a coin with probability x of heads. Suppose we toss the coin a large number n of times, and receive a payoff of $f(r)$ dollars, where r is the fraction of tosses which are heads. In the long run (as $n \to \infty$), r will approach x with probability 1, so the payoff will approach $f(x)$ since f is continuous.

If we toss the coin n times, we will get exactly k heads with probability

$$\binom{n}{k} x^k (1-x)^{n-k}$$

so the expected payoff is

$$\sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k} = B_n(f)(x) \ .$$

A direct proof of Bernstein's result can be found in [**?**] or almost any book on numerical analysis or approximation theory; a probability version of the proof can be found in [**?**]. In fact, books on numerical analysis prove the following sharpened form:

XX.10.1.4. THEOREM. Let f be a bounded real-valued function on $[0,1]$, and $B_n(f)$ the n'th Bernstein polynomial of f (XX.10.1.2.). Then

$$|f(x) - B_n(f)(x)| \leq \frac{3}{2}\,\omega\left(f, \frac{1}{\sqrt{n}}\right)$$

for all $x \in [0,1]$, where $\omega(f,\cdot)$ is the modulus of continuity () of f on $[0,1]$.

Since $\lim_{\delta\to 0}\omega(f,\delta) = 0$ if f is continuous, XX.10.1.2. is a corollary of XX.10.1.4..

Korovkin's Theorem

We give a proof of XX.10.1.2., and hence XX.10.1.1., based on a result of independent interest due to KOROVKIN [Kor53]:

XX.10.1.5. THEOREM. Let $I = [a,b]$ be a closed bounded interval, and (L_n) a sequence of positive linear operators from $C_{\mathbb{R}}(I)$ to $C_{\mathbb{R}}(I)$ (i.e. if $f \geq 0$ then $L_n f \geq 0$ for all n). Let $p_0(x) = 1$, $p_1(x) = x$, and $p_2(x) = x^2$. If $L_n p_k \to p_k$ uniformly for $k = 0,1,2$, then $L_n f \to f$ uniformly for all $f \in C_{\mathbb{R}}(I)$.

PROOF: (cf. [Bri98]) For each $t \in I$, set $g_t(x) = (x-t)^2 = t^2 p_0(x) - 2tp_1(x) + p_2(x)$. Then

$$[L_n g_t](x) = t^2[L_n p_0](x) - 2t[L_n p_1](x) + [L_n p_2](x)$$

for all n, t, x. Since $g_t \geq 0$, we have

$$0 \leq [L_n g_t](t) = t^2([L_n p_0](t) - 1) + 2t([L_n p_1](t) - t) + ([L_n p_2](t) - t^2)$$

$$\leq t^2\|L_n p_0 - p_0\| + |2t|\|L_n p_1 - p_1\| + \|L_n p_2 - p_2\|$$

and thus, since t^2 and $|2t|$ are bounded on I, for any $\eta > 0$ there is an N such that $\|L_n g_t\| < \eta$ for all $n \geq N$ and all $t \in I$.

Now fix $f \in C_{\mathbb{R}}(I)$ and $\epsilon > 0$. Since f is uniformly continuous on I, there is a $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $x, y \in I$ and $|x - y| < \delta$. Fix $t \in I$. If $x \in I$, $|x - t| \geq \delta$, we have

$$|f(t) - f(x)| \leq 2\|f\| \leq 2\|f\|\frac{(x-t)^2}{\delta^2} = \frac{2}{\delta^2}\|f\| g_t(x)$$

so for any $x \in I$ we have

$$|f(t) - f(x)| \leq \frac{2}{\delta^2}\|f\| g_t(x) + \epsilon\ .$$

Thus

$$-\epsilon p_0 - \frac{2}{\delta^2}\|f\| g_t \leq f(t)p_0 - f \leq \epsilon p_0 + \frac{2}{\delta^2}\|f\| g_t$$

and hence

$$-\epsilon L_n p_0 - \frac{2}{\delta^2}\|f\| L_n g_t \leq f(t) L_n p_0 - L_n f \leq \epsilon L_n p_0 + \frac{2}{\delta^2}\|f\| L_n g_t$$

$$|f(t)[L_n p_0](t) - [L_n f](t)| \leq \epsilon\|L_n p_0\| + \frac{2}{\delta^2}\|f\|[L_n g_t](t)\ .$$

Thus

$$|f(t) - [L_n f](t)| \le |f(t) - f(t)[L_n p_0](t)| + |f(t)[L_n p_0](t) - [L_n f](t)|$$
$$\le |f(t)||1 - [L_n p_0](t)| + \epsilon \|L_n p_0\| + \frac{2}{\delta^2}\|f\|[L_n g_t](t)$$
$$\le \|f\|\|p_0 - L_n p_0\| + \epsilon(\|p_0\| + \|p_0 - L_n p_0\|) + \frac{2}{\delta^2}\|f\|[L_n g_t](t) \ .$$

so if $\eta > 0$ and n is large enough,

$$|f(t) - [L_n f](t)| \le \|f\|\|p_0 - L_n p_0\| + \epsilon(\|p_0\| + \|p_0 - L_n p_0\|) + \frac{2\eta}{\delta^2}\|f\| \ .$$

The right side is independent of t and can be made less than 2ϵ by choosing η small enough and then n large enough. Since $\epsilon > 0$ is arbitrary, we have $L_n f \to f$ uniformly on I. 👍

XX.10.1.6. To prove XX.10.1.2., and hence XX.10.1.1., it thus suffices to show that each B_n is a positive operator (obvious) and that $B_n p_k \to p_k$ uniformly for $k = 0, 1, 2$, which is a simple calculation.

XX.10.2. The Stone-Weierstrass Theorem

In 1937 Marshall Stone formulated and proved a general version of Weierstrass's Theorem, which is now known as the Stone-Weierstrass Theorem. There are several commonly stated versions of the theorem, all of which are simple consequences of the following general version. Stone also proved a lattice version (XX.10.2.14.).

XX.10.2.1. Theorem. [Stone-Weierstrass] Let X be a compact Hausdorff space, and $C_{\mathbb{R}}(X)$ the algebra of real-valued continuous functions on X. Let A be a subalgebra of $C_{\mathbb{R}}(X)$ which is uniformly closed () and separates points (i.e. if $x, y \in X$ with $x \neq y$ there is an $f \in A$ with $f(x) \neq f(y)$). Then either $A = C_{\mathbb{R}}(X)$ or there is a $p \in X$ such that A consists of all $f \in C_{\mathbb{R}}(X)$ with $f(p) = 0$.

Here is the most commonly stated version of the Stone-Weierstrass Theorem (to prove it, consider the uniform closure of A, which is also an algebra):

XX.10.2.2. Corollary. Let X be a compact Hausdorff space, and A a subalgebra of $C_{\mathbb{R}}(X)$ which separates points and contains the constant functions. Then A is dense in $C_{\mathbb{R}}(X)$ for the uniform norm.

Weierstrass's Theorem is an immediate corollary of this, applied to $X = I$ and A the algebra of polynomials.

There is also a complex version:

XX.10.2.3. Theorem. [Complex Stone-Weierstrass] Let X be a compact Hausdorff space, and $C(X)$ the algebra of complex-valued continuous functions on X. Let A be a (complex) subalgebra of $C(X)$ which is closed under complex conjugation, uniformly closed, and separates points. Then either $A = C(X)$ or there is a $p \in X$ such that A consists of all $f \in C(X)$ with $f(p) = 0$.

A complex version of XX.10.2.2. follows immediately:

XX.10.2.4. Corollary. Let X be a compact Hausdorff space, and A a subalgebra of $C(X)$ which separates points, contains the constant functions, and is closed under complex conjugation. Then A is dense in $C(X)$ for the uniform norm.

To prove XX.10.2.3. from XX.10.2.1., note that if A is closed under complex conjugation and $f \in A$, then the real and imaginary parts $Re(f) = (f + \bar{f})/2$ and $Im(f) = (f - \bar{f})/2i$ are also in A (the converse is also true). The real-valued functions in A thus satisfy the hypotheses of XX.10.2.1..

XX.10.2.5. The hypothesis that A be closed under complex conjugation, or something similar, is necessary; see Exercise XX.10.3.10..

XX.10.2.6. Here is an important observation for the proof of XX.10.2.1.. If A is an algebra of real-valued functions on a space X, $f \in A$, and p is a polynomial with real coefficients and no constant term, then $p \circ f \in A$: if $p(t) = a_1 t + a_2 t^2 + \cdots + a_n t^n$, then $p \circ f = a_1 f + a_2 f^2 + \cdots + a_n f^n \in A$. If A contains the constant functions, the same is true if p is a polynomial with constant term. If (p_n) is a sequence of polynomials (without constant term) which converge uniformly on the range of f to a function g, then $p_n \circ f \to g \circ f$ uniformly on X, so if A is uniformly closed, $g \circ f \in A$. Here is an important application:

XX.10.2.7. Lemma. Let A be a uniformly closed algebra of real-valued bounded functions on a set X. If $f \in A$, then $|f| \in A$.

Since $|f| = g \circ f$, where $g(t) = |t|$, this follows from the next lemma:

XX.10.2.8. Lemma. For any $m \in \mathbb{N}$ there is a sequence of polynomials without constant term converging uniformly to $g(t) = |t|$ on $[-m, m]$.

Proof: This can be quickly deduced from XX.10.1.1., but we give a simple direct argument. First consider the interval $[-1, 1]$. By V.12.5.9., the Taylor polynomials $q_n(t)$ around 0 converge uniformly to $h(t) = (1 + t)^{1/2}$ on $[-1, 1]$. Set $r_n(t) = q_n(t^2 - 1)$; then r_n is a polynomial, and $r_n(t)$ converges uniformly to $h(t^2 - 1) = g(t)$ on $[-1, 1]$ (since $t^2 - 1$ maps $[-1, 1]$ into $[-1, 1]$). It is unclear whether r_n has a nonzero constant term; but if c_n is the constant term of r_n, then $c_n = r_n(0) \to g(0) = 0$, so if $p_n(t) = r_n(t) - c_n$, then p_n is a polynomial without constant term, and $p_n \to g$ uniformly on $[-1, 1]$.

For $[-m, m]$, set $\phi_n(t) = m p_n(t/m)$. Then ϕ_n is a polynomial without constant term, and $\phi_n \to g$ uniformly on $[-m, m]$. 👍

XX.10.2.9. Lemma. Let A be a uniformly closed algebra of real-valued bounded functions on a set X. If $f, g \in A$, then $f \vee g = \max(f, g)$ and $f \wedge g = \min(f, g)$ are in A.

Proof: $f \vee g = \frac{1}{2}(f + g + |f - g|)$ and $f \wedge g = \frac{1}{2}(f + g - |f - g|)$. 👍

By induction, if $f_1, \ldots, f_n \in A$, then $f_1 \vee \cdots \vee f_n$ and $f_1 \wedge \cdots \wedge f_n$ are in A.

XX.10.2.10. LEMMA. Let A be an algebra of real-valued functions on X which separates the points of X, and $x, y \in X$. If there are $f, g \in A$ with $f(x) \neq 0$ and $g(y) \neq 0$, then for any $a, b \in \mathbb{R}$ there is an $h \in A$ with $h(x) = a$ and $h(y) = b$.

PROOF: If $\phi = f^2 + g^2$, then $\phi(x) > 0$ and $\phi(y) > 0$. If $\phi(x) = \phi(y)$, let ψ be a function in A with $\psi(x) \neq \psi(y)$. Then for some (almost any) $r \in \mathbb{R}$ the function $\theta = \phi + r\psi$ has $c = \theta(x)$ and $d = \theta(y)$ distinct nonzero numbers. It is easy to find a quadratic polynomial p with no constant term such that $p(c) = a$ and $p(d) = b$ by solving two linear equations for the coefficients. Then $h = p \circ \theta$ is the desired element of A. 👍

We are thus in the situation where the following definition applies (for a more general version, see ()):

XX.10.2.11. DEFINITION. Let X be a compact Hausdorff space. A subspace L of $C_{\mathbb{R}}(X)$ is a *lattice* if $f \wedge g \in L$ for all $f, g \in L$.

A lattice L in $C_{\mathbb{R}}(X)$ has the *strong separation property* if, whenever $x, y \in X$, $x \neq y$, and $a, b \in \mathbb{R}$, there is an $f \in L$ with $f(x) = a$ and $f(y) = b$.

Since $f \vee g = -[(-f) \wedge (-g)]$, if L is a lattice and $f, g \in L$, then $f \vee g \in L$.

XX.10.2.12. A uniformly closed subalgebra A of $C_{\mathbb{R}}(X)$ is a lattice by XX.10.2.9., and, if A separates points and there is no point of X at which all the functions in A vanish, A has the strong separation property by XX.10.2.10.. Note also:

XX.10.2.13. PROPOSITION. Let X be a compact Hausdorff space, and L a lattice in $C_{\mathbb{R}}(X)$. If L separates points and the constant functions are in L, then L has the strong separation property.

PROOF: Let $x, y \in A$, $x \neq y$, and $a \leq b \in \mathbb{R}$. There is an $f \in L$ with $f(x) \neq f(y)$. Replacing f by $-f$ if necessary, we may assume $f(x) < f(y)$, and multiplying f by a scalar we may assume $f(y) - f(x) \geq b - a$. Adding a constant to f, we may assume $f(x) = a$, and thus $f(y) \geq b$. If $g = f \wedge (b1)$, then $g \in L$, $g(x) = a$, and $g(y) = b$. The case $a \geq b$ is obtained by interchanging x and y. 👍

We then have the lattice version of XX.10.2.2.:

XX.10.2.14. THEOREM. [STONE] Let X be a compact Hausdorff space, and L a lattice in $C_{\mathbb{R}}(X)$ with the strong separation property. Then L is uniformly dense in $C_{\mathbb{R}}(X)$.

PROOF: Let $f \in C_{\mathbb{R}}(X)$. Fix $\epsilon > 0$. Let $x \in X$. Let $\mathcal{U}$ be the set of all open sets U in X with the property that there is a $g \in L$ with $g(x) = f(x)$ and $|f(y) - g(y)| < \epsilon$ for all $y \in U$. Since there is a $g \in L$ with $g(x) = f(x)$, there is an open neighborhood of x in $\mathcal{U}$ by continuity. And if $z \neq x$, there is a $g \in L$ with $g(x) = f(x)$ and $g(z) = f(z)$, so there is an open neighborhood of z in $\mathcal{U}$ by continuity. Thus $\mathcal{U}$ is an open cover of X. Let $\{U_1, \ldots, U_n\}$ be a finite subcover, and g_k a function in L with $g_k(x) = f(x)$ and $|f(y) - g_k(y)| < \epsilon$ for all $y \in U_k$. Set $g_x = g_1 \wedge \cdots \wedge g_n \in L$. Then $g_x(x) = f(x)$ and $g_x(y) < f(y) + \epsilon$ for all $y \in X$.

Thus for every $x \in X$ there is an $h \in L$ with $h(x) = f(x)$ and $h(y) < f(y) + \epsilon$ for all $y \in X$. Let $\mathcal{V}$ be the collection of all open sets V in X for which there is an $h \in L$ with $h(y) < f(y) + \epsilon$ for all $y \in X$ and $|f(y) - h(y)| < \epsilon$ for all $y \in V$. By continuity every point of X has an open neighborhood in $\mathcal{V}$, so $\mathcal{V}$ is an open cover of X. Let $\{V_1, \ldots, V_m\}$ be a finite subcover, and $h_1, \ldots, h_m$ corresponding functions in L. If $h = h_1 \vee \cdots \vee h_m \in L$, then $|f(x) - h(x)| < \epsilon$ for all $x \in X$. 👍

XX.10.2.15. Corollary. Let X be a compact Hausdorff space, and L a lattice in $C_{\mathbb{R}}(X)$ which separates points and contains the constant functions. Then L is uniformly dense in $C_{\mathbb{R}}(X)$.

XX.10.2.16. The usual proof of XX.10.2.14. uses the AC. I am indebted to T. Tao for pointing out this version of the argument which does not use any choice (https://mathoverflow.net/questions/445361/stone-weierstrass-theorem-without-ac); this modification of the usual proof is similar to the modifications in the proof of XIII.11.4.1.–XIII.11.4.2. to eliminate use of the AC. I have been unable to find this argument in the literature, or even a clear statement that the Stone-Weierstrass Theorem does not depend on the AC.

XX.10.2.17. Examples. (i) Let X be a compact Hausdorff space, and $p \in X$. If L is the set of all functions in $C_{\mathbb{R}}(X)$ vanishing at p, then L is a closed subalgebra of $C_{\mathbb{R}}(X)$, hence a lattice, and separates points, but does not have the strong separation property (and is not dense in $C_{\mathbb{R}}(X)$).

(ii) A lattice which separates points does not necessarily have the strong separation property, even if there is no point where all the elements of L vanish. Let $X = \{a, b\}$ be a two-point space with the discrete topology, so that $C_{\mathbb{R}}(X) \cong \mathbb{R}^2$, and let $L = \{f : f(b) = 2f(a)\}$ (i.e. $L = \{(x, 2x) : x \in \mathbb{R}\} \subseteq \mathbb{R}^2\}$). Then L is a lattice in $C_{\mathbb{R}}(X)$ which separates points, but does not have the strong separation property. (L does not contain the constant functions and is not dense in $C_{\mathbb{R}}(X)$.)

(iii) For L to be a lattice in $C_{\mathbb{R}}(X)$ by our definition, it is not enough that L be a lattice in the induced ordering from $C_{\mathbb{R}}(X)$. Let $X = \{a, b, c\}$ be a three-point space with the discrete topology, so that $C_{\mathbb{R}}(X) \cong \mathbb{R}^3$. Let $L = \left\{f : f(c) = \frac{1}{2}(f(a) + f(b))\right\}$ (i.e. $L = \left\{\left(x, y, \frac{1}{2}(x + y)\right) : x, y \in \mathbb{R}\right\} \subseteq \mathbb{R}^3$). Then L is a subspace of $C_{\mathbb{R}}(X)$ containing the constants which has the strong separation property, and L is a lattice in the induced ordering. But if $f = \left(1, 0, \frac{1}{2}\right)$ and $g = \left(0, 1, \frac{1}{2}\right)$, the supremum of f and g in L is $(1, 1, 1)$, which is not the supremum $\left(1, 1, \frac{1}{2}\right)$ in $C_{\mathbb{R}}(X)$. And L is not dense in $C_{\mathbb{R}}(X)$.

XX.10.2.18. We now prove XX.10.2.1.. Suppose first that for every $x \in X$ there is a $g \in A$ with $g(x) \neq 0$. Then the hypotheses of XX.10.2.14. hold (XX.10.2.12.), so A is uniformly dense in $C_{\mathbb{R}}(X)$, and since A is uniformly closed, $A = C_{\mathbb{R}}(X)$.

Now suppose there is a $p \in X$ such that $g(p) = 0$ for all $g \in A$. (There cannot be more than one such p since A separates points.) Let B be the set of all functions of the form $f = g + c$, where $g \in A$ and c is a constant. Note that for such an f, $c = f(p)$. It is clear that B is an algebra of functions. Suppose (f_n) is a sequence in B, and $f_n \to f$ uniformly. Write $f_n = g_n + c_n$, with $g_n \in A$ and $c_n = f_n(p)$ a constant. Then $c_n \to c = f(p)$, and hence $g_n \to g = f - c$ uniformly. Thus $g \in A$ since A is uniformly closed, and $f = g + c \in B$. Thus B is uniformly closed. By the first part of the proof, $B = C_{\mathbb{R}}(X)$. Now

suppose $f \in C_{\mathbb{R}}(X)$ with $f(p) = 0$. Then $f \in B$, so $f = g + c$ for some $g \in A$ and constant c. But $0 = f(p) = g(p) + c = 0 + c = c$, so $f = g \in A$.

This completes the proof of XX.10.2.1.. 

XX.10.3. Exercises

XX.10.3.1. Prove the following corollary of XX.10.2.1.:

THEOREM. Let X be a locally compact Hausdorff space. Let A be a uniformly closed subalgebra of $C^{\circ}_{\mathbb{R}}(X)$ which separates points of X, such that there is no point of X at which all the functions in A vanish. Then $A = C^{\circ}_{\mathbb{R}}(X)$.

[Regard A as a subalgebra of $C_{\mathbb{R}}(X^{\dagger})$.]

XX.10.3.2. Let X be a locally compact, σ-compact Hausdorff space. Let A be an algebra of real-valued continuous functions on X which separate points of X, such that there is no point of X at which all the functions in A vanish. Show that if f is any real-valued continuous function on X, then there is a sequence (f_n) of functions in A converging u.c. to f on X.

XX.10.3.3. **General Weierstrass Approximation Theorem**. Let X be a compact subset of $\mathbb{R}^n$ for some n.

(a) Show that any continuous function from X to $\mathbb{R}$ can be uniformly approximated arbitrarily closely by a polynomial in the coordinates (in $\mathbb{R}^n$).

(b) Show that any continuous function from X to $\mathbb{R}^m$ for any m can be uniformly approximated arbitrarily closely by a function whose m coordinate functions are all polynomials in the coordinates (in $\mathbb{R}^n$).

XX.10.3.4. Let X be a compact subset of $\mathbb{C}$. Let A be the subalgebra of $C(X)$ consisting of all functions of the form $g(z) = p(z, \bar{z})$, where p is a polynomial of two variables with complex coefficients. Then A separates points of X and contains the constant functions. Show that A is closed under complex conjugation and hence is dense in $C(X)$ by XX.10.2.4..

XX.10.3.5. (a) Let X and Y be compact Hausdorff spaces. Show that every real-valued continuous function on $X \times Y$ can be uniformly approximated arbitrarily closely by (finite) sums of functions of the form $f \otimes g$, where $f \in C_{\mathbb{R}}(X)$, $g \in C_{\mathbb{R}}(Y)$, and

$$(f \otimes g)(x, y) = f(x)g(y) \ .$$

(b) Prove a similar result for complex-valued continuous functions. In fact, every complex-valued continuous function on $X \times Y$ is a uniform limit of finite sums of functions of the form $\lambda(f \otimes g)$, where λ is 1 or i and f and g are real-valued continuous functions on X and Y respectively.

(c) Use (a) to give an alternate proof of Fubini's Theorem for continuous functions (IX.3.4.7.):

THEOREM. Let $R = [a, b] \times [c, d]$ be a rectangle in $\mathbb{R}^2$, and $h : R \to \mathbb{R}$ a continuous function. Then

$$\iint_R h \, dA = \int_a^b \int_c^d h(x, y) \, dy \, dx = \int_c^d \int_a^b h(x, y) \, dx \, dy \ .$$

[First prove the theorem if $h = f \otimes g$, and take uniform limits.] Generalize to Radon measures on general compact Hausdorff spaces.

XX.10.3.6. Let $a, b \in \mathbb{R}$, $a < b$. Let $C_{sing}([a,b])$ be the set of all continuous functions $f : [a,b] \to \mathbb{R}$ such that f is differentiable a.e. and $f' = 0$ a.e. on $[a,b]$.

(a) Show that $C_{sing}([a,b])$ is a subalgebra of $C_{\mathbb{R}}([a,b])$ containing the constant functions.

(b) By scaling the Cantor function, show that $C_{sing}([a,b])$ separates points of $[a,b]$.

(c) Conclude that $C_{sing}([a,b])$ is uniformly dense in $C_{\mathbb{R}}([a,b])$.

There is an explicit proof of this result not using the Stone-Weierstrass Theorem; cf. XIX.5.8.6..

XX.10.3.7. Let f be a continuous function on an interval $[a,b]$, and suppose that all the *moments*

$$\int_a^b f(x)x^n \, dx$$

($n \in \mathbb{N}_0$) of f are zero. Prove that f is identically zero on $[a,b]$. [It suffices to show that

$$\int_a^b [f(x)]^2 \, dx = 0 \ .$$

Show that

$$\int_a^b f(x)p(x) \, dx = 0$$

for any polynomial p, and apply the Weierstrass Approximation Theorem.]

XX.10.3.8. Here is an "integral" version of Weierstrass's Theorem.

(a) If f is a polynomial with integer coefficients, regarded as an element of $C([0,1])$, then of course $f(0)$ and $f(1)$ are integers. The rest of the problem will show that if f is an element of $C([0,1])$ such that $f(0)$ and $f(1)$ are integers, then f is a uniform limit on $[0,1]$ of polynomials with integer coefficients.

(b) Let $g(x) = x + x(1-2x)(1-x) = 2x - 3x^2 + 2x^3$. Show that g is a strictly increasing function from $[0,1]$ onto $[0,1]$. The function g has fixed points 0, 1/2, and 1; if $0 < x < 1/2$, then $x < g(x) < 1/2$; if $1/2 < x < 1$, then $1/2 < g(x) < x$.

(c) Show that the iterates (g_i) of g converge pointwise on $(0,1)$ to the constant function $1/2$, and that the convergence is uniform on compact subsets (i.e. on subintervals of the form $[\epsilon, 1-\epsilon]$ for $\epsilon > 0$). Such iterates are strictly increasing.

(d) Show by induction on n that if k is any integer, $1 \le k \le 2^n - 1$, and $\epsilon > 0$, then there are integer linear combinations (ϕ_m) of powers of the g_i and $1 - g_i$ such that $0 \le \phi_m \le 1$ on $[0,1]$ and such that (ϕ_m) converges uniformly to the constant function $\frac{k}{2^n}$ on $[\epsilon, 1-\epsilon]$.

(e) If p is a polynomial with integer coefficients and $p(0) = p(1) = 0$, and $\lambda \in \mathbb{R}$, show that there is a sequence of polynomials with integer coefficients converging uniformly to λp on $[-1,1]$.

(f) If f is a continuous function on $[0,1]$ such that $f(0)$ and $f(1)$ are integers, then there is a linear polynomial q with integer coefficients with $q(x) = f(x)$ for $x = 0, 1$. Apply the Stone-Weierstrass Theorem and (e) to conclude that $f - q$, and hence f, can be approximated uniformly on $[0,1]$ by polynomials with integer coefficients.

XX.10.3.9. This problem is an extension of Exercise XX.10.3.8..

(a) Let f be a polynomial with integer coefficients, regarded as an element of $C([-1,1])$. Show that $f(-1)$, $f(0)$, and $f(1)$ are integers, and that $f(-1) \equiv f(1) \bmod 2$ (i.e. both are even or both odd). The rest of the problem will show that if f is an element of $C([-1,1])$ such that $f(-1)$, $f(0)$, and $f(1)$ are integers, and $f(-1) \equiv f(1) \bmod 2$, then f is a uniform limit on $[-1,1]$ of polynomials with integer coefficients.

(b) If p is a polynomial with integer coefficients and $p(-1) = p(0) = p(1) = 0$, and $\lambda \in \mathbb{R}$, show that there is a sequence of polynomials with integer coefficients converging uniformly to λp on $[-1,1]$. [Consider functions of the form $\psi(x) = \phi(1-x^2)$, where ϕ is one of the ϕ_m from 1(d).]

(c) If f is a continuous function on $[-1,1]$ such that $f(-1)$, $f(0)$, and $f(1)$ are integers and $f(-1) \equiv f(1) \bmod 2$, then there is a quadratic polynomial q with integer coefficients with $q(x) = f(x)$ for $x = -1, 0, 1$. Apply the Stone-Weierstrass Theorem and (b) to conclude that $f - q$, and hence f, can be approximated uniformly on $[-1,1]$ by polynomials with integer coefficients.

XX.10.3.10. (a) Let $\mathbb{T}$ be the unit circle in $\mathbb{C}$. Each polynomial with complex coefficients can be regarded as a continuous function from $\mathbb{T}$ to $\mathbb{C}$, and this set of polynomials forms a subalgebra P of $C(\mathbb{T})$. P contains the constant functions and separates points of $\mathbb{T}$. But P is not closed under complex conjugation (in fact, if p and $\bar{p}$ are both in P, then p is a constant function).

(b) Regard $\mathbb{T}$ alternatively as $[0, 2\pi]$ with the endpoints identified, and $C(\mathbb{T})$ as the algebra of continuous functions $f : [0, 2\pi] \to \mathbb{C}$ with $f(2\pi) = f(0)$. P then becomes an algebra of "trigonometric polynomials" on $[0, 2\pi]$ (but not all trigonometric polynomials), functions of the form

$$f(t) = \sum_{k=0}^{n} c_k e^{ikt}$$

for some $n \in \mathbb{N}$ (depending on f) and $\{c_k\}$ in $\mathbb{C}$.

(c) If $f \in P$, then $\int_0^{2\pi} f(t)e^{it}\,dt = 0$. The same holds if f is in the uniform closure $\bar{P}$ of P in $C(\mathbb{T})$.

(d) For $z \in \mathbb{T}$, let $f(z) = \bar{z}$. Alternatively, if $t \in [0, 2\pi]$, let $f(t) = e^{-it}$. Then $f \in C(\mathbb{T})$, and $\int_0^{2\pi} f(t)e^{it}\,dt = 2\pi$. Thus f is not in $\bar{P}$.

(e) Let $\mathbb{D}$ be the open unit disk in $\mathbb{C}$. Each $f \in P$ defines an analytic function on $\mathbb{D}$. If (f_n) is a sequence in P converging uniformly on $\mathbb{T}$ to a function $f \in \bar{P}$, then, by the Maximum Modulus Theorem, (f_n) also converges uniformly on $\bar{\mathbb{D}}$ to a function which is continuous on $\bar{\mathbb{D}}$ and analytic on $\mathbb{D}$. Conversely, if f is continuous on $\bar{\mathbb{D}}$ and analytic on $\mathbb{D}$, the Taylor polynomials (partial sums of the Taylor series) for f around 0 converge uniformly to f on $\bar{\mathbb{D}}$ (and, in particular, on $\mathbb{T}$), so f is in $\bar{P}$. Thus $\bar{P}$ consists precisely of the complex-valued continuous functions on $\mathbb{T}$ which extend to a function continuous on $\bar{\mathbb{D}}$ and analytic on $\mathbb{D}$.

(f) $\bar{P}$ also consists precisely of the functions in $C(\mathbb{T})$ whose negative Fourier coefficients vanish ().

(g) A *trigonometric polynomial* on $[0, 2\pi]$ is a function of the form

$$f(t) = \sum_{k=-n}^{n} c_k e^{ikt}$$

for some $n \in \mathbb{N}$ (depending on f) and $\{c_k\}$ in $\mathbb{C}$. Then $f \in C(\mathbb{T})$. A trigonometric polynomial can also be written

$$f(t) = \sum_{k=0}^{n} a_k \cos kt + i \sum_{k=1}^{n} b_k \sin kt$$

for $n \in \mathbb{N}$ and $a_k, b_k \in \mathbb{C}$. Show that the trigonometric polynomials are dense in $C(\mathbb{T})$.

See books on function algebras, e.g. [Bro69], for a thorough study of more sophisticated examples of this type.

XX.10.3.11. Let X be a compact Hausdorff space, and let P be a set of nonnegative real-valued continuous functions on X with the following properties:

(i) $f + g \in P$ for all $f, g \in P$.

(ii) $\alpha f \in P$ for all $f \in P$, $\alpha > 0$.

(iii) Either

(1) $f \wedge g \in P$ for all $f, g \in P$

or

(2) $f \vee g \in P$ for all $f, g \in P$

Let $L = P - P = \{f - g : f, g \in P\}$.

(a) Show that L is a lattice in $C_{\mathbb{R}}(X)$. [Use

$$(f_1 - g_1) \wedge (f_2 - g_2) = [(f_1 + g_2) \wedge (f_2 + g_1)] - (g_1 + g_2)$$

$$(f_1 - g_1) \vee (f_2 - g_2) = [(f_1 + g_2) \vee (f_2 + g_1)] - (g_1 + g_2)$$

and $f \vee g = -[(-f) \wedge (-g)]$.]

(b) Conclude from XX.10.2.15. that if P separates points and contains a nonzero constant function, then L is uniformly dense in $C_{\mathbb{R}}(X)$.

Examples of sets P satisfying the hypotheses are the set of nonnegative continuous convex functions on an interval $[a, b]$ and the set of nonnegative continuous concave functions on $[a, b]$. Thus every continuous function on $[a, b]$ can be uniformly approximated arbitrarily closely by the sum of a continuous convex function and a continuous concave function (difference of two continuous convex functions).

XX.10.3.12. Use the Stone-Weierstrass Theorem to give another proof of the Tietze Extension Theorem () for compact Hausdorff spaces: let X be a compact Hausdorff space, and Z a closed subset of X. Let A be the subalgebra of $C_{\mathbb{R}}(Z)$ consisting of the restrictions to Z of continuous functions from X to $\mathbb{R}$.

(a) Use Urysohn's Lemma to show that A is uniformly closed in $C_{\mathbb{R}}(Z)$. (This is a bit tricky.)

(b) Conclude from the Stone-Weierstrass Theorem that $A = C_{\mathbb{R}}(Z)$, i.e. that every continuous function from Z to $\mathbb{R}$ extends to X.

This proof is not easier than the proof in (), and the result is less general, but it is a different approach. A simpler argument for (a) can be given using C*-algebra theory (XX.16.2.3.).

XX.10.3.13. Prove the following converse to XX.10.2.2. (there is an analogous converse to XX.10.2.4.). If X is completely regular, let $BC_{\mathbb{R}}(X)$ be the algebra of bounded real-valued continuous functions on X. $BC_{\mathbb{R}}(X)$ is a commutative real Banach algebra under the supremum norm, isometrically isomorphic to $C_{\mathbb{R}}(\beta X)$ ().

THEOREM. Let X be a completely regular topological space. If every unital subalgebra of $BC_{\mathbb{R}}(X)$ which separates points of X is (uniformly) dense in $BC_{\mathbb{R}}(X)$, then X is compact.

[If X is not compact, choose $p \in X$ and $q \in \beta X \setminus X$, and let A be the subalgebra of $BC_{\mathbb{R}}(X) \cong C_{\mathbb{R}}(\beta X)$ of functions f with $f(p) = f(q)$; then A contains the constants and separates points of X. $A \cong C_{\mathbb{R}}(Y)$, where Y is the quotient of βX obtained by identifying p and q.]

XX.10.3.14. A $\mathcal{C}^\infty$ Weierstrass Approximation Theorem. Let f be a $\mathcal{C}^\infty$ function on the interval $I = [a, b]$.

(a) Let $n \in \mathbb{N}$ and $\epsilon > 0$. Show that there is a polynomial p such that

$$|f^{(k)}(x) - p^{(k)}(x)| < \epsilon$$

for all k, $0 \leq k \leq n$, and all $x \in I$. [Set $c = \max(1, b-a)$. Let q_n be a polynomial with $|f^{(n)}(x) - q_n(x)| < \frac{\epsilon}{c^n}$ for all $x \in I$. Successively set

$$q_{k-1}(x) = f^{(k-1)}(a) + \int_a^x q_k(t)\, dt$$

for $k = n, n-1, \ldots, 1$, and set $p = q_0$.]

(b) Show that there is a sequence (p_n) of polynomials such that $p_n^{(k)} \to f^{(k)}$ uniformly on I for each $k \in \mathbb{N} \cup \{0\}$.

(c) Conclude that the polynomials are dense in $\mathcal{C}^\infty(I)$ for its natural topology ().

(d) Show similarly that if $r \in \mathbb{N}$, then the polynomials are dense in $\mathcal{C}^r(I)$ for its natural topology ().

(e) If I is a closed box in $\mathbb{R}^n$, similarly prove that polynomials in the coordinates are dense in $\mathcal{C}^\infty(I)$ and $C^r(I)$ for their natural topologies.

(f) Formulate and prove a version of the Stone-Weierstrass Theorem giving necessary and sufficient conditions for a subalgebra of $\mathcal{C}^r(I)$ or $\mathcal{C}^\infty(I)$ to be dense for the natural topology.

XX.11. Hilbert Spaces

A Hilbert space is a vector space over $\mathbb{R}$ or $\mathbb{C}$ with an inner product or "dot product" which is complete with respect to the norm induced by the inner product. A Hilbert space can be thought of as a special kind of Banach space which has an intrinsic Euclidean geometry; in fact, every finite-dimensional subspace of a Hilbert space is a Euclidean space.

Hilbert spaces are the nicest kinds of topological vector spaces, essentially Euclidean spaces which are allowed to be infinite-dimensional. They share many pleasant properties with Euclidean spaces. In fact, they are so nice they are not themselves very interesting mathematically: there is up to isometric isomorphism exactly one of each dimension, finite or infinite. Like vector spaces, they are mainly important not in and of themselves, but as the setting for operators, which are highly nontrivial and very important.

Hilbert spaces can be either real or complex. The theories of real and complex Hilbert spaces are essentially identical and can be done in parallel. When we get to operators, however, the theories diverge somewhat, and for technical reasons (e.g. the Fundamental Theorem of Algebra) operator theory works better on complex Hilbert spaces. Thus in applications and more advanced work on operators, complex Hilbert spaces are used almost exclusively. Adhering to custom, we will take the default meaning of the term "Hilbert space" to be "complex Hilbert space."

XX.11.1. Inner Products

Although a unified treatment can be given (XX.11.1.10.), it is clearer to separately define inner products in real and complex vector spaces.

Real Inner Products

XX.11.1.1. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{R}$. A *bilinear form* on $\mathcal{V}$ is a function

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$$

(i.e. a function which assigns to each ordered pair of vectors a scalar) which is linear in each entry when the other entry is fixed, i.e.

(i) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle$ for all $x_1, x_2, y \in \mathcal{V}$.

(i′) $\langle x, y_1 + y_2 \rangle = \langle x, y_1 \rangle + \langle x, y_2 \rangle$ for all $x, y_1, y_2 \in \mathcal{V}$.

(ii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{R}$.

(ii′) $\langle x, \alpha y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{R}$.

A bilinear form is *symmetric* if

(iii) $\langle y, x \rangle = \langle x, y \rangle$ for all $x, y \in \mathcal{V}$.

A symmetric bilinear form is *positive* (or *positive semidefinite*) if

(iv) $\langle x, x \rangle \geq 0$ for all $x \in \mathcal{V}$.

A positive symmetric bilinear form is *definite* (or *positive definite*) if

(v) $\langle x, x \rangle = 0$ for $x \in \mathcal{V}$ if and only if $x = 0$.

A positive symmetric bilinear form is called a *real pre-inner product.*

A positive definite symmetric bilinear form is called a *real inner product.*

XX.11.1.2. REMARKS. (i) Note that the inner product of two vectors is a *scalar*. Thus an inner product is not really a "multiplication" on the vector space, although it has some flavor of a product.

(ii) In a symmetric form, properties (i′) and (ii′) follow automatically from (i) and (ii) respectively. Since most forms we consider are symmetric (or hermitian in the complex case), we will rarely have to separately consider (i′) and (ii′).

XX.11.1.3. **Notation.** The notation $\langle x, y\rangle$ is probably the most common notation for an inner product. Other common notations are (x, y), $\langle x|y\rangle$, and $(x|y)$. Additional notations are also occasionally used.

XX.11.1.4. EXAMPLES. (i) On $\mathbb{R}^n$, the usual dot product is a real inner product: if $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, then

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x} \cdot \boldsymbol{y} = x_1y_1 + \cdots + x_ny_n = \sum_{k=1}^{n} x_ky_k \ .$$

This is the *standard inner product* on $\mathbb{R}^n$.

(ii) On $C_{\mathbb{R}}([a, b])$, the real vector space of continuous real-valued functions on $[a, b]$, define

$$\langle f, g\rangle = \int_a^b f(t)g(t)\,dt \ .$$

then $\langle \cdot, \cdot\rangle$ is a real inner product on $C_{\mathbb{R}}([a, b])$.

These two examples obviously have considerable similarity, and can be adapted to a wide range of settings. In fact, these are the "generic" examples; every inner product looks essentially like them (XX.11.5.21.).

(iii) Write elements of $\mathbb{R}^n$ as column vectors. If A is an $n \times n$ matrix over $\mathbb{R}$, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ define

$$(\boldsymbol{x}, \boldsymbol{y}) = \langle A\boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{y}^t A\boldsymbol{x}$$

where $\langle \cdot, \cdot\rangle$ is the standard inner product. Then $(\cdot, \cdot)$ is a bilinear form on $\mathbb{R}^n$. We have

$$(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{x}^t A\boldsymbol{y} = [\boldsymbol{y}^t A^t\boldsymbol{x}]^t = \boldsymbol{y}^t A^t\boldsymbol{x} = \langle A^t\boldsymbol{x}, \boldsymbol{y}\rangle$$

since transpose is the identity on $\mathbb{R}$; thus $(\cdot, \cdot)$ is symmetric if and only if A is symmetric (i.e. $A^t = A$). The form is a pre-inner product if and only if A is positive (), or equivalently if $A = A^t$ and all eigenvalues of A are nonnegative; it is an inner product if and only if A is positive definite, i.e. positive and invertible (i.e. $A = A^t$ and all eigenvalues of A are strictly positive).

We can recover A: if A_{jk} is the (j, k) entry of A, then $A_{jk} = (\boldsymbol{e}_k, \boldsymbol{e}_j)$, where $\boldsymbol{e}_i$ is the i'th standard basis vector. Conversely, if $[\cdot, \cdot]$ is any bilinear form on $\mathbb{R}^n$, set $B_{jk} = [\boldsymbol{e}_k, \boldsymbol{e}_j]$; then a simple calculation shows that

$$[\boldsymbol{x}, \boldsymbol{y}] = \langle B\boldsymbol{x}, \boldsymbol{y}\rangle$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, where $B = (B_{jk})$, so every bilinear form on $\mathbb{R}^n$ is of this type. The matrix B is called the *matrix of the bilinear form* with respect to the standard basis of $\mathbb{R}^n$.

Complex Inner Products

XX.11.1.5. The definitions in XX.11.1.1. make sense if $\mathbb{R}$ is merely replaced by $\mathbb{C}$. However, they are not very interesting. In fact, bilinearity is essentially incompatible with positivity: if $\alpha \in \mathbb{C}$ with $\alpha \notin \mathbb{R}$, and $x \in \mathcal{V}$, if $\langle \cdot, \cdot \rangle$ is bilinear we have

$$\langle \alpha x, \alpha x \rangle = \alpha^2 \langle x, x \rangle$$

and since α^2 is not a nonnegative real number, $\langle x, x \rangle$ and $\langle \alpha x, \alpha x \rangle$ cannot both be nonnegative unless they are zero.

Fortunately, there is a fairly simple way around the difficulty, based on the fact that if $\alpha \in \mathbb{C}$, then $\alpha \bar{\alpha} \geq 0$. We replace bilinearity by "sesquilinearity" (*sesqui-* is a Latin prefix meaning "one and a half"):

XX.11.1.6. DEFINITION. Let $\mathcal{V}$ be a vector space over $\mathbb{C}$. A *sesquilinear form* on $\mathcal{V}$ is a function

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$$

(i.e. a function which assigns to each ordered pair of vectors a scalar) which is linear in the first entry when the second entry is fixed and conjugate-linear in the second entry when the first entry is fixed, i.e.

(i) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle$ for all $x_1, x_2, y \in \mathcal{V}$.

(i′) $\langle x, y_1 + y_2 \rangle = \langle x, y_1 \rangle + \langle x, y_2 \rangle$ for all $x, y_1, y_2 \in \mathcal{V}$.

(ii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{C}$.

(ii′) $\langle x, \alpha y \rangle = \bar{\alpha} \langle x, y \rangle$ for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{C}$.

A sesquilinear form is *hermitian* if

(iii) $\langle y, x \rangle = \overline{\langle x, y \rangle}$ for all $x, y \in \mathcal{V}$.

A hermitian sesquilinear form is *positive* (or *positive semidefinite*) if

(iv) $\langle x, x \rangle \geq 0$ for all $x \in \mathcal{V}$.

A positive hermitian sesquilinear form is *definite* (or *positive definite*) if

(v) $\langle x, x \rangle = 0$ for $x \in \mathcal{V}$ if and only if $x = 0$.

A positive hermitian sesquilinear form is called a *pre-inner product.*

A positive definite hermitian sesquilinear form is called a *(complex) inner product.*

XX.11.1.7. **Notation:** We have made the convention that sesquilinear forms and inner products are linear in the first variable and conjugate-linear in the second, the usual convention in mathematics. For obscure reasons, the opposite convention is common in mathematical physics, so readers should be aware of the convention used in any reference. The difference will have no practical effect, and a trivial notational change (reversing order in the inner product) will convert to the opposite convention.

XX.11.1.8. EXAMPLES. (i) On $\mathbb{C}^n$, if $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, define

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = x_1\overline{y_1} + \cdots + x_n\overline{y_n} = \sum_{k=1}^{n} x_k\overline{y_k} \ .$$

This is an inner product, called the *standard inner product* on $\mathbb{C}^n$.

(ii) On $C([a,b])$, the complex vector space of continuous complex-valued functions on $[a,b]$, define

$$\langle f, g \rangle = \int_a^b f(t)\overline{g(t)}\, dt \ .$$

then $\langle \cdot, \cdot \rangle$ is an inner product on $C([a,b])$.

These two examples obviously have considerable similarity, and can be adapted to a wide range of settings. In fact, these are the "generic" examples; every inner product looks essentially like them (XX.11.5.21.).

(iii) Write elements of $\mathbb{C}^n$ as column vectors. If A is an $n \times n$ matrix over $\mathbb{C}$, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ define

$$(\boldsymbol{x}, \boldsymbol{y}) = \langle A\boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^* A \boldsymbol{x}$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product and, for a matrix B,

$$B^* = \overline{B^t}$$

is the complex conjugate of the transpose. Then $(\cdot, \cdot)$ is a sesquilinear form on $\mathbb{C}^n$. We have

$$(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{x}^* A \boldsymbol{y} = [\boldsymbol{y}^* A^* \boldsymbol{x}]^* = \overline{\boldsymbol{y}^* A^* \boldsymbol{x}} = \overline{\langle A^* \boldsymbol{x}, \boldsymbol{y} \rangle}$$

and thus $(\cdot, \cdot)$ is hermitian if and only if A is hermitian (i.e. $A^* = A$). The form is a pre-inner product if and only if A is positive (), or equivalently if $A = A^*$ and all eigenvalues of A are nonnegative; it is an inner product if and only if A is positive definite, i.e. positive and invertible (i.e. $A = A^*$ and all eigenvalues of A are strictly positive).

We can recover A: if A_{jk} is the (j,k) entry of A, then $A_{jk} = (\boldsymbol{e}_k, \boldsymbol{e}_j)$, where $\boldsymbol{e}_i$ is the i'th standard basis vector. Conversely, if $[\cdot, \cdot]$ is any sesquilinear form on $\mathbb{C}^n$, set $B_{jk} = [\boldsymbol{e}_k, \boldsymbol{e}_j]$; then a simple calculation shows that

$$[\boldsymbol{x}, \boldsymbol{y}] = \langle B\boldsymbol{x}, \boldsymbol{y} \rangle$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$, where $B = (B_{jk})$, so every sesquilinear form on $\mathbb{C}^n$ is of this type. The matrix B is called the *matrix of the sesquilinear form* with respect to the standard basis of $\mathbb{C}^n$.

XX.11.1.9. If $\mathcal{V}$ is a complex vector space, then $\mathcal{V}$ can be regarded as a real vector space by restricting scalar multiplication. However, if $\langle \cdot, \cdot \rangle$ is a (complex) inner product on $\mathcal{V}$, it does not become a real inner product by simply restricting scalar multiplication (for one thing, it is complex-valued, not real-valued). But there is a simple conversion: if we define

$$(x, y) = Re(\langle x, y \rangle)$$

then it is easily checked that $(\cdot, \cdot)$ is a real inner product on $\mathcal{V}$ satisfying $(x,x) = \langle x, x \rangle$ for all $x \in \mathcal{V}$.

A Unified Treatment

A unified treatment of real and complex inner products can be given by making the convention that "complex conjugation" $\alpha \mapsto \bar{\alpha}$ is the identity map on $\mathbb{R}$. With this convention, we may meld the two definitions:

XX.11.1.10. DEFINITION. Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$, and $\mathcal{V}$ a vector space over $\mathbb{F}$. A *pre-inner product* on $\mathcal{V}$ is a function

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{F}$$

satisfying

(i) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle$ for all $x_1, x_2, y \in \mathcal{V}$.

(ii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$ for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{F}$.

(iii) $\langle y, x \rangle = \overline{\langle x, y \rangle}$ for all $x, y \in \mathcal{V}$.

(iv) $\langle x, x \rangle \geq 0$ for all $x \in \mathcal{V}$.

The pre-inner product is an *inner product* if it additionally satisfies

(v) $\langle x, x \rangle = 0$ for $x \in \mathcal{V}$ if and only if $x = 0$.

It is automatic that such a pre-inner product also satisfies

(i′) $\langle x, y_1 + y_2 \rangle = \langle x, y_1 \rangle + \langle x, y_2 \rangle$ for all $x, y_1, y_2 \in \mathcal{V}$.

(ii′) $\langle x, \alpha y \rangle = \bar{\alpha} \langle x, y \rangle$ for all $x, y \in \mathcal{V}$, $\alpha \in \mathbb{F}$.

XX.11.1.11. DEFINITION. A *(real or complex) inner product space* is a pair $(\mathcal{V}, \langle \cdot, \cdot \rangle)$, where $\mathcal{V}$ is a vector space over $\mathbb{R}$ or $\mathbb{C}$ and $\langle \cdot, \cdot \rangle$ is an inner product on $\mathcal{V}$.

An inner product space is a vector space $\mathcal{V}$ with specified inner product on $\mathcal{V}$. As usual, we will sometimes abuse language by saying "Let $\mathcal{V}$ be an inner product space." There will then implicitly be a specified inner product on $\mathcal{V}$. (By convention, we will take "inner product space" to mean "complex inner product space" unless otherwise qualified.)

XX.11.2. The CBS Inequality and the Norm

An inner product on a real or complex vector space defines a "norm"

$$\|x\| = \sqrt{\langle x, x \rangle}$$

but to prove that this is truly a norm, specifically to prove the triangle inequality, requires a preliminary inequality which is itself extremely important.

XX.11.2.1. THEOREM. [CBS INEQUALITY] Let $\langle \cdot, \cdot \rangle$ be a pre-inner product on a real or complex vector space $\mathcal{V}$. Then

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle$$

for all $x, y \in \mathcal{V}$. If the pre-inner product is an inner product, we have equality if and only if $\{x, y\}$ is linearly dependent.

PROOF: Fix $x, y \in \mathcal{V}$. Suppose first that $\langle y, y \rangle \neq 0$ and set

$$\alpha = -\frac{\langle x, y \rangle}{\langle y, y \rangle} \ .$$

We then have

$$\begin{aligned} 0 \leq \langle x + \alpha y, x + \alpha y \rangle &= \langle x, x \rangle + \alpha \langle y, x \rangle + \bar{\alpha} \langle x, y \rangle + \alpha \bar{\alpha} \langle y, y \rangle \\ &= \langle x, x \rangle - \frac{\langle x, y \rangle}{\langle y, y \rangle} \langle y, x \rangle - \frac{\langle y, x \rangle}{\langle y, y \rangle} \langle x, y \rangle + \frac{\langle x, y \rangle \langle y, x \rangle}{\langle y, y \rangle^2} \langle y, y \rangle \\ &= \langle x, x \rangle - \frac{|\langle x, y \rangle|^2}{\langle y, y \rangle} \end{aligned}$$

and putting the second term on the left and multiplying by $\langle y, y \rangle$ gives the inequality.

Now suppose $\langle y, y \rangle = 0$. If the pre-inner product is an inner product, then $y = 0$, so $\langle x, y \rangle = 0$ by (real) linearity in the second variable and there is nothing to prove. For the general case, set $\beta = -n\langle x, y \rangle$ for $n \in \mathbb{N}$. Then

$$\begin{aligned} 0 \leq \langle x + \beta y, x + \beta y \rangle &= \langle x, x \rangle + \beta \langle y, x \rangle + \bar{\beta} \langle x, y \rangle + \beta \bar{\beta} \langle y, y \rangle \\ &= \langle x, x \rangle + \beta \langle y, x \rangle + \bar{\beta} \langle x, y \rangle = \langle x, x \rangle - 2n |\langle x, y \rangle|^2 \ . \end{aligned}$$

This must be true for all $n \in \mathbb{N}$, so $\langle x, y \rangle = 0$.

If x or y is zero, we clearly have equality. Suppose x and y are nonzero and we have equality, and the pre-inner product is an inner product. Then there are suitable scalar multiples z and w of x and y respectively with

$$\langle z, z \rangle = \langle w, w \rangle = \langle z, w \rangle = 1 \ .$$

We then have

$$\langle z - w, z - w \rangle = \langle z, z \rangle - 2\langle z, w \rangle + \langle w, w \rangle = 0$$

so $z = w$, i.e. x and y are scalar multiples of each other. 👍

XX.11.2.2. The name "CBS inequality" is not the one most commonly used, but is the most appropriate. The names CAUCHY, BUNYAKOVSKII, and SCHWARZ are associated with the inequality, and it is commonly called by various subsets of these names (most commonly the "Cauchy-Schwarz inequality.") Here is a summary of the history:

(i) The first version of the CBS inequality to appear in print was for the standard inner product in $\mathbb{R}^n$, by CAUCHY in [Cau21] in 1821. See Exercise XX.11.11.1. for CAUCHY'S proof.

(ii) In 1859, V. BUNYAKOVSKII stated and indicated a proof of the inequality for the inner product of XX.11.1.4.(ii) [Bun59]. This paper, although written in French and not Russian (BUNYAKOVSKII was actually

Ukrainian and studied in Paris), was published in Russia and was apparently little noticed in Western Europe. See Exercise XX.11.11.3. for BUNYAKOVSKII'S statement and argument.

(iii) In 1885, H. SCHWARZ published an important and influential paper on minimal surfaces [Sch88] in which he proved, as a relatively minor technical result, a slightly more general integral version of the inequality. He was apparently unaware of BUNYAKOVSKII'S paper. Because SCHWARZ'S paper was widely read and studied, the inequality became known and was often called the "Schwarz inequality." SCHWARZ did make a substantive contribution: although he only stated the result for integrals, he gave a simple and elegant proof which works verbatim for any real inner product. See Exercise XX.11.11.4. for SCHWARZ'S proof.

(iv) It is unclear who first stated the inequality for complex inner products, or who first discovered the proof given above. The arguments of CAUCHY, BUNYAKOVSKII, and SCHWARZ do not work directly for complex inner products, although the complex case can be easily derived from the real case; see Exercise XX.11.11.5..

The website http://www-stat.wharton.upenn.edu/~steele/Publications/Books/CSMC/CSMC_index.html for the book [Ste04] contains more information and links to electronic copies of the papers by BUNYAKOVSKII and SCHWARZ. The book itself has a thorough study of the CBS inequality and many related inequalities.

The Norm

XX.11.2.3. If $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ is an inner product space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, for $x \in \mathcal{V}$ define

$$\|x\| = \sqrt{\langle x, x\rangle} \ .$$

Whenever we refer to a norm on an inner product space, we mean this induced norm unless otherwise specified.

With this notation, the CBS inequality can be rephrased:

$$|\langle x, y\rangle| \leq \|x\|\|y\|$$

for $x, y \in \mathcal{V}$. A useful consequence is:

XX.11.2.4. PROPOSITION. Let $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ be an inner product space, and $x \in \mathcal{V}$. Then

$$\|x\| = \max_{\|y\|=1} |\langle x, y\rangle| = \sup_{\|y\|=1} |\langle x, y\rangle| \ .$$

In particular, if $\langle x, y\rangle = 0$ for all y with $\|y\| = 1$, then $x = 0$.

PROOF: We have $\|x\| \geq |\langle x, y\rangle|$ for $\|y\| = 1$ by the CBS inequality, and equality is obtained for $y = \frac{x}{\|x\|}$ if $x \neq 0$ and for any y if $x = 0$. 👍

XX.11.2.5. PROPOSITION. $\|\cdot\|$ is a norm on $\mathcal{V}$.

PROOF: It is clear that $\|x\| \geq 0$ for all x and $\|x\| = 0$ if and only if $x = 0$. We have

$$\|\alpha x\| = \sqrt{\langle \alpha x, \alpha x\rangle} = \sqrt{\alpha\bar{\alpha}\langle x, x\rangle} = |\alpha|\sqrt{\langle x, x\rangle} = |\alpha|\|x\|$$

for $x \in \mathcal{V}$, $\alpha \in \mathbb{F}$ since $|\alpha| = \sqrt{\alpha\bar{\alpha}}$. For the triangle inequality,

$$\|x+y\|^2 = \langle x+y, x+y\rangle = \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2$$

for $x, y \in \mathcal{V}$. 👍

The Parallelogram Law and Polarization Identity

The norm on an inner product space satisfies the following geometric property:

XX.11.2.6. Proposition. [Parallelogram Law] Let $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ be an inner product space (real or complex), and $x, y \in \mathcal{V}$. Then

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2) \ .$$

Proof:

$$\|x+y\|^2 + \|x-y\|^2 = \langle x+y, x+y\rangle + \langle x-y, x-y\rangle$$
$$= \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle + \langle x,x\rangle - \langle x,y\rangle - \langle y,x\rangle + \langle y,y\rangle$$
$$= 2\langle x,x\rangle + 2\langle y,y\rangle = 2(\|x\|^2 + \|y\|^2) \ .$$

👍

XX.11.2.7. "Most" norms on vector spaces do not satisfy the Parallelogram Law and therefore do not come from inner products. For example, on $\mathbb{R}^2$, the norm $\|\cdot\|_2$ comes from the standard inner product; but neither $\|\cdot\|_1$ or $\|\cdot\|_\infty$ come from an inner product: since $\|\boldsymbol{e}_1\|_1 = \|\boldsymbol{e}_2\|_1 = 1$ but $\|\boldsymbol{e}_1 + \boldsymbol{e}_2\|_1 = \|\boldsymbol{e}_1 - \boldsymbol{e}_2\|_1 = 2$, the Parallelogram Law is not satisfied for $\|\cdot\|_1$, and similarly for $\|\cdot\|_\infty$.

In fact, we have a converse to XX.11.2.6., sometimes called the *Fréchet-von Neumann-Jordan Theorem*:

XX.11.2.8. Proposition. [Polarization Identity] (i) Let $\mathcal{V}$ be a real vector space, and $\|\cdot\|$ a norm on $\mathcal{V}$. Then $\|\cdot\|$ is induced from a (real) inner product if and only if it satisfies the Parallelogram Law, and the inner product giving the norm is unique and satisfies

$$\langle x,y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$$

for $x, y \in \mathcal{V}$.

(ii) Let $\mathcal{V}$ be a complex vector space, and $\|\cdot\|$ a norm on $\mathcal{V}$. Then $\|\cdot\|$ is induced from an inner product if and only if it satisfies the Parallelogram Law, and the inner product giving the norm is unique and satisfies

$$\langle x,y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2 + i\|x+iy\|^2 - i\|x-iy\|^2\right)$$

for $x, y \in \mathcal{V}$.

PROOF: We write the proof of (ii); the proof of (i) is nearly identical but easier. If the norm comes from an inner product, then a straightforward argument similar to the proof of XX.11.2.6. (Exercise XX.11.11.9.) shows that the inner product satisfies the Polarization Identity and is thus unique. Conversely, if the "inner product" defined by the Polarization Identity is a true inner product, then for $x \in \mathcal{V}$

$$\langle x, x\rangle = \frac{1}{4}\left(\|2x\|^2 - \|0\|^2 + i\|(1+i)x\|^2 - i\|(1-i)x\|^2\right) = \|x\|^2$$

since $|1+i| = |1-i|$; so the norm it defines is the given one, which therefore satisfies the Parallelogram Law. Thus it suffices to show that if the norm satisfies the Parallelogram Law, the Polarization Identity defines an inner product on $\mathcal{V}$.

So suppose the norm satisfies the Parallelogram Law, and define $\langle\cdot,\cdot\rangle$ by the Polarization Identity. A simple straightforward calculation (not using the Parallelogram Law) shows that $\langle y, x\rangle = \overline{\langle x, y\rangle}$ and $\langle x, iy\rangle = -i\langle x, y\rangle$ for any $x, y \in \mathcal{V}$, and hence $\langle ix, y\rangle = i\langle x, y\rangle$ for $x, y \in \mathcal{V}$. Next, for $x, y, z \in \mathcal{V}$,

$$\|x + y + 2z\|^2 + \|x - y\|^2 = \|(x+z) + (y+z)\|^2 + \|(x+z) - (y+z)\|^2 = 2(\|x+z\|^2 + \|y+z\|^2)$$

by the Parallelogram Law, and similarly

$$\|x + y - 2z\|^2 + \|x - y\|^2 = 2(\|x-z\|^2 + \|y-z\|^2)$$

$$\|x + y + 2iz\|^2 + \|x - y\|^2 = 2(\|x+iz\|^2 + \|y+iz\|^2)$$

$$\|x + y - 2iz\|^2 + \|x - y\|^2 = 2(\|x-iz\|^2 + \|y-iz\|^2)$$

so we have

$$4\langle x+y, 2z\rangle = \|x+y+2z\|^2 - \|x+y-2z\|^2 + i\|x+y+2iz\|^2 - i\|x+y-2iz\|^2$$

$$= 2\|x+z\|^2+2\|y+z\|^2-2\|x-z\|^2-2\|y-z\|^2+2i\|x+iz\|^2+2i\|y+iz\|^2-2i\|x-iz\|^2-2i\|y-iz\|^2 = 8(\langle x, z\rangle+\langle y, z\rangle)$$

$$\langle x+y, 2z\rangle = 2(\langle x, z\rangle + \langle y, z\rangle)$$

and in the special case $y = 0$,

$$\langle x, 2z\rangle = 2\langle x, z\rangle$$

and therefore, for general x, y, z,

$$\langle x+y, z\rangle = \frac{1}{2}\langle x+y, 2z\rangle = \langle x, z\rangle + \langle y, z\rangle$$

and $\langle\cdot,\cdot\rangle$ is additive in each variable. We have also

$$\langle 2x, y\rangle = 2\langle x, y\rangle$$

for any x, y and we get by iteration that

$$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$$

for all $x, y \in \mathcal{V}$ and every dyadic rational number α. Since

$$\langle \alpha x, y\rangle = \frac{1}{4}\left(\|\alpha x + y\|^2 - \|\alpha x - y\|^2 + i\|\alpha x + iy\|^2 - i\|\alpha x - iy\|^2\right)$$

is clearly a continuous function of α for fixed x, y, we have

$$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$$

for all $\alpha \in \mathbb{R}$. Since it is also true for $\alpha = i$, it holds for all $\alpha \in \mathbb{C}$ and $\langle\cdot,\cdot\rangle$ is an inner product. 👍

Thus the norm and inner product completely determine each other. One way of phrasing this is:

XX.11.2.9. Corollary. Let $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ and $(\mathcal{W}, (\cdot,\cdot))$ be inner product spaces over the same field, and $T : \mathcal{V} \to \mathcal{W}$ a linear transformation. Then T is an isometry (i.e. $\|Tx\| = \|x\|$ for all $x \in \mathcal{V}$) if and only if T preserves inner products (i.e. $(Tx, Ty) = \langle x, y\rangle$ for all $x, y \in \mathcal{V}$).

Continuity of the Inner Product

XX.11.2.10. The induced norm on an inner product space $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ defines a topology on $\mathcal{V}$, called the *induced topology* on $\mathcal{V}$. We will always implicitly use this topology on $\mathcal{V}$ unless otherwise specified.

The next result is also an immediate consequence of the CBS inequality:

XX.11.2.11. Proposition. Let $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ be an inner product space. Then the inner product is *jointly continuous*: if (x_n) and (y_n) are sequences in $\mathcal{V}$, $x, y \in \mathcal{V}$, and $x_n \to x$, $y_n \to y$, then

$$\langle x_n, y_n\rangle \to \langle x, y\rangle \ .$$

In particular, the inner product is *separately continuous*:

$$\langle x_n, y\rangle \to \langle x, y\rangle \text{ and } \langle x, y_n\rangle \to \langle x, y\rangle \ .$$

Proof: We have

$$\begin{aligned} |\langle x_n, y_n\rangle - \langle x, y\rangle| &\le |\langle x_n, y_n\rangle - \langle x_n, y\rangle| + |\langle x_n, y\rangle - \langle x, y\rangle| \\ = |\langle x_n, y_n - y\rangle| + |\langle x_n - x, y\rangle| &\le \|x_n\|\|y_n - y\| + \|x_n - x\|\|y\| \ . \end{aligned}$$

Since (x_n) converges, it is bounded; hence both terms go to 0 as $n \to \infty$. 👍

XX.11.3. Completeness and Hilbert Spaces

XX.11.3.1. Definition. A *(complex) Hilbert space* is a complex inner product space which is complete with respect to the induced norm. A *real Hilbert space* is a real inner product space which is complete with respect to the induced norm. In light of XX.11.2.8., we can also say, by slight abuse of terminology, that a Hilbert space is a Banach space whose norm satisfies the Parallelogram Law.

In this section, the term "Hilbert space" will denote either a real or complex Hilbert space. But in the rest of this volume, the term "Hilbert space," unless qualified with "real," will denote a *complex* Hilbert space. The letter $\mathcal{H}$ will generally be used to denote a Hilbert space.

Note that specification of an inner product (or, equivalently, a norm satisfying the Parallelogram Law) is part of the definition of a Hilbert space. But we will frequently employ the usual abuse of terminology by saying "Let $\mathcal{H}$ be a Hilbert space."

XX.11.3.2. EXAMPLES. (i) $\mathbb{C}^n$ is a Hilbert space (with respect to the standard inner product) for any n. $\mathbb{R}^n$ is a real Hilbert space.

(ii) $C([a,b])$ is not a Hilbert space with respect to the inner product of XX.11.1.8.(ii) since it is not complete (Exercise XX.11.11.10.).

(iii) The sequence space ℓ^2 of square-summable sequences of complex numbers is a Hilbert space with respect to the inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{k=1}^{\infty} x_k \overline{y_k}$$

where $\boldsymbol{x} = (x_1, x_2, \dots)$, $\boldsymbol{y} = (y_1, y_2, \dots)$ ().

(iv) More generally, $L^2(X, \mathcal{A}, \mu)$, the space of square-integrable functions on a measure space $(X, \mathcal{A}, \mu)$ (or, more precisely, of equivalence classes of square-integrable functions, with functions agreeing almost everywhere identified), with inner product

$$\langle f, g \rangle = \int_X f \overline{g} \, d\mu$$

is a Hilbert space (). If S is a set, let μ be counting measure on S, and denote $L^2(S, \mu)$ by $\ell^2(S)$. Then $\ell^2 = \ell^2(\mathbb{N})$.

XX.11.3.3. The definition and basic properties of ℓ^2 were given in 1906 by D. HILBERT, who was the first to describe something approaching the modern notion of a "Hilbert space," and developed by E. SCHMIDT, F. RIESZ, and others in immediately succeeding years. The name "Hilbert space" (apparently referring to what we today call ℓ^2) is generally attributed to SCHMIDT, but according to [Jah03, p. 385] was first used in 1908 by A. SCHÖNFLIES. The definition of an abstract Hilbert space was, however, not given until 1928 by J. VON NEUMANN [vN30].

Completion of an Inner Product Space

Just as with metric spaces () and normed vector spaces (), every inner product space can be completed to a Hilbert space:

XX.11.3.4. THEOREM. Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be a (real or complex) inner product space. Then $\mathcal{V}$ can be isometrically embedded, in a unique way up to isometric isomorphism, as a dense subspace of a (real or complex) Hilbert space, called the *completion* of $\mathcal{V}$.

PROOF: This theorem can be proved directly in the same way as () and (); see Exercise XX.11.11.12.. But there is a much easier argument using XX.11.2.8.: if $\mathcal{H}$ is the completion of $\mathcal{V}$ with respect to the induced norm, then the norm on $\mathcal{H}$ satisfies the Parallelogram Law on a dense subspace, hence by continuity on the whole space. Thus it comes from an inner product which must extend the inner product on $\mathcal{V}$ by Polarization. 👍

XX.11.3.5. As usual, there is often a concrete realization of the completion of an inner product space: if it can be identified with a subspace of a Hilbert space, the closure of this subspace is the completion.

Pre-Hilbert Spaces

XX.11.3.6. DEFINITION. A *pre-Hilbert space* is a complex vector space with a pre-inner product.

Real pre-Hilbert spaces can be defined accordingly, but we will stick to the complex case.

Any (complex) inner product space is a pre-Hilbert space. But there are other natural ones. Here is an example of a very common kind:

XX.11.3.7. EXAMPLE. Let $\mathcal{PC}([a,b])$ be the set of piecewise-continuous complex-valued functions on $[a,b]$. Define a sesquilinear form on $\mathcal{PC}([a,b])$ by

$$\langle f,g\rangle = \int_a^b f(t)\overline{g(t)}\,dt\ .$$

This is a pre-inner product, but not an inner product: if f is zero except at finitely many points, then $\langle f,f\rangle = 0$, but f is not the zero element of $\mathcal{PC}([a,b])$.

XX.11.3.8. PROPOSITION. Let $(\mathcal{V},\langle\cdot,\cdot\rangle)$ be a pre-Hilbert space. Then

(i) If

$$\mathcal{N} = \{x\in\mathcal{V} : \langle x,x\rangle = 0\}$$

then $\mathcal{N}$ is a closed subspace of $\mathcal{V}$.

(ii) The pre-inner product on $\mathcal{V}$ induces an inner product on $\mathcal{V}/\mathcal{N}$ by

$$\langle [x],[y]\rangle = \langle x,y\rangle$$

for $x,y\in\mathcal{V}$.

PROOF: This is more or less a special case of (), and is proved in a similar way. 👍

XX.11.3.9. EXAMPLE. Let $\mathcal{PC}([a,b])$ be as in XX.11.3.7.. Then $\mathcal{N}$ consists of all functions which are zero except at finitely many points. If f is piecewise-continuous, the equivalence class $[f]$ of f consists precisely of all (necessarily piecewise-continuous) functions which agree with f except at finitely many points. Let $PC([a,b])$ be the set of equivalence classes of piecewise-continuous functions. Then the pre-inner product of XX.11.3.7. gives a well-defined inner product on $PC([a,b])$. We often think of the elements of $PC([a,b])$ as functions on $[a,b]$, but they are technically equivalence classes of functions; but the distinction can often be overlooked without essential difficulty. We can naturally identify $C([a,b])$ with a (dense) subspace of $PC([a,b])$.

XX.11.3.10. DEFINITION. Let $(\mathcal{V},\langle\cdot,\cdot\rangle)$ be a pre-Hilbert space. The *completion* of $\mathcal{V}$ is the Hilbert space obtained by completing $\mathcal{V}/\mathcal{N}$.

XX.11.4. Best Approximation

Hilbert spaces have a nice property not shared by general Banach spaces which is important in applications. If A is a nonempty closed convex set in a Hilbert space $\mathcal{H}$ (e.g. a closed subspace), and $x \in \mathcal{H}$, then there is a unique vector $y \in A$ which is closest to x. This theorem was proved by Beppo Levi in 1906 [?].

We use the notation $\rho(x, y) = \|x - y\|$, and

$$\rho(x, A) = \inf_{y \in A} \rho(x, y) = \inf_{y \in A} \|x - y\| \ .$$

XX.11.4.1. Theorem. Let $\mathcal{H}$ be a Hilbert space and A a nonempty closed convex subset of $\mathcal{H}$. For any $x \in \mathcal{H}$, there is a unique $y \in A$ such that

$$\|x - y\| = \rho(x, y) = \rho(x, A) = \inf_{z \in A} \|x - z\| = \min_{z \in A} \|x - z\| \ .$$

We need a lemma which is an immediate consequence of the Parallelogram Law:

XX.11.4.2. Lemma. Let A be a nonempty convex set in an inner product space $\mathcal{V}$, $x \in \mathcal{V}$, and $\epsilon > 0$. If $y, z \in A$ with $\|x - y\|^2 < \rho(x, A)^2 + \epsilon$ and $\|x - z\|^2 < \rho(x, A)^2 + \epsilon$, then $\|y - z\|^2 < 4\epsilon$.

Proof: Set $w = \frac{1}{2}(y + z)$. Then $w \in A$ since A is convex, so $\|x - w\| \geq \rho(x, A)$. We have

$$x - w + \frac{1}{2}(y - z) = x - z$$

$$x - w - \frac{1}{2}(y - z) = x - y$$

so by the Parallelogram Law

$$\|x - z\|^2 + \|x - y\|^2 = \left\|x - w + \frac{1}{2}(y - z)\right\|^2 + \left\|x - w - \frac{1}{2}(y - z)\right\|^2 = 2\left(\|x - w\|^2 + \left\|\frac{1}{2}(y - z)\right\|^2\right)$$

$$\|y - z\|^2 = 2\|x - z\|^2 + 2\|x - y\|^2 - 4\|x - w\|^2 < 2\rho(x, A)^2 + 2\epsilon + 2\rho(x, A)^2 + 2\epsilon - 4\rho(x, A)^2 = 4\epsilon \ .$$



We now prove Theorem XX.11.4.1..

Proof: Let (y_n) be a sequence in A with $\|x - y_n\| \to \rho(x, A)$. Then Lemma XX.11.4.2. implies that (y_n) is a Cauchy sequence in $\mathcal{H}$. Let $y = \lim y_n$. Then $y \in A$ since A is closed, and

$$\|x - y\| = \lim_{n \to \infty} \|x - y_n\| = \rho(x, A) \ .$$

If $z \in A$ with $\|x - z\| = \rho(x, A)$, then the hypotheses of Lemma XX.11.4.2. are satisfied for every $\epsilon > 0$, and we conclude $y = z$. 

XX.11.4.3. Lemma XX.11.4.2. really says that an inner product space is uniformly convex (XX.11.11.15.). The argument above shows that the closest approximation theorem holds in any uniformly convex Banach space.

XX.11.4.4. Examples. (i) The closest approximation result does not hold in general Banach spaces. In $(\mathbb{R}^2, \|\cdot\|_\infty)$, if A is the closed unit ball and $x = (2, 0)$, then there is a closest vector in A to x (there must be since A is compact), but it is not unique; in fact, any vector $(1, y)$ with $|y| \leq 1$ is a closest approximation to x in A.

(ii) The proof of XX.11.4.1. uses completeness. The result fails in general in an inner product space which is not complete. For example, let $\mathcal{V} = C([0, 1])$ with the standard inner product, and let

$$A = \left\{ f \in C([0,1]) : \int_0^{1/2} f(t)\,dt - \int_{1/2}^1 f(t)\,dt = 1 \right\} .$$

Then it is easily checked that A is a closed convex set in $\mathcal{V}$. But there is no closest vector to 0 (vector of minimum norm) in A: $\rho(0, A) = 1$, but there is no f in A of norm 1. (In the completion, the unique closest vector in $\bar{A}$ to 0 is the function which is 1 on $\left[0, \frac{1}{2}\right)$ and -1 on $\left(\frac{1}{2}, 1\right]$.)

XX.11.5. Orthogonality

Hilbert spaces have a geometric structure similar to that of Euclidean space. The most important geometric notion is *orthogonality*:

XX.11.5.1. Definition. Let $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ be an inner product space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. If $x, y \in \mathcal{V}$, then x and y are *orthogonal* (with respect to $\langle\cdot,\cdot\rangle$), written $x \perp y$, if $\langle x, y\rangle = 0$.

XX.11.5.2. The geometric term *perpendicular* is sometimes used as a synonym for *orthogonal*. But in functional analysis we prefer to use *orthogonal*, especially for complex inner product spaces. Note that in a complex inner product space, "perpendicular" and "orthogonal" do not quite have the same meaning: for example, in the one-dimensional (complex) Hilbert space $\mathbb{C}^1$ with standard inner product, the vectors 1 and i are "perpendicular" but not orthogonal.

We have $0 \perp y$ for every y. Conversely, we have a useful rephrasing of the simple result XX.11.2.4.:

XX.11.5.3. Proposition. Let $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ be an inner product space, and $x \in \mathcal{V}$. Then $x \perp y$ for all $y \in \mathcal{V}$ if and only if $x = 0$.

Orthogonal Complement

XX.11.5.4. Definition. Let $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ be an inner product space, and $S \subseteq \mathcal{V}$. The *orthogonal complement* of S (in $\mathcal{V}$) is

$$S^\perp = \{x \in \mathcal{V} : \langle x, y\rangle = 0 \text{ for all } y \in S\}.$$

$S^\perp$ is a closed subspace of $\mathcal{V}$ (whether or not S is a subspace; in fact, $S^\perp = (span(S))^\perp$).

The next norm relation is fundamental.

XX.11.5.5. Proposition. [Pythagorean Theorem] Let $\mathcal{V}$ be an inner product space, and $x, y \in \mathcal{V}$ with $x \perp y$. Then

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2 \ .$$

Proof: Expand $\langle x+y, x+y\rangle$. 👍

The converse is true in a real inner product space, but not a complex one (take 1 and i in $\mathbb{C}^1$). But there is a partial converse:

XX.11.5.6. Proposition. Let $\mathcal{V}$ be an inner product space over $\mathbb{F}$, and $x, y \in \mathcal{V}$. The following are equivalent:

(i) $x \perp y$.

(ii) $\|x+\alpha y\|^2 = \|x\|^2 + |\alpha|^2\|y\|^2$ for all $\alpha \in \mathbb{F}$.

(iii) $\|x+\alpha y\| \geq \|x\|$ for all $\alpha \in \mathbb{F}$.

Proof: (i) $\Rightarrow$ (ii) is XX.11.5.5., and (ii) $\Rightarrow$ (iii) is trivial. For (iii) $\Rightarrow$ (i), we may assume $y \neq 0$. Set

$$\alpha = -\frac{\langle x, y\rangle}{\langle y, y\rangle}$$

and then a calculation as in the proof of XX.11.2.1. shows that

$$\|x+\alpha y\|^2 = \|x\|^2 - \frac{|\langle x, y\rangle|}{\|y\|^2}$$

which is only $\geq \|x\|^2$ if $\langle x, y\rangle = 0$. 👍

If $\mathcal{H}$ is a Hilbert space, $\mathcal{Y}$ is a closed subspace of $\mathcal{H}$, and $x \in \mathcal{H}$, then there is a closest vector $y \in \mathcal{Y}$ to X. It turns out that $z = x - y \in \mathcal{Y}^\perp$:

XX.11.5.7. Proposition. Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{Y}$ a closed subspace of $\mathcal{H}$. Then every $x \in \mathcal{H}$ can be uniquely written as $y + z$, where $y \in \mathcal{Y}$ and $z \in \mathcal{Y}^\perp$. The y is the closest vector in $\mathcal{Y}$ to x.

Proof: By XX.11.4.1. there is a (unique) closest vector $y \in \mathcal{Y}$ to x. Set $z = x - y$. Let $w \in \mathcal{Y}$. Then, for any $\alpha \in \mathbb{C}$,

$$\|z + \alpha w\| = \|x - (y - \alpha w)\| \geq \|x - y\| = \|z\|$$

since $y - \alpha w \in \mathcal{Y}$. Thus $z \perp w$ by XX.11.5.6.. If also $x = y' + z'$ with $y' \in \mathcal{Y}$, $z' \in \mathcal{Y}^\perp$, we have $y + z - y' - z' = 0$, so

$$y - y' = z' - z \in \mathcal{Y} \cap \mathcal{Y}^\perp = \{0\}$$

and $y = y'$, $z = z'$. 👍

XX.11.5.8. Example. Completeness is essential in XX.11.5.7.. As a slight variation of Example XX.11.4.4., let

$$\mathcal{Y} = \left\{ f \in C([0,1]) : \int_0^{1/2} f(t)\,dt - \int_{1/2}^1 f(t)\,dt = 0 \right\} .$$

Then $\mathcal{Y}$ is a closed subspace of $C([0,1])$ (with respect to the norm induced by the standard inner product). But if f is the constant function 1, there is no decomposition $f = g + h$ with $g \in \mathcal{Y}$ and $h \in \mathcal{Y}^\perp$. In fact, $\mathcal{Y}^\perp = \{0\}$ in $C([0,1])$ even though $\mathcal{Y}$ is a proper subspace of $C([0,1])$.

XX.11.5.9. If $\mathcal{V}$ is an inner product space and $S \subseteq \mathcal{V}$, then $S^{\perp\perp} = (S^\perp)^\perp$ is a closed subspace of $\mathcal{V}$ containing S. If $\mathcal{V}$ is complete, then $S^{\perp\perp}$ is the *smallest* closed subspace containing S:

XX.11.5.10. Proposition. Let $\mathcal{H}$ be a Hilbert space and $S \subseteq \mathcal{H}$. Then $S^{\perp\perp}$ is the closed linear span of S.

Proof: Let $\mathcal{Y}$ be the closed linear span of S, i.e. the closure of $span(S)$. Then $\mathcal{Y} \subseteq S^{\perp\perp}$, and $\mathcal{Y}^\perp = S^\perp$. If $x \in S^{\perp\perp}$, then $x = y + z$ with $y \in \mathcal{Y}$ and $z \in \mathcal{Y}^\perp$. But since $x, y \in S^{\perp\perp}$, $z = x - y \in S^{\perp\perp} = \mathcal{Y}^{\perp\perp}$. Thus

$$z \in \mathcal{Y}^\perp \cap \mathcal{Y}^{\perp\perp} = \{0\}$$

so $x = y \in \mathcal{Y}$, i.e. $\mathcal{Y} = S^{\perp\perp}$. 👍

XX.11.5.11. Corollary. Let $\mathcal{H}$ be a Hilbert space and $\mathcal{Y}$ a closed subspace. Then $\mathcal{Y} = \mathcal{Y}^{\perp\perp}$. In particular, $\mathcal{Y} = \mathcal{H}$ if and only if $\mathcal{Y}^\perp = \{0\}$, so if $S \subseteq \mathcal{H}$, then the closed linear span of S is $\mathcal{H}$ if and only if $S^\perp = \{0\}$.

This result can fail if the space is not complete (Example XX.11.5.8.).

Orthonormal Sets

XX.11.5.12. Definition. A set $\{e_j : j \in J\}$ of vectors in an inner product space is *orthonormal* if $\langle e_j, e_k \rangle = \delta_{jk}$ for all $j, k \in J$, i.e. it is a mutually orthogonal set of unit vectors.

Of course, any subset of an orthonormal set is orthonormal. A set is orthonormal if and only if every finite subset (in fact, every two-element subset) is orthonormal.

XX.11.5.13. Proposition. An orthonormal set in an inner product space is linearly independent.

Proof: Suppose $S = \{e_j : j \in J\}$ is an orthonormal set, and suppose

$$\alpha_1 e_{j_1} + \cdots + \alpha_n e_{j_n} = 0$$

for some $j_1, \ldots, j_n$ and scalars $\alpha_1, \ldots, \alpha_n$. Then for $1 \leq m \leq n$

$$0 = \langle \alpha_1 e_{j_1} + \cdots + \alpha_n e_{j_n}, e_{j_m} \rangle = \sum_{r=1}^{n} \alpha_r \langle e_{j_r}, e_{j_m} \rangle = \alpha_m$$

so S is linearly independent. 👍

Orthonormal sets are very nice to work with. If S is an orthonormal set, then it is easy to see how to explicitly write any vector in $span(S)$ as a convex combination of the vectors in S:

XX.11.5.14. PROPOSITION. Let $S = \{e_1, \ldots, e_n\}$ be a finite orthonormal set in an inner product space, and $x \in span(S)$. Then the unique representation of x as a linear combination of $\{e_1, \ldots, e_n\}$ is

$$x = \sum_{k=1}^{n} \langle x, e_k \rangle e_k \ .$$

We also have

$$\|x\|^2 = \sum_{k=1}^{n} |\langle x, e_k \rangle|^2 \ .$$

PROOF: Suppose

$$x = \sum_{k=1}^{n} \alpha_k e_k$$

is the unique expression of x as a linear combination of S, then, for each j,

$$\langle x, e_j \rangle = \sum_{k=1}^{n} \alpha_k \langle e_k, e_j \rangle = \alpha_j \ .$$

👍

The number $\langle x, e_j \rangle$ is often called the *Fourier coefficient* of x with respect to e_j (cf. ()).

This finite result has an important generalization to infinite orthonormal sets:

XX.11.5.15. THEOREM. Let $S = \{e_j : j \in J\}$ be an orthonormal set in an inner product space $\mathcal{V}$, and $\mathcal{Y}$ the closed linear span of S in $\mathcal{V}$. Then, for any $y \in \mathcal{Y}$:

(i) The sum

$$\sum_{j \in J} \langle y, e_j \rangle e_j$$

converges unconditionally to y in $\mathcal{V}$, i.e. the net of finite subsums converges to y.

(ii) The unordered sum

$$\sum_{j \in J} |\langle y, e_j \rangle|^2$$

converges (absolutely) to $\|y\|^2$, i.e. the net of finite subsums converges to $\|y\|^2$.

PROOF: Let $F = \{e_{j_1}, \ldots, e_{j_n}\}$ be a finite subset of S, and set

$$y_F = \sum_{k=1}^{n} \langle y, e_{j_k} \rangle e_{j_k}$$

and $z_F = y - y_F$. We have that $y_F \in span(F)$, and that

$$\langle z_F, e_{j_m} \rangle = \langle y, e_{j_m} \rangle - \langle y_F, e_{j_m} \rangle = 0$$

for $1 \leq m \leq n$; thus $z_F \in (span(F))^{\perp}$. So we have

$$\|y\|^2 = \|y_F\|^2 + \|z_F\|^2 \geq \|y_F\|^2 = \sum_{k=1}^{n} |\langle y, e_{j_k} \rangle|^2 \ .$$

Thus the sum in (ii) converges, and is $\leq \|y\|^2$.

On the other hand, if $\epsilon > 0$ there is an $x \in span(S)$ with $\|x - y\| < \epsilon$. We have $x \in span(E)$ for some finite subset E of S. If $F = \{e_{j_1}, \ldots, e_{j_n}\}$ is a finite subset of of S containing E, then

$$\|y - y_F\| \leq \|y - x\|$$

since y_F is the closest vector in $span(F)$ to y. Thus

$$\left\| y - \sum_{k=1}^{n} \langle y, e_{j_k} \rangle e_{j_k} \right\| < \epsilon$$

and the finite subsums in (i) converge to y. Also,

$$\sum_{k=1}^{n} |\langle y, e_{j_k} \rangle|^2 = \|y_F\|^2 \geq (\|y\| - \|y - y_F\|)^2 \geq (\|y\| - \epsilon)^2$$

which can be made arbitrarily close to $\|y\|^2$ by choosing ϵ sufficiently small. Thus the sum in (ii) converges to $\|y\|^2$. 👍

XX.11.5.16. COROLLARY. Let $\mathcal{H}$ be a Hilbert space, $S = \{e_j : j \in J\}$ an orthonormal set in $\mathcal{H}$, and $x \in \mathcal{H}$. Then

$$\sum_{j \in J} |\langle x, e_j \rangle|^2 \leq \|x\|^2$$

(*Bessel's inequality*). In particular, $\langle x, e_j \rangle \neq 0$ for only countably many j. We have

$$\sum_{j \in J} |\langle x, e_j \rangle|^2 = \|x\|^2$$

(*Parseval's relation*) if and only if x is in the closed linear span of S.

PROOF: Let $\mathcal{Y}$ be the closed linear span of S, and write $x = y + z$ with $y \in \mathcal{Y}$, $z \in \mathcal{Y}^{\perp}$. Then

$$\|x\|^2 = \|y\|^2 + \|z\|^2 \geq \|y\|^2 = \sum_{j \in J} |\langle y, e_j \rangle|^2 = \sum_{j \in J} |\langle x, e_j \rangle|^2$$

with equality if and only if $z = 0$. 👍

Orthonormal Bases

The next result is a combination of parts of XX.11.5.11. and XX.11.5.15.:

XX.11.5.17. COROLLARY. Let $\mathcal{H}$ be a Hilbert space, and $S = \{e_j : j \in J\}$ an orthonormal set in $\mathcal{H}$. The following are equivalent:

(i) S is a maximal orthonormal set (not contained in any strictly larger orthonormal set).

(ii) $S^\perp = \{0\}$.

(iii) The closed linear span of S is $\mathcal{H}$.

(iv) For every $x \in \mathcal{H}$, there is a unique unordered sum

$$\sum_{j\in J} \alpha_j e_j$$

which converges to x. We necessarily have that the unordered sum

$$\sum_{j\in J} |\alpha_j|^2$$

converges to $\|x\|^2$, and $\alpha_j = \langle x, e_j\rangle$.

PROOF: For the equivalence of (i) and (ii), note that if S' is an orthonormal set containing S, then $S'\backslash S \subseteq S^\perp$. Conversely, if $S^\perp \neq \{0\}$, let x be a unit vector in $S^\perp$; then $S \cup \{x\}$ is an orthonormal set strictly larger than S.

For the uniqueness in (iv), suppose the unordered sum $\sum_j \alpha_j e_j$ converges to x. Then, by linearity and continuity of the inner product, we have, for each k,

$$\langle x, e_k\rangle = \lim_{F\subseteq S} \sum_{j\in F} \langle \alpha_j e_j, e_k\rangle = \alpha_k$$

where the limit is over the net of finite subsets F of S. 👍

XX.11.5.18. DEFINITION. If $\mathcal{H}$ is a Hilbert space, an *orthonormal basis* for $\mathcal{H}$ is an orthonormal set S in $\mathcal{H}$ satisfying the equivalent conditions of XX.11.5.17..

XX.11.5.19. An orthonormal basis is linearly independent (XX.11.5.13.). If $\mathcal{H}$ is finite-dimensional, then an orthonormal basis for $\mathcal{H}$ is finite, hence a Hamel basis. But an orthonormal basis for an infinite-dimensional Hilbert space is not a Hamel basis. For example, if $\{e_{j_1}, e_{j_2}, \dots\}$ is a sequence of distinct elements of an orthonormal basis for $\mathcal{H}$, then the series

$$\sum_{n=1}^{\infty} \frac{1}{n} e_{j_n}$$

converges in $\mathcal{H}$ to a vector which cannot be written as a finite linear combination of the e_j by the uniqueness in XX.11.5.17.. A countable orthonormal basis for a Hilbert space is an unconditional Schauder basis ().

XX.11.5.20. As with Hamel bases in vector spaces (), there is nothing remotely unique about orthonormal bases for Hilbert spaces. For example, any two perpendicular unit vectors in $\mathbb{R}^2$ form an orthonormal basis for $\mathbb{R}^2$ or $\mathbb{C}^2$ (with respect to the usual inner product). $\mathbb{C}^2$ has many other orthonormal bases not contained in the subset $\mathbb{R}^2$.

If $\mathcal{H}$ is a complex Hilbert space, whether an orthonormal set is an orthonormal basis depends whether $\mathcal{H}$ is regarded as a complex Hilbert space or as a real Hilbert space with inner product $(\cdot,\cdot) = Re\langle\cdot,\cdot\rangle$. If B is an orthonormal basis for $\mathcal{H}$ as a complex Hilbert space, it is an orthonormal set, but not an orthonormal basis, for $\mathcal{H}$ as a real Hilbert space; the set

$$B \cup iB = \{e, ie : e \in B\}$$

is an orthonormal basis for $\mathcal{H}$ as a real Hilbert space.

We have a companion result slightly generalizing XX.11.5.17., which is an immediate corollary of XX.11.5.17., the continuity of the inner product, and the CBS inequality:

XX.11.5.21. Corollary. Let $\mathcal{H}$ be a Hilbert space, and $\{e_j : j \in J\}$ an orthonormal basis for $\mathcal{H}$. Let $x, y \in \mathcal{H}$. Write

$$x = \sum_{j\in J} \alpha_j e_j \ , \quad y = \sum_{j\in J} \beta_j e_j$$

as in XX.11.5.17.. Then the unordered sum

$$\sum_{j\in J} \alpha_j \overline{\beta_j}$$

converges absolutely to $\langle x, y\rangle$.

The next theorem is the main theoretical result about orthonormal bases:

XX.11.5.22. Theorem. Let $\mathcal{H}$ be a Hilbert space. Then

(i) $\mathcal{H}$ has an orthonormal basis.

(ii) Every orthonormal set in $\mathcal{H}$ is contained in an orthonormal basis.

(iii) Any two orthonormal bases for $\mathcal{H}$ have the same cardinality.

Proof: The proof of (i) is a prototype Zorn's Lemma argument (using the characterization of an orthonormal basis as a maximal orthonormal set), similar to the proof in III.11.2.32. (but easier), using XX.11.5.11., and is left to the reader. Part (ii) can be proved in a similar way: if S is an orthonormal set in $\mathcal{H}$, consider only orthonormal sets containing S. Alternately, let $\mathcal{Y}$ be the closed linear span of S. Then $\mathcal{Y}^\perp$ is a closed subspace of $\mathcal{H}$ and thus itself a Hilbert space, so has an orthonormal basis B'. Then $B = S \cup B'$ is an orthonormal basis for $\mathcal{H}$ containing S.

To prove (iii), first note that if $\mathcal{H}$ has a finite orthonormal basis B with n elements, then B is a Hamel basis for $\mathcal{H}$ and thus $\mathcal{H}$ is an n-dimensional vector space. Thus any other orthonormal basis for $\mathcal{H}$ also has exactly n elements. Now assume that one (hence every) orthonormal basis for $\mathcal{H}$ is infinite. Let $\{e_j : j \in J\}$ be an orthonormal basis for $\mathcal{H}$ of cardinality λ, and $\{f_k : k \in K\}$ another orthonormal basis for $\mathcal{H}$ of cardinality κ. Then each e_j is in the closed linear span of countably many f_k (). Every f_k occurs in one of

these, since otherwise $e_j \in \{f_k\}^\perp$ for all j. Thus K is a union of λ countable sets, so $\kappa \leq \aleph_0 \cdot \lambda$. Similarly, $\lambda \leq \aleph_0 \cdot \kappa$. Since λ and κ are infinite, $\lambda = \kappa$ (II.9.9.4.). 👍

Note that the AC is needed for this proof; in fact, (i) and (ii) may be equivalent to the AC. (The full AC is not needed for (iii), but some weaker form is (II.9.9.4.).)

XX.11.5.23. DEFINITION. Let $\mathcal{H}$ be a Hilbert space over $\mathbb{F}$ ($\mathbb{R}$ or $\mathbb{C}$). The *orthogonal dimension* (or *Hilbert dimension*) of $\mathcal{H}$ (over $\mathbb{F}$) is the cardinality of some (hence every) orthonormal basis for $\mathcal{H}$.

XX.11.5.24. The orthogonal dimension of $\mathcal{H}$ is finite if and only if $\mathcal{H}$ is a finite-dimensional vector space, in which case the orthogonal dimension is the same as the vector space dimension. The two dimensions do not coincide for infinite-dimensional Hilbert spaces in general (). We usually only consider the orthogonal dimension for a Hilbert space $\mathcal{H}$; it is often just called the *dimension* of $\mathcal{H}$ and denoted $dim(\mathcal{H})$.

The orthogonal dimension of a finite-dimensional complex Hilbert space changes (is multiplied by 2) if it is regarded as a real Hilbert space with inner product $(\cdot,\cdot) = Re\langle\cdot,\cdot\rangle$. The orthogonal dimension of an infinite-dimensional complex Hilbert space does not change if it is regarded as a real Hilbert space (since $2\kappa = \kappa$ for any infinite cardinal κ).

The orthogonal dimension classifies Hilbert spaces up to isometric isomorphism; thus the orthogonal dimension is the *only* isomorphism invariant for Hilbert spaces:

XX.11.5.25. THEOREM. Let $\mathcal{H}$ and $\mathcal{H}'$ be Hilbert spaces over $\mathbb{F}$. Then $\mathcal{H}$ and $\mathcal{H}'$ are isometrically isomorphic (over $\mathbb{F}$) if and only if they have the same orthogonal dimension. More precisely, if $B = \{e_j : j \in J\}$ is an orthonormal basis for $\mathcal{H}$ and $B' = \{f_j : j \in J\}$ is an orthonormal basis for $\mathcal{H}'$, the map $e_j \mapsto f_j$ extends to a unique isometric isomorphism $T : \mathcal{H} \to \mathcal{H}'$.

PROOF: It is obvious from the uniqueness of orthogonal dimension that Hilbert spaces of different orthogonal dimensions cannot be isometrically isomorphic. So we need only prove the last statement. The map $T_0 : span(B) \to span(B')$ defined by

$$T_0\left(\sum_{j\in F} \alpha_j e_j\right) = \sum_{j\in F} \alpha_j f_j$$

(where F is an arbitrary finite subset of J) is a well-defined linear map since B is linearly independent. T_0 is an isometry by XX.11.5.14.. Thus T_0 extends uniquely to an isometric linear isomorphism T from $\mathcal{H}$ onto $\mathcal{H}'$. The extension satisfies

$$T\left(\sum_{j\in J} \alpha_j e_j\right) = \sum_{j\in J} \alpha_j f_j$$

by continuity (cf. XX.11.5.17.). 👍

Although the AC is not used in this proof, this theorem depends on the AC since we must know that $\mathcal{H}$ and $\mathcal{H}'$ have orthonormal bases. But if they are particular Hilbert spaces for which we know the existence of orthonormal bases of the same cardinality without using the AC (e.g. XX.11.5.29.), then the isomorphism also does not depend on the AC.

XX.11.5.26. For each cardinal κ, there is a Hilbert space whose orthogonal dimension is κ, namely $\ell^2(X)$ for a set X of cardinality κ (indicator functions of singleton subsets of X form an orthonormal basis of cardinality κ). Thus (assuming AC) there is a precise classification of Hilbert spaces up to isometric isomorphism: there is exactly one for each cardinal. (There is, of course, a separate classification for real Hilbert spaces and complex Hilbert spaces.) This classification is analogous to, but not exactly the same as, the classification up to isomorphism of vector spaces over a fixed field by their dimension.

Separable Hilbert spaces

XX.11.5.27. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space. Then $\mathcal{H}$ is separable if and only if its orthogonal dimension is countable.

PROOF: If $\mathcal{H}$ has a countable orthonormal basis, then it is the closed linear span of a countable set and thus separable (). Conversely, any orthonormal basis for a Hilbert space is discrete in the relative topology (the distance between any two orthogonal unit vectors is $\sqrt{2}$), and a separable metric space cannot have an uncountable discrete subset. 👍

This proof depends on the existence of orthonormal bases and thus on the AC. But we will give a constructive proof that every separable Hilbert space has a (countable) orthonormal basis without using even the countable AC; thus this result does not depend on any form of Choice.

XX.11.5.28. Suppose $\mathcal{H}$ is a separable Hilbert space. Fix a countable dense (or just total) set in $\mathcal{H}$ and arrange it into a sequence (x_n). We will describe an explicit inductive procedure to manufacture an orthonormal basis out of the x_n, called the *Gram-Schmidt orthogonalization procedure*.

To avoid annoying notational problems, we will first go through the sequence (x_n) to discard any x_n which is a linear combination of the previous ones. Reindexing, we may thus assume the x_n are linearly independent.

In particular, $x_1 \neq 0$. Set $e_1 = \frac{x_1}{\|x_1\|}$. Then e_1 is a unit vector spanning the same one-dimensional subspace as x_1.

Now set $y_2 = x_2 - \langle x_2, e_1\rangle e_1$. Then $y_2 \neq 0$ since $\{x_1, x_2\}$ is linearly independent and hence $\{e_1, x_2\}$ is linearly independent, and

$$\langle y_2, e_1\rangle = \langle x_2, e_1\rangle - \langle x_2, e_1\rangle\langle e_1, e_1\rangle = 0$$

So $y_2 \perp e_1$. Set $e_2 = \frac{y_2}{\|y_2\|}$. Then e_2 is a unit vector orthogonal to e_1, and $span(\{e_1, e_2\}) = span(\{x_1, x_2\})$.

Continue in this manner, i.e.

$$y_3 = x_3 - \langle x_3, e_1\rangle e_1 - \langle x_3, e_2\rangle e_2$$

is a nonzero vector orthogonal to e_1 and e_2 (it is just the component of x_3 in the orthogonal complement of $span(\{x_1, x_2\})$ in $span(\{x_1, x_2, x_3\})$), and if $e_3 = \frac{y_3}{\|y_3\|}$, then $\{e_1, e_2, e_3\}$ is an orthonormal set with the same span as $\{x_1, x_2, x_3\}$. In this way we generate an orthonormal sequence (e_n), with

$$span(\{e_1, \ldots, e_n\}) = span(\{x_1, \ldots, x_n\})$$

for all n. Since the sequence (x_n) is total in $\mathcal{H}$, so is the sequence (e_n), i.e. $\{e_n : n \in \mathbb{N}\}$ is an orthonormal basis for $\mathcal{H}$.

The result is important enough that we state it as a theorem:

XX.11.5.29. THEOREM. [GRAM-SCHMIDT] Let $\mathcal{H}$ be a separable Hilbert space, and S a countable total subset of $\mathcal{H}$. Then there is a countable orthonormal basis B for $\mathcal{H}$ with $span(B) = span(S)$. The orthonormal basis B is constructed from S by an explicit algorithm.

XX.11.5.30. Several things to note about this process:

(i) The orthonormal basis obtained depends not only on the x_n, but on the order in which they are taken. Even in a two-dimensional Hilbert space the orthonormal basis is quite different in general if the order of the original vectors is changed. The reader unfamiliar with the Gram-Schmidt procedure should try some two- and three-dimensional examples to understand the process.

(ii) Nowhere in the procedure did we use that $\mathcal{H}$ is complete. Thus the process works equally well in any separable inner product space, i.e. any dense subspace of a separable Hilbert space. The conclusion is that every dense subspace of a separable Hilbert space $\mathcal{H}$ contains an orthonormal basis for $\mathcal{H}$. This is useful when $\mathcal{H}$ contains a dense subspace of nice concrete elements; cf. XX.11.5.33..

(iii) No form of Choice is used in the Gram-Schmidt process.

We thus obtain the following result without any form of AC:

XX.11.5.31. COROLLARY. Two separable Hilbert spaces over $\mathbb{F}$ are isometrically isomorphic (over $\mathbb{F}$) if and only if they have the same orthogonal dimension (over $\mathbb{F}$), which is in $\mathbb{N}_0 \cup \{\aleph_0\}$. Any separable infinite-dimensional Hilbert space (over $\mathbb{F}$) is isometrically isomorphic (over $\mathbb{F}$) to ℓ^2; hence any two separable infinite-dimensional Hilbert spaces (over $\mathbb{F}$) are isometrically isomorphic (over $\mathbb{F}$).

XX.11.5.32. Topologists generally use the term "Hilbert space" to denote the unique separable infinite-dimensional Hilbert space (i.e. ℓ^2), regarded either as a metric space or as a topological space (where it is homeomorphic to $\mathbb{R}^{\mathbb{N}}$ with the product topology (), although there is no linear homeomorphism). Thus to a topologist there is only one "Hilbert space" (up to isometry), but to an analyst there are many.

XX.11.5.33. EXAMPLE. [LEGENDRE POLYNOMIALS] Let $\mathcal{H} = L^2([-1,1])$ (we can take this simply to be the completion of $C([-1,1])$, $PC([-1,1])$, or $R([-1,1])$; cf. (), and we can work over either $\mathbb{R}$ or $\mathbb{C}$). The algebra $\mathcal{P}$ of polynomials is uniformly dense in $C([-1,1])$ by the Stone-Weierstrass Theorem (); since the uniform topology is stronger than the topology defined by the standard inner product

$$\langle f, g\rangle = \int_{-1}^{1} f(t)\overline{g(t)}\, dt$$

we have that $\mathcal{P}$ is dense in $L^2([-1,1])$. Thus, if we take $f_n(t) = t^n$, then the sequence (f_n) $(n \geq 0)$ is total in $L^2([-1,1])$.

We make an orthonormal basis for $L^2([-1,1])$ by the Gram-Schmidt procedure. Since $\|f_0\| = \sqrt{2}$, e_0 is the constant function $\frac{1}{\sqrt{2}}$. We have that

$$g_1 := f_1 - \langle f_1, e_0\rangle e_0 = t - \frac{1}{\sqrt{2}} \int_{-1}^{1} \frac{t}{\sqrt{2}}\, dt = f_1$$

(actually $f_1 \perp e_0$, so e_1 is just a unit vector in the direction of f_1). Since

$$\|g_1\| = \|f_1\| = \left[\int_{-1}^{1} |t|^2 \, dt\right]^{1/2} = \sqrt{\frac{2}{3}}$$

we have that $e_1(t) = \sqrt{\frac{3}{2}}t$.

Continuing in this manner, we obtain

$$e_n(t) = \sqrt{\frac{2n+1}{2}} p_n(t)$$

where

$$p_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n}[(t^2-1)^n]$$

is the *n'th Legendre polynomial* (this expression is called *Rodrigues's formula*). The constants are chosen so that $p_n(1) = 1$ for all n. The polynomial p_n is the solution to the *Legendre differential equation*

$$(1-t^2)y'' - 2ty' + n(n+1)y = 0$$

with $y(1) = 1$. See e.g. [**?**, 3.7] for details, and for more examples of orthonormal bases of function spaces constructed by the Gram-Schmidt procedure.

Note that, as expected, we obtain an orthonormal basis for $L^2([-1,1])$ consisting of polynomials.

XX.11.5.34. The statement (XX.11.5.30.(ii)) that a dense subspace of a Hilbert space $\mathcal{H}$ contains an orthonormal basis for $\mathcal{H}$ can be false if $\mathcal{H}$ is nonseparable. In fact, a Hilbert space of orthogonal dimension $2^{\aleph_0}$ has a dense subspace which does not contain any uncountable orthonormal set [Dix53].

A Crinkled Curve

"There's a lot of room in Hilbert space."

Traditional

XX.11.5.35. In an infinite-dimensional Hilbert space, even a separable one, there are continuous arcs which are "crinkled" in a way impossible in a finite-dimensional Euclidean space: curves which make a right-angled turn at every point. Such a curve cannot be differentiable anywhere in any reasonable sense.

XX.11.5.36. Example. In $L^2([0,1])$ (we can avoid involving measure theory by working in the dense subspace $PC([0,1])$, cf. XX.11.3.9.), for $0 \le t \le 1$ let $f_t = \chi_{[0,t]}$, the indicator function of the subinterval $[0,t]$. It is easy to check that $t \mapsto f_t$ is a one-to-one continuous function from $[0,1]$ into $L^2([0,1])$. If $0 \le a < b < c \le 1$, then the chord $f_b - f_a$ from f_a to f_b is orthogonal to the chord $f_c - f_b$ from f_b to f_c. Thus any two consecutive chords (or indeed any two nonoverlapping chords) are orthogonal! In fact, the entire arcs from 0 to b and from b to 1 lie in orthogonal affine subspaces, i.e. the curve makes a right-angled turn at b.

See XX.11.11.13. for another sense in which "There's a lot of room in Hilbert space."

XX.11.6. Projections

The theory of subspaces and orthogonal complements in a Hilbert space has a nice alternate description in terms of orthogonal projection operators (projections).

XX.11.6.1. Definition. Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{Y}$ a closed subspace. For any vector $x \in \mathcal{H}$, write $x = y + z$ with $y \in \mathcal{Y}$ and $z \in \mathcal{Y}^\perp$, and define $P_\mathcal{Y}(x)$ to be y. Then $\mathcal{P}_\mathcal{Y}$ is a function from $\mathcal{H}$ to $\mathcal{H}$, with range $\mathcal{Y}$. $P_\mathcal{Y}$ is called the *orthogonal projection operator*, or just *projection*, of $\mathcal{H}$ onto $\mathcal{Y}$.

Note that $P_\mathcal{Y}(x)$ is well defined by the uniqueness of the decomposition $x = y + z$ (XX.11.5.7.). We have $P_{\{0\}} = 0$ and $P_\mathcal{H} = I$.

XX.11.6.2. Proposition. Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{Y}$ a closed subspace. Then $P_\mathcal{Y}$ is a bounded linear operator on $\mathcal{H}$ with operator norm $\|P_\mathcal{Y}\| = 1$ (unless $\mathcal{Y} = \{0\}$), and $\mathcal{N}(P_\mathcal{Y}) = \mathcal{Y}^\perp$.

Proof: If $x_1, x_2 \in \mathcal{H}$, and $x_1 = y_1 + z_1$, $x_2 = y_2 + z_2$ with $y_j \in \mathcal{Y}$, $z_j \in \mathcal{Y}^\perp$, then

$$x_1 + x_2 = (y_1 + y_2) + (z_1 + z_2)$$

must be the unique decomposition for $x_1 + x_2$, and thus

$$P_\mathcal{Y}(x_1 + x_2) = y_1 + y_2 = P_\mathcal{Y}(x_1) + P_\mathcal{Y}(x_2) \ .$$

Similarly, if $x \in \mathcal{H}$ and $\alpha \in \mathbb{F}$, and $x = y + z$ is the decomposition of x, then $\alpha x = \alpha y + \alpha z$ is the decomposition of αx, and thus

$$P_\mathcal{Y}(\alpha x) = \alpha y = \alpha P_\mathcal{Y}(x) \ .$$

Thus $P_\mathcal{Y}$ is a linear operator. If $x \in \mathcal{H}$, $x = y + z$, then by XX.11.5.5.

$$\|P_\mathcal{Y}(x)\|^2 = \|y\|^2 \leq \|y\|^2 + \|z\|^2 = \|x\|^2$$

so $P_\mathcal{Y}$ is a bounded operator and $\|P_\mathcal{Y}\| \leq 1$. But if $\mathcal{Y} \neq \{0\}$, and y is a nonzero vector in $\mathcal{Y}$, then $P_\mathcal{Y}(y) = y$, so $\|P_\mathcal{Y}\| = 1$.

It is obvious that $P_\mathcal{Y}(x) = 0$ if and only if $x \in \mathcal{Y}^\perp$. 👍

XX.11.6.3. Orthogonal projection operators are the simplest kinds of non-scalar bounded operators on a Hilbert space. They are the "building blocks" for general self-adjoint operators via the Spectral Theorem ().

XX.11.6.4. Proposition. Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{Y}$ a closed subspace. Then $P_\mathcal{Y}$ is an idempotent operator on $\mathcal{H}$, i.e. $P_\mathcal{Y}^2 = P_\mathcal{Y}$. Every idempotent bounded operator P on $\mathcal{H}$ with $\|P\| \leq 1$ is an orthogonal projection operator, i.e. $P = P_\mathcal{X}$ for some closed subspace $\mathcal{X}$ of $\mathcal{H}$.

Proof: If $x \in \mathcal{H}$, $x = y + z$, then

$$P_\mathcal{Y}^2(x) = P_\mathcal{Y}(P_\mathcal{Y}(x)) = P_\mathcal{Y}(y) = y = P_\mathcal{Y}(x) \ .$$

Conversely, if $P^2 = P$, let $\mathcal{X}$ be the range of P. If $x \in \mathcal{X}$, then $x = P(v)$ for some $v \in \mathcal{H}$, so $P(x) = P^2(v) = P(v) = x$. If $w = v - x$, we then have $w \in \mathcal{N}(P)$ and $v = x + w$; thus, if v is any vector in $\mathcal{H}$, we can write $v = P(v) + w$ with $w \in \mathcal{N}(P)$. If $x \in \mathcal{X}$ and $w \in \mathcal{N}(P)$, then, for any $\alpha \in \mathbb{F}$,

$$\|x\| = \|P(x + \alpha w)\| \leq \|x + \alpha w\|$$

if $\|P\| \leq 1$; thus $x \perp w$ by XX.11.5.6., i.e. $\mathcal{N}(P) \subseteq \mathcal{X}^\perp$. But if $v \in \mathcal{X}^\perp$, we can write $v = P(v) + w$ with $w \in \mathcal{N}(P)$. Since $P(v) \in \mathcal{X}$ and $w \in \mathcal{X}^\perp$, we must have $v = w$ by XX.11.5.7., i.e. $v \in \mathcal{N}(P)$, $\mathcal{N}(P) = \mathcal{X}^\perp$. Since $\mathcal{X} = \{x \in \mathcal{H} : P(x) = x\}$ and P is bounded, if $x_n \in \mathcal{X}$ and $x_n \to x$, then $x_n = P(x_n) \to P(x)$, so $x = P(x)$, $x \in \mathcal{X}$, and $\mathcal{X}$ is closed (this can be seen more easily by noting that $\mathcal{X}$ is the null space of the bounded operator $I - P$). Thus, if $v \in \mathcal{H}$, and $v = x + w$ for $x \in \mathcal{X}$, $w \in \mathcal{X}^\perp$, we have $P(v) = x = P_{\mathcal{X}}(v)$. 👍

Thus the projections have a simple algebraic characterization: they are exactly the idempotent bounded operators on $\mathcal{H}$ of norm ≤ 1.

Here are some other elementary properties of projections:

XX.11.6.5. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{Y}$ a closed subspace. Then $P_{\mathcal{Y}^\perp} = I - P_{\mathcal{Y}}$.

XX.11.6.6. The projection $I - P$ is often written $P^\perp$. Thus $P_{\mathcal{Y}^\perp} = P_{\mathcal{Y}}^\perp$. We say two projections P and Q are *orthogonal*, written $P \perp Q$, if $PQ = 0$. We have $P_{\mathcal{X}} \perp P_{\mathcal{Y}}$ if and only if $\mathcal{Y} \subseteq \mathcal{X}^\perp$. Also, if $P \perp Q$, then $P + Q$ is a projection (and conversely). We have $P \perp P^\perp$ for any P.

XX.11.6.7. PROPOSITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be closed subspaces of $\mathcal{H}$. The following are equivalent:

(i) $\mathcal{X} \subseteq \mathcal{Y}$.

(ii) $P_{\mathcal{X}} P_{\mathcal{Y}} = P_{\mathcal{Y}} P_{\mathcal{X}} = P_{\mathcal{X}}$.

(iii) $P_{\mathcal{Y}} - P_{\mathcal{X}}$ is a projection (it is $P_{\mathcal{Y} \cap \mathcal{X}^\perp}$).

XX.11.6.8. This result suggests that we can put a reasonable partial order on the projections on $\mathcal{H}$ by setting $P \leq Q$ if $PQ = QP = P$. Thus $P_{\mathcal{X}} \leq P_{\mathcal{Y}}$ if and only if $\mathcal{X} \subseteq \mathcal{Y}$, i.e. $P \leq Q$ if and only if $\mathcal{R}(P) \subseteq \mathcal{R}(Q)$. There is a "largest" projection I and a "smallest" projection 0 under this partial order. Any collection $\{P_j\}$ of projections has an infimum $P = \bigwedge_j P_j$ and a supremum $Q = \bigvee_j P_j$: if $P_j = P_{\mathcal{X}_j}$, then $P = P_{\mathcal{X}}$, where $\mathcal{X} = \cap_j \mathcal{X}_j$, and $Q = P_{\mathcal{Y}}$, where $\mathcal{Y}$ is the closed linear span of $\cup_j \mathcal{X}_j$. In particular, if P and Q are projections, with $P = P_{\mathcal{X}}$ and $Q = P_{\mathcal{Y}}$, then $P \wedge Q$ is $P_{\mathcal{X} \cap \mathcal{Y}}$ and $P \vee Q$ is $P_{\{\mathcal{X} + \mathcal{Y}\}^-}$. Thus the projections on $\mathcal{H}$ form a complete lattice ().

Using the complement $P \mapsto P^\perp = I - P$, the projections on $\mathcal{H}$ form a complete Boolean algebra.

XX.11.6.9. **Caution:** Two projections P and Q do not commute (i.e. $PQ \neq QP$) in general: for example, consider projections from $\mathbb{R}^2$ onto two one-dimensional subspaces which are distinct but not orthogonal. $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ commute if and only if the subspaces $\mathcal{X}$ and $\mathcal{Y}$ are "nicely situated" with respect to each other. If $PQ = QP$, we have $P \wedge Q = PQ$ and $P \vee Q = P + Q - PQ$. In general, there are no simple algebraic formulas for $P \wedge Q$ or $P \vee Q$. If $P \perp Q$, i.e. if $PQ = 0$, then also $QP = 0$ and $P \wedge Q = 0$, $P \vee Q = P + Q$.

XX.11.7. Dual Spaces and Weak Topology

Bounded linear functionals on a Hilbert space are easy to describe:

XX.11.7.1. THEOREM. [RIESZ-FRÉCHET] Let $\mathcal{H}$ be a Hilbert space, and ϕ a bounded linear functional on $\mathcal{H}$. Then there is a (unique) vector $\xi \in \mathcal{H}$ with $\phi(\eta) = \langle \eta, \xi \rangle$ for all $\eta \in \mathcal{H}$; and $\|\xi\| = \|\phi\|$.

PROOF: We may assume $\|\phi\| = 1$. $\mathcal{N}(\phi)$ is a closed subspace of $\mathcal{H}$ of codimension one. Thus there is a unique unit vector $\xi \in \mathcal{N}(\phi)^\perp$ with $\phi(\xi) > 0$. Set $t = \phi(\xi)$. Then $t \leq 1$. If $\eta \in \mathcal{H}$, then

$$\eta = \langle \eta, \xi \rangle \xi + \zeta$$

for some $\zeta \in \mathcal{N}(\phi)$, and thus $\phi(\eta) = \langle \eta, \xi \rangle t$. Let $\epsilon > 0$. There is a unit vector $\eta \in \mathcal{H}$ with $|\phi(\eta)| > 1 - \epsilon$. Then, by the CBS inequality,

$$t \geq |\langle \eta, \xi \rangle| t = |\phi(\eta)| > 1 - \epsilon \ .$$

Since $\epsilon > 0$ is arbitrary, $t = 1$. If $\phi(\eta) = \langle \eta, \xi_1 \rangle$ for all η, then $\langle \eta, \xi - \xi_1 \rangle = 0$ for all η, and thus $\xi_1 = \xi$. 👍

So the dual space $\mathcal{H}^*$ of $\mathcal{H}$ may be identified with $\mathcal{H}$ (the identification is conjugate-linear), and the weak (= weak-*) topology is given by the inner product: $\xi_i \to \xi$ weakly if $\langle \xi_i, \eta \rangle \to \langle \xi, \eta \rangle$ for all η. In particular, a Hilbert space is a reflexive Banach space and the unit ball is therefore weakly compact. A useful way of stating this is:

XX.11.7.2. COROLLARY. Let $\mathcal{H}$ be a Hilbert space. If (ξ_i) is a bounded weak Cauchy net in $\mathcal{H}$ (i.e. $(\langle \xi_i, \eta \rangle)$ is Cauchy for all $\eta \in \mathcal{H}$), then there is a (unique) vector $\xi \in \mathcal{H}$ such that $\xi_i \to \xi$ weakly, i.e. $\langle \xi_i, \eta \rangle \to \langle \xi, \eta \rangle$ for all $\eta \in \mathcal{H}$.

The weak topology is, of course, distinct from the norm topology (strictly weaker) if $\mathcal{H}$ is infinite-dimensional: for example, an orthonormal sequence of vectors converges weakly to 0. The next proposition gives a connection.

XX.11.7.3. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, $\xi_i, \xi \in \mathcal{H}$, with $\xi_i \to \xi$ weakly. Then $\|\xi\| \leq \liminf \|\xi_i\|$, and $\xi_i \to \xi$ in norm if and only if $\|\xi_i\| \to \|\xi\|$.

PROOF: $\|\xi\|^2 = \lim \langle \xi_i, \xi \rangle \leq \|\xi\| \liminf \|\xi_i\|$ by the CBS inequality.

$$\|\xi_i - \xi\|^2 = \langle \xi_i, \xi_i \rangle - \langle \xi_i, \xi \rangle - \langle \xi, \xi_i \rangle + \langle \xi, \xi \rangle$$

which goes to zero if and only if $\|\xi_i\| \to \|\xi\|$. 👍

XX.11.7.4. If $\mathcal{H}$ is infinite-dimensional, then the weak topology on $\mathcal{H}$ is not first countable, and a weakly convergent net need not be (norm-)bounded. It is an easy consequence of Uniform Boundedness (XX.8.4.1.) that a weakly convergent sequence is bounded. For example, if $\{\xi_n\}$ is an orthonormal sequence in $\mathcal{H}$, then 0 is in the weak closure of $\{\sqrt{n}\xi_n\}$, but no subsequence converges weakly to 0 (cf. [Hal82, Problem 28], [**?**]). If $\mathcal{H}$ is separable, then the restriction of the weak topology to the unit ball of $\mathcal{H}$, and hence to any bounded set, is metrizable [Hal82, Problem 24].

XX.11.8. Direct Sums and Amplifications

There are two standard constructions on Hilbert spaces which are used repeatedly, direct sum and tensor product. Tensor products are discussed in XX.11.9..

XX.11.8.1. If $\{\mathcal{H}_i : i \in \Omega\}$ is a set of Hilbert spaces, we can form the Hilbert space direct sum, denoted $\bigoplus_\Omega \mathcal{H}_i$, as the set of "sequences" (indexed by Ω) $(\cdots \xi_i \cdots)$, where $\xi_i \in \mathcal{H}_i$ and $\sum_i \|\xi_i\|^2 < \infty$ (so $\xi_i \neq 0$ for only countably many i). The inner product is given by

$$\langle(\cdots \xi_i \cdots), (\cdots \eta_i \cdots)\rangle = \sum_i \langle \xi_i, \eta_i \rangle.$$

This is the completion of the algebraic direct sum with respect to this inner product. If all $\mathcal{H}_i$ are the same $\mathcal{H}$, the direct sum is called the *amplification* of $\mathcal{H}$ by $card(\Omega)$. The amplification of $\mathcal{H}$ by n (the direct sum of n copies of $\mathcal{H}$) is denoted by $\mathcal{H}^n$; the direct sum of countably many copies of $\mathcal{H}$ is denoted $\mathcal{H}^\infty$ (or sometimes $l^2(\mathcal{H})$).

Unless otherwise qualified, a "direct sum" of Hilbert spaces will always mean the Hilbert space direct sum.

XX.11.9. Tensor Products of Inner Product Spaces

We can form the (algebraic) tensor product of any (finite) collection of vector spaces over a field $\mathbb{F}$ (). If $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$ and each of the vector spaces has a specified inner product, there is a canonical way to form an inner product on the tensor product.

XX.11.9.1. **Notation:** In functional analysis, it is customary to use the symbol $\odot$ instead of $\otimes$ for the algebraic tensor product of Hilbert spaces. The reason is that the algebraic tensor product of infinite-dimensional Hilbert spaces is not complete with respect to the induced inner product (it is if the spaces are finite-dimensional). But it can be completed to form a Hilbert space which is called the Hilbert space tensor product; the symbol $\otimes$ is used for the Hilbert space tensor product. We will use $\odot$ for the algebraic tensor product in this section, so our notation will be inconsistent with ().

XX.11.9.2. Definition. Let $(\mathcal{V}_k, \langle\cdot,\cdot\rangle_k)$ $(1 \leq k \leq n)$ be a finite collection of inner product spaces over $\mathbb{F}$ ($\mathbb{R}$ or $\mathbb{C}$). On the algebraic tensor product $\mathcal{V} = \mathcal{V}_1 \odot \cdots \odot \mathcal{V}_n$ over $\mathbb{F}$ define an $\mathbb{F}$-valued function on elementary tensors by

$$\langle \xi_1 \otimes \cdots \otimes \xi_n, \eta_1 \otimes \cdots \otimes \eta_n \rangle = \langle \xi_1, \eta_1 \rangle_1 \cdot \langle \xi_2, \eta_2 \rangle_2 \cdots \langle \xi_n, \eta_n \rangle_n$$

and extend to all of $\mathcal{V}$ by linearity.

XX.11.9.3. PROPOSITION. The function $\langle \cdot, \cdot \rangle$ is well defined on $\mathcal{V}$ and is an inner product.

XX.11.9.4. DEFINITION. If $\mathcal{H}_1$, $\mathcal{H}_2$ are Hilbert spaces, the Hilbert space tensor product $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the completion of $\mathcal{H}_1 \odot \mathcal{H}_2$. Tensor products of finitely many Hilbert spaces are defined similarly. (Infinite tensor products are trickier and will be discussed in **??**.)

There is a nice way to view the tensor product of two (or more) Hilbert spaces:

XX.11.9.5. PROPOSITION. If $\mathcal{H}$ and $\mathcal{H}'$ are Hilbert spaces and $\{\eta_j : j \in \Omega\}$ is an orthonormal basis for $\mathcal{H}'$, then there is an isometric isomorphism from $\mathcal{H} \otimes \mathcal{H}'$ to the amplification (XX.11.8.1.) of $\mathcal{H}$ by $card(\Omega)$, given by $\sum_i \xi_i \otimes \eta_i \mapsto (\cdots \xi_i \cdots)$. In particular, $\mathcal{H}^n$ is naturally isomorphic to $\mathcal{H} \otimes \mathbb{C}^n$ and $\mathcal{H}^\infty \cong \mathcal{H} \otimes l^2$.

XX.11.9.6. If $\{\xi_j : j \in \Omega\}$ and $\{\eta_k : k \in \Lambda\}$ are orthonormal bases for $\mathcal{H}$ and $\mathcal{H}'$ respectively, then

$$\{\xi_j \otimes \eta_k : (j,k) \in \Omega \times \Lambda\}$$

is an orthonormal basis for $\mathcal{H} \otimes \mathcal{H}'$, and similarly for general finite tensor products. If $\mathcal{H}$ or $\mathcal{H}'$ is finite-dimensional, there is no distinction between $\mathcal{H} \odot \mathcal{H}'$ and $\mathcal{H} \otimes \mathcal{H}'$ since $\mathcal{H} \odot \mathcal{H}'$ is already complete. If both are finite-dimensional, $dim(\mathcal{H}) = n$ and $dim(\mathcal{H}') = m$, then $dim(\mathcal{H} \otimes \mathcal{H}') = nm$.

XX.11.9.7. The usual commutative and associative isomorphisms $\mathcal{H}_1 \otimes \mathcal{H}_2 \cong \mathcal{H}_2 \otimes \mathcal{H}_1$ and $(\mathcal{H}_1 \otimes \mathcal{H}_2) \otimes \mathcal{H}_3 \cong \mathcal{H}_1 \otimes (\mathcal{H}_2 \otimes \mathcal{H}_3) \cong \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_3$, etc., are all isometric.

XX.11.9.8. EXAMPLE. If $(\mathcal{X}, \mu)$ is a measure space and $\mathcal{H}$ is a separable Hilbert space, then the Hilbert space $L^2(X,\mu) \otimes \mathcal{H}$ is naturally isomorphic to $L^2(X, \mu, \mathcal{H})$, the set of weakly measurable functions $f : X \to \mathcal{H}$ (i.e. $x \mapsto \langle f(x), \eta \rangle$ is a complex-valued measurable function for all $\eta \in \mathcal{H}$) such that $\int_X \|f(x)\|^2 \, d\mu(x) < \infty$, with inner product

$$\langle f, g \rangle = \int_X \langle f(x), g(x) \rangle \, d\mu(x).$$

The isomorphism sends $f \otimes \eta$ to $(x \mapsto f(x)\eta)$.

XX.11.10. Real Hilbert Spaces

One can also consider *real inner product spaces*, real vector spaces with a (bilinear, real-valued) inner product $(\cdot, \cdot)$ satisfying the properties of **??** for $\alpha \in \mathbb{R}$. A *real Hilbert space* is a complete real inner product space. A real Hilbert space $\mathcal{H}_\mathbb{R}$ can be complexified to $\mathbb{C} \otimes_\mathbb{R} \mathcal{H}_\mathbb{R}$; conversely, a (complex) Hilbert space can be regarded as a real Hilbert space by restricting scalar multiplication and using the real inner product $(\xi, \eta) = Re\langle \xi, \eta \rangle$. (Note, however, that these two processes are not quite inverse to each other.) All of the results of this section, and many (but by no means all) results about operators, hold verbatim or have obvious exact analogs in the case of real Hilbert spaces.

Throughout this volume, the term "Hilbert space," unless qualified with "real," will always denote a *complex* Hilbert space; and "linear" will mean "complex-linear."

XX.11.11. Exercises

XX.11.11.1. Prove the CBS inequality for $\mathbb{R}^n$ as follows. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$.

(a) Show that

$$\left(\sum_{k=1}^{n} x_k y_k\right)^2 + \sum_{j<k} (x_j y_k - x_k y_j)^2 = \left(\sum_{k=1}^{n} x_k^2\right)\left(\sum_{k=1}^{n} y_k^2\right) .$$

(b) Use that all the terms in the second sum are nonnegative to deduce the CBS inequality.

(c) Equality holds in the CBS inequality if and only if all terms in the second sum are zero. Show that this implies that $\{\boldsymbol{x}, \boldsymbol{y}\}$ is linearly dependent.

This was CAUCHY'S proof of the CBS inequality from [Cau21].

XX.11.11.2. Give another proof of the CBS inequality for $\mathbb{R}^n$ as follows.

(a) Show that for $x, y \in \mathbb{R}$, $|xy| \leq \frac{1}{2}(x^2 + y^2)$, with equality if and only if $y = \pm x$. [Use that $(x+y)^2$ and $(x-y)^2$ are nonnegative.]

(b) Show that it suffices to prove the CBS inequality for $\boldsymbol{x}, \boldsymbol{y}$ of norm 1.

(c) if $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ have norm 1, then

$$\left|\sum_{k=1}^{n} x_k y_k\right| \leq \sum_{k=1}^{n} |x_k y_k| \leq \frac{1}{2}\left(\sum_{k=1}^{n} x_k^2 + \sum_{k=1}^{n} y_k^2\right) = 1$$

with equality if and only if $\boldsymbol{y} = \pm \boldsymbol{x}$.

XX.11.11.3. Let f and g be real-valued continuous functions on $[a, b]$. Use Riemann sums for $\int_a^b f(t)g(t)\,dt$, $\int_a^b [f(t)]^2\,dt$, and $\int_a^b [g(t)]^2\,dt$, and the CBS inequality for $\mathbb{R}^n$, to prove

$$\left(\int_a^b f(t)g(t)\,dt\right)^2 \leq \left(\int_a^b [f(t)]^2\,dt\right)\left(\int_a^b [g(t)]^2\,dt\right)$$

with equality if and only if $\{f, g\}$ is linearly dependent. This was BUNYAKOVSKII'S inequality and proof from [Bun59].

XX.11.11.4. Let $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ be a real pre-inner product space, and $x, y \in \mathcal{V}$.

(a) Write $\langle x + \alpha y, x + \alpha y \rangle$ as a quadratic function of α.

(b) Use the quadratic formula and the fact that $\langle x + \alpha y, x + \alpha y \rangle \geq 0$ for all α to deduce the CBS inequality.

This was essentially SCHWARZ'S proof from [Sch88] (he only explicitly considered the case where the inner product was given by a (multiple) integral as in XX.11.1.4.(ii).)

XX.11.11.5. Deduce the CBS inequality for complex pre-inner products from the inequality for real pre-inner products. [Consider the real inner product $(x, y) = Re(\langle x, y \rangle)$ and suitable scalar multiples.] This gives a simpler proof for indefinite complex inner products.

XX.11.11.6. [Ste04] Prove the CBS inequality for the standard inner product on $\mathbb{R}^n$ by induction, starting with the case $n = 2$ (it is trivial for $n = 1$).

XX.11.11.7. (a) Justify the name "Parallelogram Law" by giving a geometric interpretation in terms of the lengths of the diagonals of a parallelogram. [If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$, consider the parallelogram with vertices $\mathbf{0}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{x} + \boldsymbol{y}$.] See VI.3.1.22.. Show that the Pythagorean Theorem is the special case of the Parallelogram Law for a rhombus. [Given a right triangle, assemble four copies into a rhombus.]

(b) Show that the Parallelogram Law is equivalent to the statement that

$$\|x + y\|^2 - \|x - y\|^2 = 2(\|x + y\|^2 - \|x\|^2 - \|y\|^2)$$

for all x, y.

(c) Show that in an inner product space,

$$\langle x, y\rangle = \frac{1}{2}[(\|x + y\|^2 - \|x\|^2 - \|y\|^2) + i(\|x + iy\|^2 - \|x\|^2 - \|y\|^2)]$$

for all x, y. Give a similar formula for real inner products.

XX.11.11.8. Let $\mathcal{V}$ be a normed vector space over $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. Show that the norm on $\mathcal{V}$ comes from an inner product if and only if every two-dimensional subspace of $\mathcal{V}$ is isometrically isomorphic to $\mathbb{F}^2$ with the induced norm from the standard inner product.

XX.11.11.9. Suppose $(\mathcal{V}, \langle\cdot,\cdot\rangle)$ is an inner product space. Expand the right side of the Polarization Identity (XX.11.2.8.) to show that the identity is satisfied.

XX.11.11.10. (a) In $C([0, 1])$, let f_n $(n \geq 2)$ be defined by

$$f_n(t) = \begin{cases} 0 & \text{if} \quad 0 \leq t \leq \frac{1}{2} \\ \text{linear} & \text{if} \quad \frac{1}{2} \leq t \leq \frac{1}{2} + \frac{1}{n} \\ 1 & \text{if} \quad \frac{1}{2} \leq t \leq 1 \end{cases} .$$

Show that (f_n) is a uniformly bounded sequence in $C([0, 1])$ which is a Cauchy sequence with respect to the norm induced by the inner product, but which does not converge in $C([0, 1])$ in the induced norm. (Note that we have referred here to the uniform norm $\|\cdot\|_\infty$ () as well as the norm induced by the inner product; in this problem and succeeding problems we discuss Cauchy sequences and convergence only with respect to the norm induced by the inner product. The sequence (f_n) is bounded but not Cauchy in the uniform norm.)

(b) It is slightly harder to show that $PC([0, 1])$ (XX.11.3.9.) is not complete. Let g_n $(n \geq 2)$ be defined by

$$g_n(t) = \begin{cases} 0 & \text{if} \quad 0 \leq t \leq \frac{1}{n} \\ \sin\left(\frac{\pi}{t}\right) & \text{if} \quad \frac{1}{n} < t \leq 1 \end{cases} .$$

Show that (g_n) defines a uniformly bounded Cauchy sequence in $C([0, 1])$ which does not converge in $PC([0, 1])$.

(c) It is easier to give an example of a nonconvergent Cauchy sequence which is not uniformly bounded. Let

$$h_n(t) = \min\left(n, \frac{1}{\sqrt[3]{t}}\right)$$

for $0 \leq t \leq 1$. Then (h_n) is a Cauchy sequence in $C([0, 1])$ which does not converge even in $PC([0, 1])$.

XX.11.11.11. Let $\mathcal{R}([a,b])$ be the set of complex-valued Riemann-integrable functions on $[a,b]$.

(a) Show that the formula

$$\langle f,g\rangle = \int_a^b f(t)\overline{g(t)}\,dt$$

(where the integral is a Riemann integral) defines a pre-inner product on $\mathcal{R}([a,b])$.

(b) Explicitly describe the space $\mathcal{N}$.

(c) Let $R([a,b]) = \mathcal{R}([a,b])/\mathcal{N}$. Show that $PC([a,b])$ can be naturally identified with a dense subspace of $R([a,b])$.

(d) The Cauchy sequence (h_n) of XX.11.11.10.(c) does not converge in $R([0,1])$.

(e) Find a uniformly bounded Cauchy sequence in $R([a,b])$ that does not converge. [A Cantor set of positive measure () is helpful.]

(f) Try to describe the completion of $R([a,b])$ (which will also be the completion of $C([a,b])$ and $PC([a,b])$ since these are dense in $R([a,b])$). The example in (d) shows that we must at least add all nonnegative improperly Riemann square-integrable functions on $[a,b]$, but (e) shows that even linear combinations of these do not give enough. The answer involves Lebesgue measure and integration (); this example gives good motivation for development of the Lebesgue theory.

XX.11.11.12. Let $(\mathcal{V},\langle\cdot,\cdot\rangle)$ be a (real or complex) inner product space. Let $\mathcal{H}$ be the set of equivalence classes of Cauchy sequences in $\mathcal{V}$ as in the proof of (). Show that

$$\langle[(x_n)],[(y_n)]\rangle = \lim_{n\to\infty}\langle x_n,y_n\rangle$$

gives a well-defined inner product on $\mathcal{H}$ inducing the norm defined in ().

XX.11.11.13. Let $\mathcal{H}$ be an infinite-dimensional Hilbert space (real or complex).

(a) Show that for every $r>0$, an open ball of radius $3r$ in $\mathcal{H}$ contains infinitely many pairwise disjoint open balls of radius r. If $\mathcal{H}$ is nonseparable, an open ball of radius $3r$ contains uncountably many pairwise disjoint open balls of radius r. [Consider open balls centered at points of an orthonormal basis.] This is one justification for the statement "There's a lot of room in Hilbert space."

(b) Use (a) to show that there is no translation-invariant Borel measure on $\mathcal{H}$ for which open balls have finite positive measure.

(c) Show that an open ball of radius $2r$ in a Hilbert space (finite- or infinite-dimensional, real or complex) cannot contain more than two disjoint open balls of radius r. [If there are three such balls, consider the two-dimensional real affine subspace spanned by the centers.]

(d) In (a), 3 may be replaced by $1+\sqrt{2}\approx 2.414$ and in (c), 2 may be replaced by any number less than $1+\frac{2}{\sqrt{3}}\approx 2.1547$. (These numbers are surprisingly close together.) In fact, for any $n\in\mathbb{N}$, an open ball of radius $1+\sqrt{2\left(1-\frac{1}{n}\right)}$ is the smallest open ball containing n pairwise disjoint open balls of radius 1 in any Hilbert space of (real) dimension $n-1$ or greater; cf. https://mathoverflow.net/questions/369444/balls-in-hilbert-space (the unit balls are centered at the vertices of a regular $(n-1)$-simplex, and the large ball at the centroid). The question of the smallest open ball in a Hilbert space of (real) dimension $d<n-1$ containing n pairwise disjoint open balls of radius 1 appears to be much harder; it is essentially the optimal sphere-packing problem for fixed d and $n\gg d$ (XIII.19.4.11.).

XX.11.11.14. Let $\{\mathcal{X}_i : i \in I\}$ be an indexed collection of Banach spaces. Define the *ℓ^2-direct sum* $\bigoplus_{\ell^2} \mathcal{X}_i$ to be the subset of $\prod_{i\in I} \mathcal{X}_i$ consisting of all elements $(\cdots x_i \cdots)$ for which $\sum_{i\in I} \|x_i\|^2 < \infty$. (Note that $\bigoplus_{\ell^2} \mathcal{X}_i = \prod_{i\in I} \mathcal{X}_i$ if I is finite.)

(a) Let $\mathcal{X} = \bigoplus_{\ell^2} \mathcal{X}_i$. Show that $\mathcal{X}$ is a vector subspace of $\prod_{i\in I} \mathcal{X}_i$ and that $\|\cdot\|$ defined on $\mathcal{X}$ by

$$\|(\cdots x_i \cdots)\| = \left[\sum_{i\in I} \|x_i\|^2\right]^{1/2}$$

is a norm on $\mathcal{X}$.

(b) Show that $(\mathcal{X}, \|\cdot\|)$ is complete, i.e. a Banach space.

(c) Show that if each $\mathcal{X}_i$ is a Hilbert space (i.e. the norm satisfies the Parallelogram Law), then $\mathcal{X}$ is also a Hilbert space.

XX.11.11.15. Uniformly Convex Banach Spaces. Let $\mathcal{X}$ be a Banach space. $\mathcal{X}$ is *strictly convex* if, whenever $x, y \in \mathcal{X}$, $x \neq y$, $\|x\| = \|y\| = 1$, we have $\left\|\frac{1}{2}(x+y)\right\| < 1$. $\mathcal{X}$ is *uniformly convex* if for every $\epsilon > 0$ there is a $\delta > 0$ such that, whenever $x, y \in X$, $\|x\| = \|y\| = 1$, $\|x-y\| > \epsilon$, we have $\left\|\frac{1}{2}(x+y)\right\| < 1-\delta$. Note that a uniformly convex Banach space is strictly convex.

(a) Show that $\mathcal{X}$ is strictly convex if and only if every element of $\mathcal{X}$ of norm 1 is an extreme point of the closed unit ball of $\mathcal{X}$ ().

(b) Show that any (closed) subspace of a strictly [resp. uniformly] convex Banach space is strictly [resp. uniformly] convex.

(c) Use the Parallelogram Law to show that any Hilbert space is uniformly convex (cf. XX.11.4.2.).

(d) Prove *Clarkson's Inequality* [**?**]: if $z, w \in \mathbb{C}$, $p \in [2, \infty)$, $q = \frac{p}{p-1}$ (XX.15.1.1.), then

$$|z+w|^p + |z-w|^p \leq 2(|z|^q + |w|^q)^{p/q}$$

$$|z+w|^q + |z-w|^q \geq 2(|z|^p + |w|^p)^{q/p}$$

(the case $p = 2$ is just the Parallelogram Law).

(e) Show that $(\mathbb{R}^n, \|\cdot\|_p)$ is uniformly convex for $1 < p < \infty$, but that $(\mathbb{R}^n, \|\cdot\|_1)$ and $(\mathbb{R}^n, \|\cdot\|_\infty)$ are not strictly convex. What about ℓ^p or $L^p(X, \mathcal{A}, \mu)$ for a measure space $(X, \mathcal{A}, \mu)$?

(f) Show that a strictly convex finite-dimensional Banach space is uniformly convex.

(g) Show that any ℓ^2-direct sum (XX.11.11.14.) of strictly convex Banach spaces is strictly convex.

(h) Show that $\bigoplus_{\ell^2}(\mathbb{R}^2, \|\cdot\|_p)$ for $p \in \mathbb{N}$, $p \geq 2$, is strictly convex but not uniformly convex. Thus (g) is not true if "strictly" is replaced by "uniformly." (But it is if the index set is finite.)

(i) Formulate and prove a precise statement to the effect that $\bigoplus_{\ell^2} \mathcal{X}_i$ is uniformly convex if and only if the $\mathcal{X}_i$ are "uniformly uniformly convex." Deduce the last statement of (h).

(j) Adapt the proof of XX.11.4.1. to show that every uniformly convex Banach space has the best approximation property.

It is known that a uniformly convex Banach space is always reflexive (Milman-Pettis Theorem). A strictly convex Banach space need not be reflexive; in fact, every separable Banach space has an equivalent strictly convex norm (cf. https://math.stackexchange.com/questions/488790/is-there-a-non-reflexive-banach-space-whi rq=1).

XX.11.11.16. Let $\mathcal{H}$ be a Hilbert space (real or complex).

(a) Show that the norm topology and weak topology coincide on the unit sphere of $\mathcal{H}$. [Use XX.11.7.3..]

(b) Show that this fails in most Banach spaces, even Banach spaces isomorphic to a Hilbert space. [If $\mathcal{H}$ is a Hilbert space, consider $\mathbb{F} \oplus \mathcal{H}$ with max norm $\|(\lambda, x)\| = \max(|\lambda|, \|x\|)$, and recall that an orthonormal sequence in $\mathcal{H}$ converges to 0 weakly.] Is it true in a uniformly convex Banach space?

XX.11.12. Sesquilinear Forms and Operators

XX.11.12.1. If $\mathcal{X}$ is a normed complex vector space, $\mathcal{H}$ a Hilbert space, and $T : \mathcal{X} \to \mathcal{H}$ a bounded operator, set

$$B(x, y) = \langle Tx, y\rangle \ .$$

Then it is easily checked that B is a sesquilinear form on $\mathcal{X} \times \mathcal{H}$, which is *bounded*, i.e. there is a K such that

$$|B(x, y)| \leq K\|x\|\|y\|$$

for all $x \in \mathcal{X}$, $y \in \mathcal{H}$. The smallest such K, the *norm* $\|B\|$ of B, is equal to $\|T\|$.

The following converse establishes a one-one correspondence between bounded operators and bounded sesquilinear forms. We state it for complex spaces; there is an essentially identical analogous version for real spaces, replacing "sesquilinear" with "bilinear."

XX.11.12.2. Proposition. Let $\mathcal{X}$ be a normed vector space, $\mathcal{H}$ a Hilbert space, and B a bounded sesquilinear form on $\mathcal{X} \times \mathcal{H}$. Then there is a unique bounded operator $T : \mathcal{X} \to \mathcal{H}$ such that

$$B(x, y) = \langle Tx, y\rangle$$

for all $x \in \mathcal{X}$, $y \in \mathcal{H}$. $\|T\| = \|B\|$.

Proof: Fix $x \in \mathcal{X}$. Define $\phi : \mathcal{H} \to \mathbb{C}$ by $\phi(y) = B(x, y)$. Then ϕ is a conjugate-linear functional on $\mathcal{H}$, and ϕ is bounded ($\|\phi\| \leq \|B\|\|x\|$). Thus by XX.11.7.1. there is a unique $z \in \mathcal{H}$ with

$$\phi(y) = B(x, y) = \langle z, y\rangle \ .$$

Set $Tx = z$. In this way a function $T : \mathcal{X} \to \mathcal{H}$ is defined. It is an easy consequence of the linearity of B in the first variable that T is linear. If $x \in \mathcal{X}$, we have

$$\|Tx\| = \sup_{\|y\|=1} |\langle Tx, y\rangle| = \sup_{\|y\|=1} |B(x, y)| \leq \|B\|\|x\|$$

and thus T is bounded and $\|T\| \leq \|B\|$. By taking the supremum over all x with $\|x\| = 1$, we obtain $\|T\| = \|B\|$. 👍

XX.11.12.3. There is thus a one-one correspondence between bounded operators and bounded sesquilinear forms. Each point of view has its advantages. Some early authors such as Hilbert used the form approach exclusively instead of working with operators, but some parts of the theory become almost impossibly difficult and cumbersome (e.g. composition of operators) from this point of view. I. Fredholm and F. Riesz emphasized the operator point of view, which is superior for most purposes.

The following special case of XX.11.12.2. is especially useful in the theory of partial differential equations:

XX.11.12.4. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, and B a bounded sesquilinear form on $\mathcal{H} \times \mathcal{H}$ for which there is an $\epsilon > 0$ such that

$$|B(x,x)| \geq \epsilon \|x\|^2$$

for all $x \in \mathcal{H}$. Then there is a unique $T \in \mathcal{B}(\mathcal{H})$ with

$$B(x,y) = \langle Tx, y \rangle$$

for all $x, y \in \mathcal{H}$. T is invertible and $\|T\| = \|B\|$. Thus there is a unique $S \in \mathcal{B}(\mathcal{H})$, namely $S = T^{-1}$, such that

$$B(Sx,y) = \langle x, y \rangle$$

for all $x, y \in \mathcal{H}$. S is invertible and $\|S\| \leq \frac{1}{\epsilon}$.

There is an analogous result for real Hilbert spaces, with "sesquilinear" replaced by "bilinear."

XX.11.12.5. COROLLARY. Let $\mathcal{H}$ be a Hilbert space, and B a bounded sesquilinear form on $\mathcal{H} \times \mathcal{H}$ for which there is an $\epsilon > 0$ such that

$$|B(x,x)| \geq \epsilon \|x\|^2$$

for all $x \in \mathcal{H}$. If ϕ is any bounded linear functional on $\mathcal{H}$, then there is a unique $z \in \mathcal{H}$ such that $\phi(x) = B(x,z)$ for all $x \in \mathcal{H}$.

PROOF: Let T and $S = T^{-1}$ be as in XX.11.12.4. (so $(S^*)^{-1} = T^*$). There is a unique $y \in \mathcal{H}$ with $\phi(x) = \langle x, y \rangle$ for all $x \in \mathcal{H}$. Then $z = S^*y$ is such a z:

$$B(x,z) = \langle Tx, z \rangle = \langle x, T^*z \rangle = \langle x, y \rangle = \phi(x)$$

for all x. Conversely, if z is such a vector, then $\phi(x) = \langle x, T^*z \rangle$ for all x, so $T^*z = y$, $z = S^*y$. 👍

XX.11.12.6. For the proof of XX.11.12.4., everything follows immediately from XX.11.12.2. except invertibility of T, which is a consequence of the next elementary result, and uniqueness of S, which is immediate [if S' is another such operator, then for each $x \in \mathcal{H}$ we have $B(Sx,y) = B(S'x,y)$ and hence $B((S-S')x,y) = 0$ for all $y \in \mathcal{H}$, and in particular for $y = (S-S')x$, so

$$0 = |B((S-S')x,(S-S')x)| \geq \epsilon \|(S-S')x\|^2$$

and $(S-S')x = 0$.]

XX.11.12.7. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$ with the property that there is an $\epsilon > 0$ such that $\langle Tx, x \rangle \geq \epsilon \|x\|^2$ for all $x \in \mathcal{H}$. Then T is invertible. $\|T^{-1}\| \leq \frac{1}{\epsilon}$.

PROOF: By the CBS inequality, $\|Tx\| \geq \epsilon \|x\|$ for all $x \in \mathcal{H}$, i.e. T is bounded below. Thus T is one-to-one with closed range (). If $x \perp \mathcal{R}(T)$, then $\langle Tx, x \rangle = 0$, so $x = 0$, and $\mathcal{R}(T)$ is dense in $\mathcal{H}$. Thus $\mathcal{R}(T) = \mathcal{H}$. We have, for any $x \in \mathcal{H}$, $\|x\| = \|T(T^{-1}x)\| \geq \epsilon \|T^{-1}x\|$. 👍

XX.11.12.8. In many references, Proposition XX.11.12.4., or sometimes XX.11.12.5., is called the "Lax-Milgram Theorem" (or sometimes "the famous Lax-Milgram Theorem") since it appeared in [LM54]. Although it is useful, it does not seem to merit named status, let alone fame, since

(1) It is such an easy and obvious consequence of XX.11.7.1. (although the proofs in [LM54] and in some other references such as [Yos80] are unnecessarily complicated).

(2) The more general result XX.11.12.2. was well known to Hilbert, Riesz, and other early workers in Hilbert space, long before Lax and Milgram came along.

Not everything published by prominent mathematicians deserves hagiography.

XX.12. Banach Algebras and Spectral Theory

Most of the advanced study of operators is connected with the spectrum, which is a generalization of the set of eigenvalues. While the most immediate applications of spectral theory are to operators, the theory itself is done most efficiently in the more general context of Banach algebras, and spectral theory has numerous applications beyond the case of operators. We will thus first set up the theory of Banach algebras and develop spectral theory there.

XX.12.1. Banach Algebras

A Banach algebra is a Banach space which has a multiplication with nice compatibility properties:

XX.12.1.1. Definition. Let $\mathbb{F}$ be $\mathbb{R}$ or $\mathbb{C}$. A *normed algebra* over $\mathbb{F}$ is an algebra () A over $\mathbb{F}$ with a norm satisfying

$$\|xy\| \leq \|x\|\|y\|$$

for all $x, y \in A$ (*submultiplicative norm*). If A is a normed algebra which is a Banach space (i.e. is complete), A is a *Banach algebra*.

We thus have real Banach algebras and complex Banach algebras. For technical reasons, we will usually consider complex Banach algebras, especially for spectral theory; thus we will make the convention that the term "Banach algebra" without the qualification "real" will denote a *complex* Banach algebra.

There are a great many important Banach algebras with varied properties. For example:

XX.12.1.2. Examples. (i) Let $\mathcal{X}$ be a Banach space (over $\mathbb{F}$). Then $\mathcal{B}(\mathcal{X})$, with the usual operations and operator norm (), is a Banach algebra (over $\mathbb{F}$).

(ii) Let X be a compact Hausdorff space, and let $C(X)$ be the set of complex-valued continuous functions on X. Give $C(X)$ the usual operations of pointwise addition, pointwise multiplication, and pointwise scalar multiplication, and the supremum norm (). Then $C(X)$ is a (complex) Banach algebra. Similarly, $C_{\mathbb{R}}(X)$ is a real Banach algebra. More generally, if X is locally compact, the set $C_0(X)$ of complex-valued continuous functions vanishing at infinity () is a Banach algebra with the same operations and norm.

(iii) As a special case of (i), $M_n(\mathbb{F}) = \mathcal{B}(\mathbb{F}^n)$ becomes a Banach algebra when given the operator norm from any norm on $\mathbb{F}^n$. Thus there are many norms on $M_n(\mathbb{F})$ making it into a Banach algebra. As a special case, or as a special case of (ii) with X a one-point space, $\mathbb{F}$ is a Banach algebra over $\mathbb{F}$ with norm $\|\alpha\| = |\alpha|$.

(iv) If $\mathcal{X}$ is any Banach space, define $xy = 0$ for all $x, y \in \mathcal{X}$. Then $\mathcal{X}$ is a Banach algebra. (This example is not very interesting, to be sure.)

These examples only scratch the surface of the possibilities, even the interesting possibilities.

Joint Continuity of Multiplication

In a Banach space (or just normed vector space), addition and scalar multiplication are jointly continuous. In a normed algebra, multiplication is also jointly continuous:

XX.12.1.3. PROPOSITION. Let A be a normed algebra. If $x_n, y_n, x, y \in A$ and $x_n \to x$, $y_n \to y$, then $x_n y_n \to xy$.

PROOF: By the usual argument,

$$\|x_n y_n - xy\| = \|x_n y_n - x_n y + x_n y - xy\| \leq \|x_n y_n - x_n y\| + \|x_n y - xy\| \leq \|x_n\|\|y_n - y\| + \|y\|\|x_n - x\| \ .$$

Since (x_n) converges, it is bounded (). 👍

Completion of a Normed Algebra

XX.12.1.4. Similarly, if A is a normed algebra, then its completion () has a natural multiplication making it a Banach algebra. For if (x_n) and (y_n) are Cauchy sequences in A, then, for any n and m,

$$\|x_n y_n - x_m y_m\| = \|x_n y_n - x_n y_m + x_n y_m - x_m y_m\| \leq \|x_n y_n - x_n y_m\| + \|x_n y_m - x_m y_m\| \leq \|x_n\|\|y_n - y_m\| + \|y_m\|\|x_n - x_m\| \ .$$

Since Cauchy sequences are bounded (), it follows that $(x_n y_n)$ is a Cauchy sequence.

Unital Banach Algebras and Unitization

XX.12.1.5. Examples XX.12.1.2.(ii) and (iv) show that a Banach algebra need not have a multiplicative identity. A Banach algebra with a multiplicative identity is a *unital Banach algebra*. The multiplicative identity, if it exists, is generally denoted 1. We usually assume $\|1\| = 1$; while this is not necessarily true in general, the norm can be replaced by an equivalent Banach algebra norm with this property (Exercise XX.12.6.1.).

If a Banach algebra is not unital, a unit can be added in a systematic way:

XX.12.1.6. PROPOSITION. Let A be a nonunital Banach algebra over $\mathbb{F}$. Let $A^\dagger = A \oplus \mathbb{F}$. Define addition and scalar multiplication coordinatewise, and multiplication and a norm by

$$(x, \lambda) \cdot (y, \mu) = (xy + \lambda y + \mu x, \lambda\mu)$$

$$\|(x, \lambda)\| = \|x\| + |\lambda| \ .$$

Then $A^\dagger$ is a unital Banach algebra containing a copy of A as an ideal () of codimension 1, called the *unitization* of A.

The rule for multiplication can be remembered by thinking of (x, λ) as $x + \lambda 1$. (The multiplicative identity is $(0, 1)$.)

PROOF: Checking the algebraic rules for an algebra is routine but tedious, and is left to the reader. It is clear that $\|\cdot\|$ is a norm on $A^\dagger$. For submultiplicativity, if $x, y \in A$, $\lambda, \mu \in \mathbb{F}$,

$$\|(x, \lambda)(y, \mu)\| = \|xy + \lambda y + \mu x\| + |\lambda\mu| \leq \|xy\| + \|\lambda y\| + \|\mu x\| + |\lambda\mu|$$

$$\leq \|x\|\|y\| + |\lambda|\|y\| + |\mu|\|x\| + |\lambda\mu| = (\|x\| + |\lambda|)(\|y\| + |\mu|) \ .$$

The map $x \mapsto (x, 0)$ identifies A with a subalgebra (ideal) of $A^\dagger$. To see that $A^\dagger$ is complete, if (a_n, λ_n) is a Cauchy sequence in $A^\dagger$, then (a_n) is a Cauchy sequence in A and (λ_n) a Cauchy sequence in $\mathbb{F}$. These converge to some a and λ, and $(a_n, \lambda_n) \to (a, \lambda)$. 👍

XX.12.1.7. Many aspects of Banach algebra theory are technically simpler for unital Banach algebras; XX.12.1.6. allows the nonunital case to be largely reduced to the unital case. We do not want to just restrict to unital Banach algebras, however, since many naturally occurring Banach algebras are nonunital. Many, but not all, nonunital Banach algebras have an approximate unit () which can be used as a substitute.

XX.12.2. Invertibility, Ideals, and Quotients

One of the remarkable and appealing aspects of the theory of Banach algebras is the intricate interplay between the topology and the algebraic structure. One of the most basic, and most important, examples of this is the use of power series in Banach algebras to construct inverses (and, later, other elements).

XX.12.2.8. Recall that an element x in a unital algebra A is *invertible* if there is a $y \in A$ with $xy = yx = 1$. **Caution:** Since a Banach algebra is not necessarily commutative, for x to be invertible it is not generally enough that there is a y with $yx = 1$ (such an x is *left invertible*), or that there is a z with $xz = 1$ (such an x is *right invertible*); cf. (). However, as in any unital ring, if x is both left and right invertible, the left and right inverses coincide and x is invertible (and conversely): if $yx = xz = 1$, then

$$y = y(xz) = (yx)z = z \ .$$

XX.12.2.9. THEOREM. Let A be a unital Banach algebra, and $x \in A$. If the power series

$$\sum_{k=0}^{\infty} x^k$$

(where $x^0 = 1$ by convention) converges in A, the sum is a multiplicative inverse for $1 - x$. If $\|x\| < 1$, then the power series converges to a multiplicative inverse for $1 - x$, and

$$\|(1-x)^{-1}\| \leq \frac{1}{1 - \|x\|} \ .$$

PROOF: Suppose the series converges, and let y be the sum. If

$$s_n = \sum_{k=0}^{n} x^k$$

is the partial sum to n, so that $s_n \to y$, we have by a simple calculation that

$$s_n(1-x) = (1-x)s_n = 1 - x^{n+1} \ .$$

The left side converges to $y(1-x) = (1-x)y$, and the right side to 1 since $x^{n+1} \to 0$. For the second statement, since $\|x^k\| \leq \|x\|^k$ for every k by submultiplicativity, we have that

$$\sum_{k=1}^{\infty} \|x^k\|$$

converges by comparison to the geometric series with ratio $\|x\|$, and hence $\sum_{k=0}^{\infty} x^k$ converges (). By the triangle inequality,

$$\left\| \sum_{k=0}^{\infty} x^k \right\| \leq \sum_{k=0}^{\infty} \|x^k\| \leq \sum_{k=0}^{\infty} \|x\|^k = \frac{1}{1-\|x\|} \ .$$

XX.12.2.10. Corollary. Let A be a unital Banach algebra, and $u \in A$. If $\|u-1\| < 1$, then u is invertible in A, and

$$\|u^{-1}\| \leq \frac{1}{1-\|u-1\|} \ .$$

Proof: Apply the theorem with $x = 1-u$.

XX.12.2.11. Corollary. Let A be a unital Banach algebra, and $u, v \in A$. If u is invertible in A and

$$\|u-v\| < \frac{1}{\|u^{-1}\|}$$

then v is also invertible in A, and

$$\|v^{-1}\| \leq \frac{\|u^{-1}\|}{1-\|u^{-1}\|\|u-v\|} \ .$$

Thus the open ball around u of radius $\frac{1}{\|u^{-1}\|}$ consists entirely of invertible elements. In particular, the set of invertible elements in A is an open set.

Proof: If $\|u-v\| < \frac{1}{\|u^{-1}\|}$, then

$$\|1-u^{-1}v\| = \|u^{-1}(u-v)\| \leq \|u^{-1}\|\|u-v\| < 1$$

so $u^{-1}v$ is invertible and v is left invertible. Similarly, $\|1-vu^{-1}\| < 1$ so v is right invertible. We have

$$\|v^{-1}u\| = \|(u^{-1}v)^{-1}\| \leq \frac{1}{1-\|u^{-1}v-1\|}$$

so

$$\|v^{-1}\| = \|(v^{-1}u)u^{-1}\| \leq \|v^{-1}u\|\|u^{-1}\| \leq \frac{\|u^{-1}\|}{1-\|u^{-1}v-1\|} = \frac{\|u^{-1}\|}{1-\|u^{-1}(v-u)\|} \leq \frac{\|u^{-1}\|}{1-\|u^{-1}\|\|v-u\|} \ .$$

XX.12.2.12. Note that the bound

$$\|v^{-1}\| \leq \frac{\|u^{-1}\|}{1 - \|u^{-1}v - 1\|}$$

in the proof is generally better than the bound in the statement of the result; but the bound in the statement is usually easier to compute.

Continuity of Inversion

XX.12.2.13. If A is a unital algebra, write A^{-1} for the set of invertible elements of A. Note that A^{-1} is a group () under multiplication, since a product of invertible elements is invertible.

XX.12.2.14. Corollary. Let A be a unital Banach algebra. Then the map $x \mapsto x^{-1}$ is a continuous function from A^{-1} to A^{-1}.

Proof: By the sequential criterion, it suffices to show that if (u_n) is a sequence in A^{-1} converging to $u \in A^{-1}$, then $u_n^{-1} \to u^{-1}$. There is an N such that

$$\|u_n - u\| < \frac{1}{2\|u^{-1}\|}$$

for all $n \geq N$. Thus, for $n \geq N$, we have

$$\|u_n^{-1}\| \leq \frac{\|u^{-1}\|}{1 - \|u^{-1}\|\|u_n - u\|} < 2\|u^{-1}\| \ .$$

So, for $n \geq N$,

$$\|u_n^{-1} - u^{-1}\| = \|u_n^{-1}(u - u_n)u^{-1}\| \leq \|u_n^{-1}\|\|u_n - u\|\|u^{-1}\| < 2\|u^{-1}\|^2\|u_n - u\|$$

which goes to 0 as $n \to \infty$. 👍

XX.12.2.15. Corollary. If A is a unital Banach algebra, then A^{-1} (with the norm topology) is a topological group () under multiplication.

XX.12.2.16. If I is an ideal in a Banach algebra A, it follows from continuity of multiplication that the closure $\bar{I}$ is also an ideal. If I is a proper ideal, it is not immediately obvious that $\bar{I}$ is a proper ideal (i.e. I could be dense in A); in fact, this can happen if A is nonunital (). But it cannot happen if A is a unital Banach algebra:

XX.12.2.17. Corollary. Let A be a unital Banach algebra, and I proper ideal of A. Then $\bar{I}$ is also a proper ideal of A.

Proof: An ideal in a unital algebra is a proper ideal if and only if it contains no invertible elements. Thus if I is a proper ideal, it is contained in the set S of noninvertible elements of A. But S is closed in A, so $\bar{I} \subseteq S$ and $\bar{I}$ is a proper ideal. 👍

XX.12.3. Spectrum

We now define the crucial notion of the spectrum of an element of a Banach algebra. For reasons partially explained in XX.12.3.5., we will consider only *complex* Banach algebras. For simplicity, we will also only develop the theory for unital Banach algebras. Thus, in this section (except (XX.12.3.2.)), all Banach algebras will be unital and complex; subalgebras will be assumed to contain the multiplicative identity. All homomorphisms will be assumed unital.

XX.12.3.1. DEFINITION. Let A be a unital Banach algebra, $x \in A$. The *spectrum* of x in A is

$$\sigma_A(x) = \{\lambda \in \mathbb{C} : x - \lambda 1 \text{ is not invertible}\} .$$

The complement of $\sigma_A(x)$ in $\mathbb{C}$ is called the *resolvent set* of x in A, denoted $\rho_A(x)$.

Thus $\sigma_A(x)$ is a subset of $\mathbb{C}$, which depends on x, and in general also on A (XX.12.3.11.). In many instances the A is understood (or, more importantly, the spectrum is independent of the choice of A, e.g. ()), and we just write $\sigma(x)$ for $\sigma_A(x)$.

XX.12.3.2. If A is a nonunital Banach algebra and $x \in A$, we can define $\sigma_A(x) = \sigma_{A^\dagger}(x)$. We then have $0 \in \sigma_A(x)$ for every $x \in A$. Although much of the theory we develop works in the nonunital case with occasional technical complications, for simplicity we will not consider spectrum in a nonunital Banach algebra from now on.

Here are the two most important examples, which together underly and motivate spectral theory:

XX.12.3.3. EXAMPLES. (i) Let $\mathcal{X}$ be a complex Banach space, and $T \in \mathcal{B}(\mathcal{X})$. Then $\sigma(T) = \sigma_{\mathcal{B}(\mathcal{X})}(T)$ is the set of $\lambda \in \mathbb{C}$ such that $T - \lambda I$ is not invertible, i.e. is either not injective or not surjective (the Bounded Inverse Theorem implies that if $T - \lambda I$ is injective and surjective, then $(T - \lambda I)^{-1} \in \mathcal{B}(\mathcal{X})$). If $T - \lambda I$ is not injective, then λ is by definition an eigenvalue of T. Thus $\sigma(T)$ can be thought of as the set of *generalized eigenvalues* of T.

If $\mathcal{X}$ is finite-dimensional, then by () $T - \lambda I$ is surjective if and only if it is injective. Thus in this case $\sigma(T)$ is exactly the set of eigenvalues of T, a finite nonempty subset of $\mathbb{C}$.

(ii) Let X be a compact Hausdorff space, and $A = C(X)$. If $f \in A$, then f is invertible in A if and only if f never takes the value 0. Thus $f - \lambda 1$ is invertible if and only if f never takes the value λ, i.e. $\sigma_A(f)$ is precisely the range of f, a nonempty compact subset of $\mathbb{C}$.

XX.12.3.4. Note that the spectrum of an element of a Banach algebra is a purely algebraic concept, and the norm plays no role in its definition. We may more generally define the spectrum of an element in any unital algebra over $\mathbb{C}$ in an identical way. (Thus, if $\mathcal{V}$ is a finite-dimensional complex vector space and $T \in \mathcal{L}(\mathcal{V})$, then $\sigma(T)$ is the set of eigenvalues of T.) However, the interplay between the topological and algebraic structure in a Banach algebra has important consequences for the *properties* of the spectrum.

XX.12.3.5. Example XX.12.3.3.(i) shows one basic reason why we restrict to complex algebras: operators on real vector spaces can fail to have (real) eigenvalues since $\mathbb{R}$ is not algebraically closed, for example a rotation in $\mathbb{R}^2$. There are deeper reasons why we restrict to complex algebras, involving applications of Complex Analysis (e.g. XX.12.3.27.).

Topological Properties of the Spectrum and Resolvent

XX.12.3.6. PROPOSITION. Let A be a unital Banach algebra, and $x \in A$. Then $\sigma_A(x)$ is closed in $\mathbb{C}$, i.e. $\rho_A(x)$ is open in $\mathbb{C}$.

PROOF: This follows immediately from the fact that A^{-1} is open in A, since $\lambda \mapsto x - \lambda 1$ is a continuous function from $\mathbb{C}$ to A. 👍

XX.12.3.7. PROPOSITION. Let A be a unital Banach algebra, and $x \in A$. If $\lambda \in \sigma_A(x)$, then $|\lambda| \leq \|x\|$. Thus $\sigma_A(x)$ is bounded in $\mathbb{C}$, hence compact.

PROOF: The case $x = 0$ is trivial ($\sigma_A(0) = \{0\}$), so suppose $x \neq 0$. Let $\alpha \in \mathbb{C}$. If $|\alpha| < \frac{1}{\|x\|}$, then the series

$$\sum_{k=0}^{\infty} (\alpha x)^k$$

converges to an inverse for $1 - \alpha x$ by XX.12.2.9.. So if $\lambda \in \mathbb{C}$, $|\lambda| > \|x\|$, we have that $1 - \frac{1}{\lambda}x$ is invertible, and hence $x - \lambda 1 = -\lambda(1 - \frac{1}{\lambda}x)$ is also invertible and $\lambda \notin \sigma_A(x)$. 👍

XX.12.3.8. We will see later (XX.12.3.27.) that $\sigma_A(x)$ is always nonempty. A more careful analysis (XX.12.3.32.) along the lines of the argument in the proof of XX.12.3.7. will give a more precise description of the size of the set $\sigma_A(x)$.

XX.12.3.9. Let A be a unital Banach algebra, and $x \in A$. Then continuity of inversion implies that the function $R : \rho_A(x) \to A^{-1}$ given by $R(\lambda) = (\lambda 1 - x)^{-1} = -(x - \lambda 1)^{-1}$ is continuous. R is called the *resolvent function* of x, with $R(\lambda)$ usually denoted $R_\lambda(x)$; it is actually far more than just continuous in λ (XX.12.3.25.). (The sign convention for $R_\lambda(x)$ is not universal.)

Dependence of Spectrum on the Algebra

XX.12.3.10. The spectrum of an element can depend on which containing Banach algebra is considered: if A is a (unital) Banach algebra and B a (unital) closed subalgebra of A, hence itself a Banach algebra, and $x \in B$, then x can be invertible in A but not in B, i.e. we can have $x^{-1} \notin B$. Obviously $\sigma_A(x) \subseteq \sigma_B(x)$, but the containment can be proper:

XX.12.3.11. EXAMPLE. Let $\mathbb{T}$ be the unit circle in $\mathbb{C}$, and $A = C(\mathbb{T})$. Let f be the "identity function" $f(z) = z$ from $\mathbb{T}$ to $\mathbb{C}$ (note that the term "identity function" can be confusing here since f is not the multiplicative identity of A). Then $\sigma_A(f)$ is the range of f (XX.12.3.3.(ii)), i.e. $\sigma_A(f) = \mathbb{T}$.

Let B be the subset of A consisting of functions which extend to functions continuous on the closed unit disk $\bar{\mathbb{D}}$ and holomorphic on $\mathbb{D}$ (B is exactly the algebra $\bar{P}$ of Exercise XX.10.3.10.). By the Maximum Modulus Theorem (), such an extension, if it exists, is unique. By the Maximum Modulus Theorem and XII.4.1.2., B is a closed unital subalgebra of A, and $f \in B$. However, if $|\lambda| \leq 1$, the unique extension g_λ of

$f - \lambda 1$ to $\bar{\mathbb{D}}$ is $g_\lambda(z) = z - \lambda$, which is not invertible in B. If $|\lambda| > 1$, then $\frac{1}{g_\lambda}$ is holomorphic on $\mathbb{C} \setminus \{\lambda\}$, hence on $\bar{\mathbb{D}}$, so $f - \lambda 1$ is invertible in B. Thus $\sigma_B(f) = \bar{\mathbb{D}}$.

However, $\sigma_B(x)$ cannot be too much larger than $\sigma_A(x)$:

XX.12.3.12. PROPOSITION. Let A be a unital Banach algebra and B a closed unital subalgebra of A, and let $x \in B$. Then $\sigma_A(x) \subseteq \sigma_B(x)$ and $\partial(\sigma_B(x)) \subseteq \partial(\sigma_A(x)) \subseteq \sigma_A(x)$, where ∂ denotes topological boundary ().

PROOF: We have already made the trivial observation that $\sigma_A(x) \subseteq \sigma_B(x)$. Suppose $\lambda \in \partial(\sigma_B(x))$. Then $\lambda \in \sigma_B(x)$ since $\sigma_B(x)$ is closed, and there is a sequence (λ_n) in $\rho_B(x)$ with $\lambda_n \to \lambda$. Since $\rho_B(x) \subseteq \rho_A(x)$, $\lambda_n \in \rho_A(x)$ for all n, so either $\lambda \in \partial(\sigma_A(x))$ (if $\lambda \in \sigma_A(x)$) or $\lambda \in \rho_A(x)$. But if $\lambda \in \rho_A(x)$ (i.e. $x - \lambda 1$ is invertible in A), then $(x - \lambda_n 1)^{-1} \to (x - \lambda 1)^{-1}$ by XX.12.3.9., and $(x - \lambda_n 1)^{-1} \in B$ for all n, so $(x - \lambda 1)^{-1} \in B$ since B is closed in A, contradicting that $\lambda \in \sigma_B(x)$. Thus $\lambda \in \partial(\sigma_A(x))$. 👍

XX.12.3.13. Thus, if A is a unital Banach algebra and B a closed unital subalgebra of A, and $x \in B$, then $\sigma_B(x)$ is the union of $\sigma_A(x)$ and possibly some bounded connected components of the complement of $\sigma_A(x)$. If $\sigma_A(x)$ does not divide $\mathbb{C}$ (i.e. if the complement is connected), then necessarily $\sigma_B(x) = \sigma_A(x)$. This is also true if $\sigma_B(x)$ has empty interior.

Spectral Mapping Properties

XX.12.3.14. If f is a polynomial with complex coefficients, and x is an element of a unital algebra A, then there is an obvious way to apply f to x to obtain an element $f(x) \in A$: if

$$f(X) = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \cdots + \alpha_n X^n$$

then

$$f(x) = \alpha_0 1 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_n x^n \ .$$

If f and g are polynomials with complex coefficients, and $x \in A$, then $(f + g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$, i.e. the map $f \mapsto f(x)$ is a homomorphism () from the polynomial ring $\mathbb{C}[X]$ to A.

XX.12.3.15. PROPOSITION. [SPECTRAL MAPPING THEOREM FOR POLYNOMIALS] Let A be a unital algebra over $\mathbb{C}$, $x \in A$, and f a polynomial with complex coefficients. Then

$$\sigma_A(f(x)) = \{f(\lambda) : \lambda \in \sigma_A(x)\} \ .$$

PROOF: Let $\mu \in \mathbb{C}$. Since $\mathbb{C}$ is algebraically closed, the polynomial $f(X) - \mu$ factors into linear factors in $\mathbb{C}[X]$:

$$f(X) - \mu = (X - \lambda_1)(X - \lambda_2) \cdots (X - \lambda_n)$$

for some $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ (not necessarily distinct). We then have

$$f(x) - \mu 1 = (x - \lambda_1 1) \cdots (x - \lambda_n 1)$$

in A. Then by the next lemma $f(x) - \mu 1$ is invertible in A if and only if $x - \lambda_k 1$ is invertible in A for all k, i.e. $\mu \in \sigma_A(f(x))$ if and only if $\lambda_k \in \sigma_A(x)$ for some k. Since the numbers $\lambda_1, \ldots, \lambda_n$ are precisely the complex numbers λ such that $f(\lambda) = \mu$, the result follows. 👍

XX.12.3.16. Lemma. Let A be a unital ring (e.g. a unital algebra over a field $\mathbb{F}$), and $a_1, \ldots, a_n \in A$. If the a_k commute, then $a = a_1 a_2 \cdots a_n$ is invertible in A if and only if each a_k is invertible in A.

Proof: If the a_k are invertible, then a is invertible, and $a^{-1} = a_n^{-1} \cdots a_1^{-1}$ (this direction holds even if the a_k do not commute). Conversely, if the a_k commute and a is invertible, we have for each k

$$a_k(a_1 a_2 \cdots a_{k-1} a_{k+1} \cdots a_n a^{-1}) = 1$$

$$(a^{-1} a_1 a_2 \cdots a_{k-1} a_{k+1} \cdots a_n) a_k = 1$$

so a_k is both left and right invertible in A. 👍

Note that the conclusion can be false if the a_k do not commute ().

XX.12.3.17. Proposition. [Spectral Mapping Theorem for Inverses] Let A be a unital algebra over $\mathbb{C}$, and $x \in A$. If x is invertible in A (i.e. $0 \notin \sigma_A(x)$), then

$$\sigma_A(x^{-1}) = \left\{ \frac{1}{\lambda} : \lambda \in \sigma_A(x) \right\} .$$

Proof: Since x^{-1} is invertible, $0 \notin \sigma_A(x^{-1})$. If $\lambda \neq 0$, we have

$$x^{-1} - \frac{1}{\lambda} 1 = -\frac{1}{\lambda} x^{-1}(x - \lambda 1)$$

and x^{-1} and $x - \lambda 1$ commute, so $x^{-1} - \frac{1}{\lambda} 1$ is invertible if and only if $x - \lambda 1$ is invertible. 👍

XX.12.3.18. The two previous spectral mapping theorems are special cases of the following result. If A is a unital algebra over $\mathbb{C}$, $x \in A$, and f is a rational function in one variable with coefficients in $\mathbb{C}$ and no poles in $\sigma_A(x)$, so f defines a continuous function on $\sigma_A(x)$, write $f = \frac{p}{q}$ with p and q complex polynomials, in lowest terms, i.e. p and q have no common roots and q has no roots in $\sigma_A(x)$. Then $p(x)$ and $q(x)$ are defined in A, and $q(x)$ is invertible by XX.12.3.15.. We can then define $f(x) = p(x)q(x)^{-1} \in A$. (This is well defined since p and q are unique up to scalar multiples.) It is easily checked that $f \mapsto f(x)$ is an algebra homomorphism.

XX.12.3.19. PROPOSITION. [SPECTRAL MAPPING THEOREM FOR RATIONAL FUNCTIONS] Let A be a unital algebra over $\mathbb{C}$, $x \in A$, and f a rational function with complex coefficients and no poles in $\sigma_A(x)$. Then

$$\sigma_A(f(x)) = \{f(\lambda) : \lambda \in \sigma_A(x)\} .$$

PROOF: Let $\mu \in \mathbb{C}$. Then, for $z \in \mathbb{C}$,

$$f(z) - \mu = \frac{p(z)}{q(z)} - \mu = \frac{p(z) - \mu q(z)}{q(z)}$$

where p and q are polynomials with no common roots. Note that $q(z)$ and $p(z) - \mu q(z)$ also have no common roots. Let $\alpha_1, \ldots, \alpha_n$ be the roots of $p - \mu q$ and $\beta_1, \ldots, \beta_m$ the roots of q. We have

$$f(x) - \mu 1 = c(x - \alpha_1 1) \cdots (x - \alpha_n 1)(x - \beta_1 1)^{-1} \cdots (x - \beta_m 1)^{-1}$$

where c is a nonzero constant. Since the last m factors are invertible and the factors commute, $f(x) - \mu 1$ is invertible if and only if none of the α_k are in $\sigma_A(x)$. But the α_k are precisely the $z \in \mathbb{C}$ for which $p(z) - \mu q(z) = 0$, which are precisely the z for which $f(z) = \mu$ (since none of the β_k are roots of $p - \mu q$). 👍

Here is a cute fact related in spirit:

XX.12.3.20. PROPOSITION. Let A be a unital algebra over $\mathbb{C}$, $x, y \in A$. Then

$$\sigma_A(xy) \cup \{0\} = \sigma_A(yx) \cup \{0\} .$$

PROOF: Suppose $\lambda \neq 0$ and $xy - \lambda 1$ is invertible (i.e. $\lambda \notin \sigma_A(xy)$). Then, if $r = (xy - \lambda 1)^{-1}$, we have

$$(xy - \lambda 1)r = r(xy - \lambda 1) = 1$$

$$xyr = rxy = 1 - \lambda r$$

So we have

$$-\frac{1}{\lambda}(1 - yrx)(yx - \lambda 1) = -\frac{1}{\lambda}[yx - yrxyx + \lambda yrx - \lambda 1] = -\frac{1}{\lambda}[yx - y(1 - \lambda r)x + \lambda yrx - \lambda 1] = 1$$

and similarly $-\frac{1}{\lambda}(yx - \lambda 1)(1 - yrx) = 1$, so $yx - \lambda 1$ is invertible and $\lambda \notin \sigma_A(yx)$. The x and y can be interchanged in the argument. 👍

XX.12.3.21. So, while $0 \in \sigma_A(xy)$ (i.e. xy not invertible) does not imply $0 \in \sigma_A(yx)$ (i.e. yx not invertible), the rest of the two spectra are identical. If A is a Banach algebra and xy is invertible but yx is not invertible, then 0 is an isolated point of the spectrum of yx (since $\sigma_A(xy) = \sigma_A(yx) \setminus \{0\}$ is closed in $\mathbb{C}$).

Holomorphic Properties of the Resolvent

Recall (XX.12.3.9.) that if x is an element of a unital Banach algebra A, the resolvent function of x is defined to be

$$R_\lambda(x) = (\lambda 1 - x)^{-1}$$

for $\lambda \in \rho_A(x) \subseteq \mathbb{C}$. In what follows, we keep A and x fixed, and just write R_λ for $R_\lambda(x)$.

XX.12.3.22. PROPOSITION.

$$\lim_{|\lambda|\to\infty} R_\lambda = 0 \ .$$

PROOF: Note that since $\sigma_A(x)$ is bounded (XX.12.3.7.), R_λ is defined for $|\lambda|$ sufficiently large, so the limit potentially exists. If $|\lambda|$ is large, we have

$$R_\lambda = (\lambda 1 - x)^{-1} = \alpha(1 - \alpha x)^{-1}$$

where $\alpha = \frac{1}{\lambda}$. We have

$$\lim_{\alpha\to 0}(1 - \alpha x) = 1$$

and by XX.12.2.10. and XX.12.2.14. we also have

$$\lim_{\alpha\to 0}(1 - \alpha x)^{-1} = 1 \ .$$

Thus

$$\lim_{\alpha\to 0} \alpha(1 - \alpha x)^{-1} = 0 \ .$$



XX.12.3.23. PROPOSITION. Let $\lambda, \mu \in \rho_A(x)$. Then

$$R_\mu = (1 - y)^{-1} R_\lambda$$

where $y = (\lambda - \mu)R_\lambda$.

PROOF: We have

$$\mu 1 - x = (\lambda 1 - x) - (\lambda - \mu) = (\lambda 1 - x)(1 - y)$$

so the formula follows by taking inverses. 

Putting this together with XX.12.2.9., we obtain:

XX.12.3.24. COROLLARY. Let $\lambda_0 \in \rho_A(x)$, and $\lambda \in \mathbb{C}$. If the series

$$\sum_{k=0}^{\infty}(\lambda_0 - \lambda)^k R_{\lambda_0}^{k+1}$$

converges in A, then $\lambda \in \rho_A(x)$ and the sum is R_λ. In particular, if $|\lambda - \lambda_0| < \frac{1}{\|R_{\lambda_0}\|}$, $\lambda \in \rho_A(x)$ and the series converges to R_λ.

PROOF: Apply XX.12.3.23. with (λ, μ) replaced by (λ_0, λ), and apply XX.12.2.9. to $y = (\lambda_0 - \lambda)R_{\lambda_0}$. 👍

XX.12.3.25. Thus R_λ is locally representable as a power series, and thus is an "analytic function" of λ. There is a theory of Banach-space-valued analytic functions () which applies. However, we will need only the following weaker form which reduces consideration to complex-valued functions:

XX.12.3.26. COROLLARY. Let $\phi \in A^*$, and define $f_\phi(\lambda) = \phi(R_\lambda)$ for $\lambda \in \rho_A(x)$. Then f_ϕ is a holomorphic (analytic) complex-valued function on $\rho_A(x)$.

PROOF: Fix $\lambda_0 \in \rho_A(x)$. If $|\lambda - \lambda_0| < \frac{1}{\|R_{\lambda_0}\|}$, then the infinite series

$$\sum_{k=0}^{\infty}\phi(R_{\lambda_0}^{k+1})(\lambda_0 - \lambda)^k = \sum_{k=0}^{\infty}(-1)^k\phi(R_{\lambda_0}^{k+1})(\lambda - \lambda_0)^k$$

converges to $f_\phi(\lambda)$ by XX.12.3.24. and continuity of ϕ. Thus f_ϕ is locally representable by a convergent power series. 👍

The Spectrum is Nonempty!

The first major consequence is:

XX.12.3.27. THEOREM. Let A be a unital Banach algebra, and $x \in A$. Then $\sigma_A(x)$ is nonempty.

PROOF: Suppose $\sigma_A(x) = \emptyset$, so $\rho_A(x) = \mathbb{C}$. Let $\phi \in A^*$. Then f_ϕ is an entire function (). We have

$$\lim_{|\lambda|\to\infty} f_\phi(\lambda) = 0$$

by XX.12.3.22. and continuity of ϕ. Thus f_ϕ is bounded, hence constant by Liouville's Theorem (XII.3.3.10.). Since the limit at infinity is 0, f_ϕ is identically 0, i.e. $\phi(R_\lambda) = 0$ for all $\lambda \in \mathbb{C}$ and all $\phi \in A^*$. By () this implies that $R_\lambda = 0$ for all λ, a contradiction. 👍

A simple corollary is:

XX.12.3.28. Corollary. [Gelfand-Mazur Theorem] The only (complex) Banach division algebra is $\mathbb{C}$.

Proof: Recall that A is a division algebra if every nonzero element is invertible. If $x \in A$, there is a $\lambda \in \sigma_A(x)$. For this λ, $x - \lambda 1$ is not invertible; hence $x - \lambda 1 = 0$, $x = \lambda 1$. Thus every element of A is a scalar multiple of 1. 👍

Spectral Radius

Let x be an element of a unital Banach algebra A. Since $\sigma_A(x)$ is a nonempty compact subset of the plane, it makes sense to make the following definition:

XX.12.3.29. Definition. The *spectral radius* of x is

$$r_\sigma(x) = \sup_{\lambda \in \sigma_A(x)} |\lambda| = \max_{\lambda \in \sigma_A(x)} |\lambda| \ .$$

XX.12.3.30. The notation $r_\sigma(x)$ does not reflect the choice of the Banach algebra A containing x, on which the spectral radius potentially depends. But it is an immediate consequence of XX.12.3.12. that if B is a unital Banach subalgebra of A and $x \in B$, then the same spectral radius for x is obtained whether $\sigma_A(x)$ or $\sigma_B(x)$ is used in the definition, so $r_\sigma(x)$ in fact does not depend on which Banach algebra containing x is used. We could indeed just use the *smallest* Banach subalgebra of A containing x, i.e. the closed linear span of $\{x^n : n \geq 0\}$.

XX.12.3.31. Note also that the spectral radius of x depends only on the *algebraic* structure of A and not on the norm. However, we needed to know that A was a Banach algebra with respect to *some* norm in order to conclude that $\sigma_A(x)$ is nonempty and bounded so that the spectral radius is well defined.

There is nonetheless a tight relationship between the spectral radius and norm:

XX.12.3.32. Theorem. Let A be a Banach algebra and $x \in A$. Then

$$r_\sigma(x) = \lim_{n \to \infty} \|x^n\|^{1/n} = \inf_n \|x^n\|^{1/n} \ .$$

The theorem will be proved in several steps.

XX.12.3.33. Part of the content of the theorem is that

$$\lim_{n \to \infty} \|x^n\|^{1/n}$$

exists and equals

$$\inf_n \|x^n\|^{1/n} \ .$$

This can be proved directly using IV.1.12.5., setting $\alpha_n = \|x^n\|$, since $\|x^{n+m}\| = \|x^n x^m\| \leq \|x^n\|\|x^m\|$ for any $n, m \in \mathbb{N}$ by submultiplicativity. It will not be necessary for us to prove this since it will follow automatically from our other arguments. The sequence $(\|x^n\|^{1/n})$ is not necessarily nonincreasing ().

The first observation for the proof of XX.12.3.32. is that $r_\sigma(x) \leq \|x\|$ by XX.12.3.7.. We can do better by noting the following consequence of the Spectral Mapping Theorem for polynomials (XX.12.3.15.):

XX.12.3.34. Proposition. Let x be an element of a unital Banach algebra A, and f a polynomial with complex coefficients. Then

$$r_\sigma(f(x)) = \max_{\lambda \in \sigma_A(x)} |f(\lambda)| \ .$$

In particular, $r_\sigma(x^n) = [r_\sigma(x)]^n$ for any $n \in \mathbb{N}$.

XX.12.3.35. Corollary. Let x be an element of a unital Banach algebra A. Then

$$r_\sigma(x) \leq \inf_n \|x^n\|^{1/n} \ .$$

Proof: By XX.12.3.34. and XX.12.3.7. we have, for any n,

$$[r_\sigma(x)]^n = r_\sigma(x^n) \leq \|x^n\| \ .$$



XX.12.3.36. So to finish the proof of Theorem XX.12.3.32., it suffices to show that

$$r_\sigma(x) \geq \limsup_{n \to \infty} \|x^n\|^{1/n} \ .$$

This will take some work using Complex Analysis. Our argument follows [Ped89]; roughly speaking, it is based on XII.4.4.6..

XX.12.3.37. Fix $\phi \in A^*$ and let $f_\phi(\lambda) = \phi(R_\lambda)$. Then f_ϕ is analytic on $\rho_A(x)$, and in particular on $\Omega = \{\lambda \in \mathbb{C} : |\lambda| > r_\sigma(x)\}$. We have that

$$f_\phi(\lambda) = \frac{1}{\lambda}\phi\left(\left[1 - \frac{1}{\lambda}x\right]^{-1}\right) = \frac{1}{\lambda}\phi\left(\sum_{k=0}^{\infty}\left[\frac{1}{\lambda}x\right]^k\right) = \frac{1}{\lambda}\sum_{k=0}^{\infty}\phi\left(\left[\frac{1}{\lambda}x\right]^k\right) = \sum_{k=0}^{\infty}\lambda^{-k-1}\phi(x^k)$$

for $|\lambda| > \|x\|$; thus this is the Laurent series for f_ϕ on Ω and converges to f_ϕ u.c. on Ω (XII.4.4.12.).

XX.12.3.38. Now fix $s > r_\sigma(x)$. For any $n \in \mathbb{N}$, the series

$$\sum_{k=0}^{\infty} \lambda^{n-k} \phi(x^k)$$

converges uniformly to $g_n(\lambda) = \lambda^{n+1} f_\phi(\lambda)$ on the circle γ of radius s around 0, so we can compute the integral of g_n around γ by integrating term-by-term:

$$\int_0^{2\pi} s^{n+1} e^{i(n+1)\theta} \phi(se^{i\theta})\, d\theta = \sum_{k=0}^{\infty} \left[\int_0^{2\pi} s^{n-k} e^{i(n-k)\theta} \phi(x^k)\, d\theta \right] = \int_0^{2\pi} \phi(x^n)\, d\theta = 2\pi\phi(x^n)$$

since the integrals for $k \neq n$ are zero. Thus we obtain that

$$|\phi(x^n)| \leq s^{n+1} M(s) \|\phi\|$$

where $M(s) = \max_{|\lambda|=s} \|R_\lambda\|$. Since $\phi \in A^*$ is arbitrary, by the Hahn-Banach Theorem () we conclude that

$$\|x^n\| \leq s^{n+1} M(s)$$

$$\|x^n\|^{1/n} \leq s^{\frac{n+1}{n}} M(s)^{1/n}$$

for every n. Taking lim sup as $n \to \infty$, we obtain

$$\limsup_{n\to\infty} \|x^n\|^{1/n} \leq s \ .$$

This holds for all $s > r_\sigma(x)$, so we conclude

$$\limsup_{n\to\infty} \|x^n\|^{1/n} \leq r_\sigma(x) \ .$$

This completes the proof of Theorem XX.12.3.32.. 

XX.12.4. Holomorphic Functional Calculus

XX.12.4.1. If $f(t) = \alpha_1 t + \cdots + \alpha_n t^n$ is a polynomial with complex coefficients, without constant term (i.e. $f(0) = 0$), and x is an element of an algebra A over $\mathbb{C}$, then there is an obvious way to apply f to x to obtain an element $f(x) = \alpha_1 x + \cdots + \alpha_n x^n \in A$. If A is a Banach algebra, there is a very important way of extending this procedure to holomorphic functions, called *functional calculus*.

XX.12.4.2. If X is a compact subset of $\mathbb{C}$, denote by $H(X)$ the algebra of functions holomorphic in a neighborhood of X and vanishing at 0 if $0 \in X$, with functions identified if they agree on a neighborhood of X. Functional calculus gives a homomorphism from $H(\sigma_A(x))$ to the Banach subalgebra of A generated by x extending the map for polynomials. The image of f is denoted $f(x)$. The element $f(x)$ can be defined using the Cauchy Integral Formula, but in some cases (e.g. if f is entire) it is also given by a power series. If A is unital, functional calculus is also defined for holomorphic functions not vanishing at zero. Functional calculus has the following properties, which (along with the elementary definition for polynomials) determine it uniquely:

XX.12.4.3. PROPOSITION. Let A be a Banach algebra and $x \in A$. Then

(i) For any $f \in H(\sigma_A(x))$, $\sigma_A(f(x)) = \{f(\lambda) : \lambda \in \sigma_A(x)\}$.

(ii) If $f \in H(\sigma_A(x))$ and $g \in H(f(\sigma_A(x))) = H(\sigma_A(f(x)))$, so $g \circ f \in H(\sigma_A(x))$, then $(g \circ f)(x) = g(f(x))$.

(iii) If $f_n, f \in H(\sigma_A(x))$ and $f_n \to f$ uniformly on a neighborhood of $\sigma_A(x)$, then $f_n(x) \to f(x)$.

(iv) If B is a Banach algebra and $\phi : A \to B$ a continuous (bounded) homomorphism, then $\phi(f(x)) = f(\phi(x))$ for any $f \in H(\sigma_A(x))$.

PROOF: (i): Suppose A is unital. Let f be analytic on $\mathcal{U}$ containing $\sigma_A(x)$. If $\lambda \in \sigma_A(x)$, then $f(z) - f(\lambda) = (z - \lambda)g(z)$ with g analytic on $\mathcal{U}$; then $f(x) - f(\lambda)1 = (x - \lambda 1)g(x)$, and since $x - \lambda 1$ and $g(x)$ commute and $x - \lambda 1$ is not invertible, $f(\lambda) \in \sigma_A(f(x))$. Conversely, if $\mu \notin \{f(\lambda) : \lambda \in \sigma_A(x)\}$, then $h(z) = (f(z) - \mu)^{-1}$ is analytic on $\{z \in \mathcal{U} : f(z) \neq \mu\}$, which contains $\sigma_A(x)$, and $h(x) = (f(x) - \mu 1)^{-1}$.
(ii)-(iv) are straightforward. 👍

XX.12.4.4. An especially important case of functional calculus uses the exponential function $f(z) = e^z$. If x is any element of a unital Banach algebra A, then $f(x)$ is defined and denoted e^x. The element e^x is given by the power series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$. For any x, e^x is invertible, with inverse e^{-x}; $e^{x+y} = e^x e^y$ if x and y commute. If A is a Banach *-algebra, then $(e^x)^* = e^{(x^*)}$. Conversely, if $\sigma_A(x)$ is contained in a simply connected open set not containing 0, let $f(z)$ be a branch of the logarithm holomorphic in a neighborhood of $\sigma_A(x)$; then $y = f(x)$ satisfies $e^y = x$. In particular, if $\|1 - x\| < 1$, then there is a $y \in A$ with $\|y\| < \pi/2$ with $e^y = x$ (use the principal branch of log).

XX.12.4.5. If A is a unital Banach algebra, write A^{-1} (also often written $\mathrm{GL}_1(A)$) for the (open) set of invertible elements in A, and A_o^{-1} (or $\mathrm{GL}_1(A)_o$) for the connected component of the identity in A^{-1}. Then

$$exp(A) = \{e^y : y \in A\} \subseteq A^{-1}$$

is path-connected, and by the above the subgroup of A^{-1} generated algebraically by $exp(A)$ is a connected open subgroup, hence equal to A_o^{-1}. In particular, A_o^{-1} is an open subgroup of A^{-1}, and every element of A_o^{-1} is a finite product of exponentials.

If A and B are unital Banach algebras and ϕ is a bounded homomorphism from A onto B, and $y \in B^{-1}$, then there is not necessarily an $x \in A^{-1}$ with $\phi(x) = y$:

XX.12.4.6. EXAMPLE. Let $\mathbb{D}$ be the closed unit disk in $\mathbb{C}$, and $\mathbb{T}$ its boundary. There is a *-homomorphism $\phi : C(\mathbb{D}) \to C(\mathbb{T})$ given by restriction. But if $g \in C(\mathbb{T})$ with $g(z) = z$, then $g \in C(\mathbb{T})^{-1}$, but there is no $f \in C(\mathbb{D})^{-1}$ with $\phi(f) = g$.

The image in the Calkin algebra of a Fredholm operator of nonzero index is another example.
However:

XX.12.4.7. PROPOSITION. If $\phi : A \to B$ is a surjective bounded homomorphism of unital Banach algebras, then $\phi(A_o^{-1}) = B_o^{-1}$.

This can be proved either by an application of the Open Mapping Theorem or by noting that if $x \in A$, then $\phi(e^x) = e^{\phi(x)}$.

Homomorphisms

We have a generalization of XX.12.3.10.:

XX.12.4.8. PROPOSITION. Let A and B be unital Banach algebras, and $\phi : A \to B$ a unital homomorphism. If $x \in A$, then $\sigma_B(\phi(x)) \subseteq \sigma_A(x)$.

PROOF: If $y \in A$ is invertible in A, then $\phi(y)$ is invertible in B $[\phi(y)^{-1} = \phi(y^{-1})]$. 👍

XX.12.5. Spectral Theory for Bounded Operators

In this section, we examine the spectrum and its properties for bounded operators on Banach spaces, concentrating almost exclusively on the case of operators on Hilbert spaces.

We will essentially follow the first comprehensive development and exposition of this theory, given by F. Riesz [?] in 1913, abstracted to Banach spaces and Hilbert spaces (which did not exist in 1913!) and with some modern technical simplifications.

XX.12.5.1. If $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ ($\mathcal{X}, \mathcal{Y}$ Banach spaces) is invertible in the algebraic sense, i.e. T is one-to-one and onto, then T^{-1} is automatically bounded by the Open Mapping Theorem. So there are two ways a bounded operator T can fail to be invertible:

T is not bounded below (recall that T is *bounded below* if $\exists \epsilon > 0$ such that $\|Tx\| \geq \epsilon \|x\|$ for all $x \in \mathcal{X}$)

$\mathcal{R}(T)$ is not dense.

These possibilities are not mutually exclusive. For example, if T is a normal operator on a Hilbert space, then $\mathcal{N}(T) = \mathcal{R}(T)^{\perp}$ (), so if $\mathcal{R}(T)$ is not dense, then T is not one-to-one and hence not bounded below. Thus a normal operator on a Hilbert space is invertible if and only if it is bounded below.

XX.12.5.2. Proposition. Let T be a bounded operator on a Hilbert space $\mathcal{H}$. Then T is invertible if and only if both T and T^* are bounded below.

Proof: If T is invertible, then for any x we have $\|x\| = \|T^{-1}(Tx)\| \leq \|T^{-1}\|\|Tx\|$, so T is bounded below. Since T^* is also invertible, T^* is bounded below too.

Conversely, suppose T and T^* are bounded below. In particular, T^* is one-to-one, so $\mathcal{N}(T^*) = \mathcal{R}(T)^{\perp} = \{0\}$, and thus $\mathcal{R}(T)$ is dense in $\mathcal{H}$. If $y \in \mathcal{H}$, then there is a sequence (y_n) in $\mathcal{R}(T)$ with $y_n \to y$. For each n there is an $x_n \in \mathcal{H}$ (unique, since T is one-to-one) with $y_n = Tx_n$. Fix $\epsilon > 0$ such that $\|Tx\| \geq \epsilon \|x\|$ for all $x \in \mathcal{H}$. Then, for any n and m,

$$\|x_n - x_m\| \leq \epsilon^{-1} \|Tx_n - Tx_m\| = \epsilon^{-1} \|y_n - y_m\|$$

and since (y_n) is a Cauchy sequence, so is (x_n). Let $x_n \to x$. We then have $Tx_n \to Tx$ since T is continuous (bounded), so $y = Tx$ and T is surjective. T^{-1} is automatically bounded by the Bounded Inverse Theorem. ☝

Recall the following definition from XX.12.3.1., rewritten in the context of operators and expanded:

XX.12.5.3. Definition. Let $\mathcal{X}$ be a normed vector space, $T \in \mathcal{B}(\mathcal{X})$. The *spectrum* of T is

$$\sigma(T) = \{\lambda \in \mathbb{C} : T - \lambda I \text{ is not invertible}\}$$

The *point spectrum* of T is

$$\sigma_p(T) = \{\lambda \in \mathbb{C} : T - \lambda I \text{ is not one-to-one}\}$$

(thus $\sigma_p(T)$ is exactly the set of eigenvalues of T). The *continuous spectrum* $\sigma_c(T)$ of T is the set of $\lambda \in \mathbb{C}$ such that $T - \lambda I$ is one-to-one with dense range, but not bounded below. The *residual spectrum* $\sigma_r(T)$ is the set of $\lambda \in \mathbb{C}$ such that $T - \lambda I$ is one-to-one but does not have dense range.

Note that $\sigma(T)$ is the disjoint union of $\sigma_p(T)$, $\sigma_c(T)$, and $\sigma_r(T)$. One or more of these sets can be empty (XX.12.5.8.). But we have the fundamental fact from XX.12.3.6., XX.12.3.27., and XX.12.3.32.:

XX.12.5.4. THEOREM. Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{B}(\mathcal{X})$. Then

(i) $\sigma(T)$ is a nonempty compact subset of $\mathbb{C}$.

(ii) $r_\sigma(T) = \max_{\lambda\in\sigma(T)} |\lambda| = \lim_{n\to\infty} \|T^n\|^{1/n} = \inf_n \|T^n\|^{1/n}$.

XX.12.5.5. The spectrum of T can be thought of as the set of "generalized eigenvalues" of T (note that every actual eigenvalue of T is in the spectrum). The first definition (not the same as the one here – essentially the reciprocal) and name "spectrum" were given by HILBERT; the modern definition is due to F. RIESZ.

XX.12.5.6. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$. Then $\sigma(T^*) = \{\bar\lambda : \lambda \in \sigma(T)\}$.

PROOF: This follows immediately from the fact that if $S \in \mathcal{B}(\mathcal{H})$, then S is invertible if and only if S^* is invertible, and $(T + \lambda I)^* = T^* + \bar\lambda I$. 👍

The relation between $\sigma_p(T)$, $\sigma_c(T)$, $\sigma_r(T)$ and $\sigma_p(T^*)$, $\sigma_c(T^*)$, $\sigma_r(T^*)$ is slightly more complicated:

XX.12.5.7. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$. If $\lambda \in \mathbb{C}$, then

(i) If $\lambda \in \sigma_p(T)$, then $\bar\lambda \in \sigma_p(T^*) \cup \sigma_r(T^*)$.

(ii) If $\lambda \in \sigma_r(T)$, then $\bar\lambda \in \sigma_p(T^*)$.

(iii) $\lambda \in \sigma_p(T) \cup \sigma_r(T)$ if and only if $\bar\lambda \in \sigma_p(T^*) \cup \sigma_r(T^*)$.

(iv) $\lambda \in \sigma_c(T)$ if and only if $\bar\lambda \in \sigma_c(T^*)$.

PROOF: Parts (i) and (ii) follow immediately from the relation $\mathcal{N}(S) = \mathcal{R}(S^*)^\perp$; (iii) follows trivially from (i) and (ii), and (iv) follows from (iii) and XX.12.5.6. by process of elimination. 👍

XX.12.5.8. EXAMPLES. (i) If $\mathcal{X}$ is finite-dimensional, then $\sigma(T) = \sigma_p(T)$ is the set of eigenvalues of T.
(ii) If T is a normal operator on a Hilbert space, then $\mathcal{N}(T) = \mathcal{R}(T)^\perp$ (), so if $\mathcal{R}(T)$ is not dense, then T is not one-to-one and hence not bounded below. Thus a normal operator on a Hilbert space is invertible if and only if it is bounded below, and $\sigma_r(T) = \emptyset$.
(iii) Let (α_n) be a bounded sequence in $\mathbb{C}$. Define $T : \ell^2 \to \ell^2$ by

$$T(\xi_1, \xi_2, \dots) = (\alpha_1\xi_1, \alpha_2\xi_2, \dots) \ .$$

Then $T \in \mathcal{B}(\ell^2)$, $\|T\| = \sup_n |\alpha_n|$. We have $\sigma_p(T) = \{\alpha_n : n \in \mathbb{N}\}$, $\sigma_r(T) = \emptyset$ since T is normal, and $\sigma(T)$ is the closure of $\sigma_p(T)$, so if $\sigma_p(T)$ is not closed we have $\sigma_c(T) = \sigma(T) \setminus \sigma_p(T) \neq \emptyset$.

(iv) Let T be the *unilateral shift* on ℓ^2:

$$T(\xi_1, \xi_2, \dots) = (0, \xi_1, \xi_2, \dots) .$$

Then T is an isometry, hence $\|T\| = 1$. T is one-to-one, but $\mathcal{R}(T)$ is not dense in ℓ^2; thus $0 \in \sigma_r(T)$. It is easy to see that $\sigma_p(T) = \emptyset$. By XX.12.5.6., XX.12.5.7., and (v) below, $\sigma(T)$ is the closed unit disk $\bar{\mathbb{D}}$ in $\mathbb{C}$, $\sigma_r(T)$ is the open unit disk, and $\sigma_c(T)$ is the unit circle.

(v) Let T^* be the adjoint of the unilateral shift of (iv). Then T^* is the *left shift*:

$$T^*(\xi_1, \xi_2, \dots) = (\xi_2, \xi_3, \dots) .$$

If $\lambda \in \mathbb{C}$, $|\lambda| < 1$, then the vector $x_\lambda = (1, \lambda, \lambda^2, \dots)$ is in ℓ^2 and $T^*(x_\lambda) = \lambda x_\lambda$; thus $\lambda \in \sigma_p(T^*)$. It is easy to see that no complex number of absolute value 1 can be an eigenvalue of T^*, so $\sigma_p(T^*)$ is the open unit disk $\mathbb{D}$. Thus $\sigma(T^*)$ contains $\bar{\mathbb{D}}$ since it is closed. But $\|T^*\| = 1$, so $r_\sigma(T^*) \le 1$ and $\sigma(T^*) \subseteq \bar{\mathbb{D}}$. Thus $\sigma(T^*) = \bar{\mathbb{D}}$. By XX.12.5.6., XX.12.5.7., and (iv), $\sigma_c(T^*)$ is the unit circle and $\sigma_r(T^*) = \emptyset$.

Numerical Range

XX.12.5.9. The *numerical range* of an operator T on a Hilbert space $\mathcal{H}$ is

$$W(T) = \{\langle Tx, x\rangle : \|x\| = 1\}.$$

Set

$$w(T) = \sup\{|\lambda| : \lambda \in W(T)\}.$$

It is obvious that $W(T^*) = \{\bar{\lambda} : \lambda \in W(T)\}$, so $w(T^*) = w(T)$; if $T = T^*$, then $W(T) \subseteq \mathbb{R}$. (Conversely, if $W(T) \subseteq \mathbb{R}$, then $T = T^*$: for $\xi, \eta \in \mathcal{H}$, expand $\langle T(\xi + \eta), \xi + \eta\rangle \in \mathbb{R}$ and $\langle T(\xi + i\eta), \xi + i\eta\rangle \in \mathbb{R}$.)

XX.12.5.10. Proposition. If $T \in \mathcal{B}(\mathcal{H})$, then $\sigma(T) \subseteq \overline{W(T)}$ and $r_\sigma(T) \le w(T) \le \|T\|$.

Proof: To see that $\sigma(T) \subseteq \overline{W(T)}$, note that if $\lambda \in \sigma(T)$, then by XX.12.5.2. either $T - \lambda I$ or $(T - \lambda I)^*$ is not bounded below, so there is a sequence (x_n) of unit vectors such that either $(T - \lambda I)x_n \to 0$ or $(T - \lambda I)^* x_n \to 0$. In either case, we obtain that $\langle (T - \lambda I)x_n, x_n\rangle \to 0$ by the CBS inequality. The inequality $r_\sigma(T) \le w(T)$ follows immediately, and $w(T) \le \|T\|$ is immediate from the CBS inequality. ☝

XX.12.5.11. Examples. (i) Note that $r_\sigma(T) = w(T) = \|T\|$ does not hold for general operators: for example, let T be the (nilpotent) operator on $\mathbb{C}^2$ given by the matrix $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$. Then $\sigma(T) = \{0\}$, so $r_\sigma(T) = 0$; but $\|T\| = 1$. It can be shown by an easy calculation that $W(T) = \{\lambda : |\lambda| \le 1/2\}$, so $w(T) = 1/2$.

(ii) If $\mathcal{H}$ is finite-dimensional and $T \in \mathcal{B}(\mathcal{H})$, then $W(T)$ is closed since the unit sphere in $\mathcal{H}$ is compact, and thus $\sigma(T) \subseteq W(T)$. But if $\mathcal{H}$ is infinite-dimensional, neither statement is true in general. Let $T \in \mathcal{B}(\ell^2)$ be defined by

$$T(\xi_1, \xi_2, \dots) = (\xi_1, \xi_2/2, \dots, \xi_n/n, \dots) .$$

Then $\sigma(T) = \{0\} \cup \{\frac{1}{n} : n \in \mathbb{N}\}$. But $\langle Tx, x\rangle \ne 0$ if $x \ne 0$, so $0 \notin W(T)$. It is not hard to compute that $W(T) = (0, 1]$.

XX.12.5.12. The expression $\langle Tx, x\rangle$ is not linear in x, and anyway the set of unit vectors is not convex. However, perhaps remarkably, the numerical range $W(T)$ *is* always convex (Hausdorff-Toeplitz Theorem). See XX.12.6.2. for a proof. It is also true that $w(T)$ is a norm on $\mathcal{B}(\mathcal{H})$ for a Hilbert space $\mathcal{H}$ (XX.12.6.3.).

Self-Adjoint and Positive Operators

XX.12.5.13. Definition. A (necessarily self-adjoint; cf. XX.12.6.4.) bounded operator T on a Hilbert space $\mathcal{H}$ is *positive* if $\langle Tx, x\rangle \geq 0$ for all $x \in \mathcal{H}$ (i.e. $W(T) \subseteq [0, +\infty)$). Write $T \geq 0$ if T is positive, and write $\mathcal{B}(\mathcal{H})_+$ for the set of positive operators on $\mathcal{H}$.

XX.12.5.14. If $T \in \mathcal{B}(\mathcal{H})$, then $T^*T \geq 0$ (this is also true if T is conjugate-linear); if $T = T^*$, then $\lambda I - T \geq 0$ for $\lambda \geq \|T\|$. Sums and nonnegative scalar multiples of positive operators are positive.

If S, T are self-adjoint operators, write $S \leq T$ if $T - S \geq 0$, i.e. if $\langle Sx, x\rangle \leq \langle Tx, x\rangle$ for all x. If $T = T^*$, we have $-w(T)I \leq T \leq w(T)I$, and in particular $-\|T\|I \leq T \leq \|T\|I$.

XX.12.5.15. Let S be a positive operator, and let m and M be the inf and sup of $W(S)$ respectively. The form $(x, y) = \langle Sx, y\rangle$ is positive semidefinite, and thus satisfies the CBS inequality, yielding, for all x, y,

$$|\langle Sx, y\rangle|^2 \leq \langle Sx, x\rangle\langle Sy, y\rangle \leq \langle Sx, x\rangle M\|y\|^2 \leq M^2\|x\|^2\|y\|^2.$$

Setting $y = Sx$ and dividing through by $\|Sx\|^2$, we obtain, for all x,

$$\langle S^2x, x\rangle = \langle Sx, Sx\rangle = \|Sx\|^2 \leq M\langle Sx, x\rangle \leq M^2\|x\|^2.$$

Since $0 \leq M \leq \|S\|$, we conclude:

XX.12.5.16. Proposition. Let S be a positive operator, and let m and M be the inf and sup of $W(S)$ respectively. Then

(i) $\|S\| = M$ and $S^2 \leq MS$.

(ii) If S is bounded below, then $m > 0$.

[For (ii), if $\|Sx\| \geq \epsilon\|x\|$ for all x, then $M > 0$, and for $\|x\| = 1$ we have

$$\epsilon^2 = \epsilon^2\|x\|^2 \leq \langle Sx, Sx\rangle \leq M\langle Sx, x\rangle$$

so $m \geq \frac{\epsilon^2}{M}$. In fact, $\epsilon = m$ works and this is the largest possible ϵ.]

XX.12.5.17. Proposition. If T is a self-adjoint operator on a Hilbert space H, and m and M are the inf and sup of $W(T)$ respectively, then $\sigma(T) \subseteq [m, M] \subseteq \mathbb{R}$ and $m, M \in \sigma(T)$. Thus $r_\sigma(T) = w(T)$.

Proof: $\sigma(T) \subseteq [m, M]$ by XX.12.5.10.. Note that $mI \leq T \leq MI$. Applying XX.12.5.16.(ii) to $T - mI$ and $MI - T$, we obtain $m, M \in \sigma(T)$. 👍

XX.12.5.18. PROPOSITION. Let T be a self-adjoint element of $\mathcal{B}(\mathcal{H})$. Then $r_\sigma(T) = \|T\| (= w(T))$.

PROOF: By scaling, we may assume $r_\sigma(T) = w(T) = 1$, so $-I \leq T \leq I$. Set $S = (T+I)/2$; then $0 \leq S \leq I$. By XX.12.5.16., $S^2 \leq S$, so $I - T^2 = 4(S - S^2) \geq 0$, $0 \leq T^2 \leq I$, $\|T^2\| = \|T\|^2 \leq 1$ by () and XX.12.5.16.. 👍

XX.12.5.19. In particular, if $T \geq 0$, then $\sigma(T) \subseteq [0, \|T\|]$ and $\|T\| \in \sigma(T)$.

XX.12.5.20. The result of XX.12.5.18. also follows easily from the spectral radius formula, since if T is self-adjoint we have $\|T^2\| = \|T\|^2$ (), so by iteration $\|T^{2^n}\| = \|T\|^{2^n}$ and thus

$$r_\sigma(T) = \lim_{m\to\infty} \|T^m\|^{1/m} = \lim_{n\to\infty} \|T^{2^n}\|^{1/2^n} = \|T\| \ .$$

However, the proof in XX.12.5.18. is more elementary, not requiring ().

We can by a similar argument extend XX.12.5.18.:

XX.12.5.21. PROPOSITION. Let $\mathcal{H}$ be a Hilbert space, and $T \in \mathcal{B}(\mathcal{H})$. If T is normal, then $r_\sigma(T) = w(T) = \|T\|$.

PROOF: We have $\|T\|^2 = \|T^*T\|$ (), and for each n, $\|T^n\|^2 = \|(T^n)^*T^n\| = \|(T^*T)^n\|$ since T is normal. But T^*T is self-adjoint, so $\|(T^*T)^{2^n}\| = \|T^*T\|^{2^n} = \|T\|^{2^{n+1}}$. Thus we have

$$r_\sigma(T) = \lim_{n\to\infty} \|T^{2^n}\|^{1/2^n} = \lim_{n\to\infty} [\|(T^*T)^{2^n}\|^{1/2}]^{1/2^n} = \lim_{n\to\infty} \|(T^*T)^{2^n}\|^{1/2^{n+1}} = \|T\| \ .$$

👍

XX.12.5.22. PROPOSITION. Let T be a positive operator on a Hilbert space $\mathcal{H}$, and $x, y \in \mathcal{H}$. Then $\langle T(x+y), x+y\rangle \leq [\langle Tx, x\rangle^{1/2} + \langle Ty, y\rangle^{1/2}]^2$.

PROOF: We have $\langle T(x+y), x+y\rangle = \langle Tx, x\rangle + 2Re\langle Tx, y\rangle + \langle Ty, y\rangle$, and by the CBS inequality $Re\langle Tx, y\rangle \leq [\langle Tx, x\rangle\langle Ty, y\rangle]^{1/2}$. 👍

XX.12.5.23. COROLLARY. Let T be a positive operator on $\mathcal{H}$, and $\{x_i\}$ a total set of vectors in $\mathcal{H}$. If $\langle Tx_i, x_i\rangle = 0$ for all i, then $T = 0$.

XX.12.5.24. If X is a compact subset of $\mathbb{C}$ not containing 0, and $k \in \mathbb{N}$, there is in general no bound on the norm of T^{-1} as T ranges over all operators with $\|T\| \leq k$ and $\sigma(T) \subseteq X$. For example, let $S_n \in \mathcal{B}(l^2)$ be the truncated shift:

$$S_n(\alpha_1, \alpha_2, \dots) = (0, \alpha_1, \alpha_2, \dots, \alpha_n, 0, 0, \dots)$$

and let $T_n = I - S_n$. $\|S_n\| = 1$, so $\|T_n\| \leq 2$ for all n. Since S_n is nilpotent, $\sigma(S_n) = \{0\}$, so $\sigma(T_n) = \{1\}$ for all n. T_n is invertible, with $T_n^{-1} = I + S_n + S_n^2 + \cdots + S_n^n$, and $\|T_n^{-1}\xi_1\| = \sqrt{n+1}$, so $\|T_n^{-1}\| \geq \sqrt{n+1}$.

If T is restricted to normal operators, then there is a bound: if T is an invertible normal operator, then $\|T^{-1}\| = d^{-1}$, where d is the distance from 0 to $\sigma(T)$. This follows immediately from II.1.6.3 and the elementary fact that $\sigma(T^{-1}) = \{\lambda^{-1} : \lambda \in \sigma(T)\}$ (cf. II.1.5.2(i)).

XX.12.6. Exercises

XX.12.6.1. (a) If A is a unital normed algebra, show that $\|1\| \geq 1$. [$1 = 1 \cdot 1$.]

(b) Give an example of a Banach algebra norm $\|\cdot\|$ on $M_2(\mathbb{C})$ for which $\|I\| > 1$. [Try $\|[a_{ij}]\| = \sum_{i,j} |a_{ij}|$. There are many other possibilities.]

(c) Let A be a unital Banach algebra with norm $\|\cdot\|$. For $x \in A$, define

$$|||a||| = \sup_{\|x\|=1} \|ax\|$$

(i.e. define $L_a \in \mathcal{B}(A)$ by $L_a(x) = ax$, and set $|||a||| = \|L_a\|_{op}$). Show that $|||\cdot|||$ is a Banach algebra norm on A equivalent to $\|\cdot\|$, and that $|||1||| = 1$.

(d) [Kat04, VIII.1.4] Show, more generally, that if A is a Banach space with norm $\|\cdot\|$ which is a unital algebra under a multiplication which is separately continuous, then the norm $|||\cdot|||$ defined in (c) is a Banach algebra norm on A equivalent to $\|\cdot\|$, with $|||1||| = 1$, and in particular multiplication is jointly continuous. [First observe that $|||a||| \geq \frac{\|a\|}{\|1\|}$ for any a. Then show the map $a \mapsto L_a$ from A to $\mathcal{B}(A)$ has closed graph. Conclude that $(A, |||\cdot|||)$ is complete and (using the Closed Graph Theorem) that $|||\cdot|||$ is equivalent to $\|\cdot\|$.]

(e) What if A is nonunital? See XX.16.1.2. for a nonunital version.

XX.12.6.2. This problem gives a proof of the Hausdorff-Toeplitz Theorem:

THEOREM. Let $\mathcal{V}$ be a (complex) inner product space, and $T \in \mathcal{L}(\mathcal{V})$ (not necessarily bounded). Then $W(T)$ is a convex subset of $\mathbb{C}$.

(a) Show that $W(\alpha T + \beta I) = \{\alpha\lambda + \beta : \lambda \in W(T)\}$ for $\alpha, \beta \in \mathbb{C}$, and $\langle Tx, x\rangle = \langle T(\lambda x), \lambda x\rangle$ for $|\lambda| = 1$.

(b) If α and β are complex numbers, show that there is a $\lambda \in \mathbb{C}$, $|\lambda| = 1$, with $\lambda\alpha + \bar{\lambda}\beta$ a nonnegative real number. [Consider $\pm\frac{\bar{\alpha}-\beta}{|\bar{\alpha}-\beta|}$ if $\beta \neq \bar{\alpha}$.] Conclude that if $x, y \in \mathcal{V}$, there is a $\lambda \in \mathbb{C}$, $|\lambda| = 1$, such that $\langle T(\lambda x), y\rangle + \langle Ty, \lambda x\rangle$ is a nonnegative real number.

(c) Conclude from (a) and (b) that to show that $W(T)$ is convex for all T, it suffices to show that if $\langle Tx, x\rangle = 0$ and $\langle Ty, y\rangle = 1$ for $x, y \in \mathcal{V}$, $\|x\| = \|y\| = 1$, $\langle Tx, y\rangle + \langle Ty, x\rangle \geq 0$, then for any $\lambda \in [0, 1]$ there is a $z \in \mathcal{V}$, $\|z\| = 1$, with $\langle Tz, z\rangle = \lambda$.

(d) Let $x, y \in \mathcal{V}$ be as in (c). Show that if $0 \leq \alpha \leq 1$, then $(1-\alpha)x + \alpha y \neq 0$.

(e) For $0 \leq \alpha \leq 1$, set

$$z_\alpha = \frac{(1-\alpha)x + \alpha y}{\|(1-\alpha)x + \alpha y\|}$$

and $f(\alpha) = \langle Tz_\alpha, z_\alpha\rangle$. Show that f is a real-valued continuous function on $[0, 1]$.

(f) Since $f(0) = 0$ and $f(1) = 1$, f takes all values in $[0, 1]$ by the Intermediate Value Theorem.

XX.12.6.3. (a) Let $\mathcal{V}$ be a (complex) inner product space, and $T \in \mathcal{L}(\mathcal{V})$ (not assumed bounded). If $w(T) = 0$, i.e. if $\langle Tx, x\rangle = 0$ for all $x \in \mathcal{V}$, show that $T = 0$. [If $x, y \in \mathcal{V}$, expand $\langle T(x+y), x+y\rangle$ and $T(x+iy), x+iy\rangle$ and conclude that $\langle Tx, y\rangle = 0$.]

(b) Show that the result in (a) is not necessarily true in a real inner product space. [Consider a 90° rotation in $\mathbb{R}^2$.]

(c) Let $\mathcal{H}$ be a Hilbert space. If $S, T \in \mathcal{B}(\mathcal{H})$, show that $w(S+T) \leq w(S) + w(T)$. Thus w is a norm on $\mathcal{B}(\mathcal{H})$. Is it equivalent to the operator norm? Is $\mathcal{B}(\mathcal{H})$ complete with respect to the norm w?

XX.12.6.4. Let $\mathcal{V}$ be a (complex) inner product space, and $T \in \mathcal{L}(\mathcal{V})$ (not necessarily bounded). Suppose $W(T) \subseteq \mathbb{R}$, i.e. $\langle Tx, x\rangle \in \mathbb{R}$ for all $x \in \mathcal{V}$. Show that $\langle Tx, y\rangle = \langle x, Ty\rangle$ for all $x, y \in \mathcal{V}$. [As in XX.12.6.3., expand $\langle T(x+y), x+y\rangle$ and $\langle T(x+iy), x+iy\rangle$.]

XX.12.6.5. [] (a) Let T be the operator on $\mathbb{C}^2$ with triangular matrix

$$\begin{bmatrix} a_1 & b \\ 0 & a_2 \end{bmatrix}.$$

Note that a_1 and a_2 are the eigenvalues of T. Show that $W(T)$ is the ellipse with foci at a_1 and a_2 and minor axis $|b|$. (If $a_1 = a_2$, this ellipse is a circle with radius $|b|/2$, and if $b = 0$, so T is normal, $W(T)$ is just the line segment between a_1 and a_2.) XX.12.5.11. is a special case.

(b) Show that every operator T on $\mathbb{C}^2$ is unitarily equivalent to one with a triangular matrix of this form (with $b \geq 0$), and thus $W(T)$ is an ellipse. (Things are not nearly so simple even for $\mathbb{C}^3$.)

XX.12.6.6. (a) Let $T \in \mathcal{B}(\mathbb{R}^2)$ have matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

with respect to the standard basis. Show that

$$\|T\| = \|A\| = \sqrt{\frac{a^2+b^2+c^2+d^2+\sqrt{a^4+6a^2b^2-2a^2c^2-2a^2d^2+8abcd+b^4-2b^2c^2-2b^2d^2+c^4+6c^2d^2+d^4}}{2}}.$$

[Use that $\|A\|$ is the square root of the largest eigenvalue of A^tA.]

(b) Find a similar formula if $T \in \mathcal{B}(\mathbb{C}^2)$, i.e. $a, b, c, d \in \mathbb{C}$.

(c) (For masochists only, or for someone adept at using computer algebra programs) Find a similar formula for 3×3 and 4×4 matrices.

(d) Show that if $n \geq 5$, there is no algebraic formula for the norm of an $n \times n$ matrix A in terms of the entries of A. Specifically, show that the norm of the matrix

$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 3 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

which in this case is just the largest eigenvalue, approximately 3.917, cannot be written in terms of radicals and rational numbers, i.e. its characteristic polynomial is not solvable by radicals over $\mathbb{Q}$ (D. Zare; cf. https://mathoverflow.net/questions/268453/is-there-an-algebraic-formula-for-the-eigenvalues-of-a-symmetric-n
For larger n just add rows and columns of 0's.

The formula in (a) is not for serious use; it is only to indicate that even in the 2×2 case, the norm of a matrix or linear transformation does not have a simple algebraic formula.

XX.12.6.7. Prove the following nonlinear version of XX.12.2.9.:

THEOREM. Let $\mathcal{X}$ be a Banach space, and $f : \mathcal{X} \to \mathcal{X}$ a strict contraction (a Lipschitz function with Lipschitz constant < 1). Let $\iota : \mathcal{X} \to \mathcal{X}$ be the identity function. Then $g = \iota - f$ is a homeomorphism from $\mathcal{X}$ onto $\mathcal{X}$.

[Reduce to the case $f(0) = 0$ (). Set $f^0 = \iota$ and $f^k = f \circ f \circ \cdots \circ f$ (k times) for $k > 0$, and $h_n = \sum_{k=0}^{n} f^k$. Show that (h_n) converges pointwise on X, and uniformly on bounded sets, to an inverse for g.]

XX.13. Fredholm Operators and Fredholm Index

Fredholm operators are a very important class of "almost invertible" operators. To each Fredholm operator there is associated an integer "index" which roughly speaking measures the extent to which the operator fails to be invertible (or, more precisely, to have a "small" invertible perturbation). Arguably the most important fact about Fredholm operators is that a compact perturbation of a Fredholm operator is also Fredholm, with the same index; in particular, any compact perturbation of an invertible operator is Fredholm, with index 0.

The name "Fredholm operator" is in honor of Swedish mathematician I. Fredholm. It is argued in [Lax02] that there is "no historical justification for this," and that the notions of Fredholm operator and Fredholm index should be attributed to Fritz Noether (son of Max Noether and brother of Emmy Noether). However, Fredholm, who worked around the turn of the 20th century, before Banach spaces were defined, developed some of the crucial basic properties of such operators in his work on integral equations (nicely described in [Lax02]), although he did indeed only consider operators of index 0. The names "Fredholm operator" and "Fredholm index" are also well established so we will use them instead of the names "Noether operator" and "Noether index" used in [Lax02].

The theory of Fredholm operators and Fredholm index is not only important for applications, but (in retrospect) was a pioneering example of bringing algebraic topology into the theory of operator algebras. In fact, it was one of the first examples of the process called "noncommutative topology," which underlies much of the modern theory of operator algebras. See the introduction to [Con94] for an excellent survey of noncommutative topology and geometry.

XX.13.1. Banach Space Preliminaries

The first elementary result uses the quotient norm: if $\mathcal{Y}$ is a closed subspace of a normed vector space $\mathcal{X}$, the *quotient norm* on $\mathcal{X}/\mathcal{Y}$ is the norm $\|x+\mathcal{Y}\| = \inf\{\|x-y\| : y \in \mathcal{Y}\} = \inf\{\|z\| : z+\mathcal{Y} = x+\mathcal{Y}\}$.

XX.13.1.1. Proposition. Let $\mathcal{Y}$ be a proper closed subspace of a normed vector space $\mathcal{X}$, and $\epsilon > 0$. Then there is a $x \in \mathcal{X}$ with $\|x\| \leq 1+\epsilon$ and $\|x-y\| \geq 1$ for all $y \in \mathcal{Y}$.

Proof: Let x be a suitable preimage of a unit vector in $\mathcal{X}/\mathcal{Y}$. 👍

Next we obtain a useful fact which is a simple consequence of the Hahn-Banach Theorem:

XX.13.1.2. Proposition. Let $\mathcal{X}$ be a normed vector space, and $\mathcal{Y}$ a finite-dimensional subspace (which is automatically closed.) Then $\mathcal{Y}$ has a closed complementary subspace $\mathcal{Z}$ (i.e. $\mathcal{Y}+\mathcal{Z} = \mathcal{X}$ and $\mathcal{Y} \cap \mathcal{Z} = \{0\}$.) If $\mathcal{R}$ is a closed subspace of $\mathcal{X}$ with $\mathcal{Y} \cap \mathcal{R} = \{0\}$, then $\mathcal{Z}$ may be chosen to contain $\mathcal{R}$.

Proof: Let $\{x_1, \ldots, x_n\}$ be a basis for $\mathcal{Y}$, and define linear functionals ϕ_i $(1 \leq i \leq n)$ on $\mathcal{Y}$ by $\phi_i(\sum_{j=1}^n \alpha_j x_j) = \alpha_i$ (i.e. $\phi_i(x_j) = \delta_{ij}$.) Then each ϕ_i is bounded since $\mathcal{Y}$ is finite-dimensional. If $\mathcal{W}$ is the subspace of $\mathcal{X}$ spanned by $\mathcal{Y}$ and $\mathcal{R}$, then $\mathcal{W}$ is closed (the image of $\mathcal{W}$ in $\mathcal{X}/\mathcal{R}$ is the same as the image of $\mathcal{Y}$, hence is finite-dimensional and therefore closed), so $\mathcal{W} \cong \mathcal{Y} \oplus \mathcal{R}$. Extend ϕ_i to $\mathcal{W}$ by setting it equal to 0 on $\mathcal{R}$. Then extend ϕ_i to a bounded linear functional ψ_i on $\mathcal{X}$ by the Hahn-Banach Theorem. Define $P \in \mathcal{B}(\mathcal{X})$ by $Py = \sum_{i=1}^n \psi_i(y)x_i$. Then $\mathcal{R}(P) = \mathcal{Y}$, and $P|_{\mathcal{Y}}$ is the identity (P is a projection onto $\mathcal{Y}$.) Then $\mathcal{Z} = \mathcal{N}(P)$ is a closed subspace of $\mathcal{X}$ complementary to $\mathcal{Y}$, containing $\mathcal{R}$. 👍

We recall the definitions of some standard terms from operator theory which will be used in this section.

XX.13.1.3. DEFINITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be normed vector spaces, and $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$. Then T is *bounded below* if there is an $\epsilon > 0$ such that $\|Tx\| \geq \epsilon \|x\|$ for all $x \in \mathcal{X}$.

An operator which is bounded below is one-to-one, and its range is closed if $\mathcal{X}$ is a Banach space.

XX.13.1.4. DEFINITION. An operator $T \in \mathcal{B}(\mathcal{X})$ has *finite ascent* n if the chain $\mathcal{N}(T) \subseteq \mathcal{N}(T^2) \subseteq \cdots$ stabilizes at n, i.e. $\mathcal{N}(T^n) = \mathcal{N}(T^{n+1}) = \cdots$.

T has *finite descent* m if the chain $\mathcal{R}(T) \supseteq \mathcal{R}(T^2) \supseteq \cdots$ stabilizes at m.

XX.13.1.5. Note that if $\mathcal{N}(T^n) = \mathcal{N}(T^{n+1})$, then $\mathcal{N}(T^n) = \mathcal{N}(T^{n+m})$ for all m: if $x \in \mathcal{N}(T^{r+1}) \backslash \mathcal{N}(T^r)$ for some $r > n$, then $T^{r-n}x \in \mathcal{N}(T^{n+1}) \setminus \mathcal{N}(T^n)$. Similarly, if $\mathcal{R}(T^n) = \mathcal{R}(T^{n+1})$, then $\mathcal{R}(T^n) = \mathcal{R}(T^{n+m})$ for all m.

XX.13.1.6. It is fairly exceptional for an operator to have finite ascent or descent, but note that an injective operator has finite ascent 1, and a surjective operator finite descent 1. Every operator on a finite-dimensional space has finite ascent and descent. A normal operator () on a Hilbert space has finite ascent 1, but not necessarily finite descent, even if it is compact (Exercise XX.13.6.1.).

XX.13.1.7. DEFINITION. If $T : \mathcal{X} \to \mathcal{Y}$ is a bounded operator, the *adjoint* of T is the operator $T^* : \mathcal{Y}^* \to \mathcal{X}^*$ defined by $T^*(\phi) = \phi \circ T$. ($\mathcal{X}^*$ is the usual Banach space dual of $\mathcal{X}$, the space of bounded linear functionals on $\mathcal{X}$.)

XX.13.1.8. $T^* \in \mathcal{B}(\mathcal{Y}^*, \mathcal{X}^*)$ and $\|T^*\| = \|T\|$ by the Hahn-Banach Theorem. If $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces, then $\mathcal{X}^*$ and $\mathcal{Y}^*$ can be identified with $\mathcal{X}$ and $\mathcal{Y}$ in the standard way, and then T^* agrees with the usual adjoint operator defined by the inner product: $\langle Tx, y\rangle = \langle x, T^*y\rangle$ for $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

XX.13.1.9. PROPOSITION. If $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ is finite-rank, then T^* is finite-rank.

PROOF: T factors through $\mathcal{R}(T)$, so T^* factors through $\mathcal{R}(T)^*$, which is finite-dimensional. 👍

XX.13.2. Compact Operators

XX.13.2.1. DEFINITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces. An operator $T : \mathcal{X} \to \mathcal{Y}$ is *compact* if T sends bounded subsets of $\mathcal{X}$ to precompact subsets of $\mathcal{Y}$.

XX.13.2.2. In other words, if B is a bounded subset of $\mathcal{X}$, then $\overline{T(B)}$ is a compact subset of $\mathcal{Y}$. Equivalently, whenever (x_n) is a bounded sequence of vectors in $\mathcal{X}$, then (Tx_n) has a convergent subsequence. T is compact if and only if $T(\mathcal{B}_\mathcal{X})$ is totally bounded. Since a compact metric space is separable, if $T : \mathcal{X} \to \mathcal{Y}$ is compact, $\overline{\mathcal{R}(T)}$ is a separable subspace of $\mathcal{Y}$.

XX.13.2.3. In many references, particularly older ones, compact operators are called *completely continuous operators.*

XX.13.2.4. It is obvious that a compact operator is bounded, and the composition (in either order) of a compact operator with a bounded operator is compact. It is easily seen using the total boundedness criterion that a (norm-)limit of compact operators is compact. The set of compact operators from $\mathcal{X}$ to $\mathcal{Y}$, denoted $\mathcal{K}(\mathcal{X},\mathcal{Y})$, is a closed subspace of $\mathcal{B}(\mathcal{X},\mathcal{Y})$; $\mathcal{K}(\mathcal{X}) = \mathcal{K}(\mathcal{X},\mathcal{X})$ is a closed two-sided ideal in $\mathcal{B}(\mathcal{X})$. Every finite-rank bounded operator is compact; for many, but not all, Banach spaces, every compact operator is a limit of finite-rank operators (see the discussion below.)

XX.13.2.5. It is easily seen that a bounded operator is weakly continuous (cf. [?].) So if $\mathcal{X}$ is a Hilbert space (or, more generally, a reflexive Banach space), $\mathcal{B}_\mathcal{X}$ is weakly compact, so if $T \in \mathcal{B}(\mathcal{X},\mathcal{Y})$, $T(\mathcal{B}_\mathcal{X})$ is weakly compact and thus (norm-)closed. In particular, if $T \in \mathcal{K}(\mathcal{X},\mathcal{Y})$, then $T(\mathcal{B}_\mathcal{X})$ is already closed and therefore compact.

XX.13.2.6. PROPOSITION. If $T \in \mathcal{K}(\mathcal{X},\mathcal{Y})$, then $T^* \in \mathcal{K}(\mathcal{Y}^*,\mathcal{X}^*)$.

PROOF: Let $\epsilon > 0$. Since $T(\mathcal{B}_\mathcal{X})$ is totally bounded, there are $x_1,\cdots,x_n \in \mathcal{B}_\mathcal{X}$ such that, for each $x \in \mathcal{B}_\mathcal{X}$, there is an i with $\|Tx - Tx_i\| < \epsilon/3$. Give $\mathbb{F}^n$ the infinity norm, i.e. $\|(\alpha_1,\ldots,\alpha_n)\| = \max|\alpha_i|$, and define $S : \mathcal{Y}^* \to \mathbb{F}^n$ by $S\phi = (\phi(Tx_1),\ldots,\phi(Tx_n))$. S is finite-rank and bounded and hence compact, so $S(\mathcal{B}_{\mathcal{Y}^*})$ is totally bounded; let $\phi_1,\cdots,\phi_m \in \mathcal{B}_{\mathcal{Y}^*}$ be such that, for every $\phi \in \mathcal{B}_{\mathcal{Y}^*}$, there is a j such that $\|S\phi - S\phi_j\| < \epsilon/3$. If $\phi \in \mathcal{B}_{\mathcal{Y}^*}$, choose a corresponding ϕ_j; then, for any $x \in \mathcal{B}_\mathcal{X}$, with corresponding x_i,

$$|T^*\phi(x) - T^*\phi_j(x)| = |\phi(Tx) - \phi_j(Tx)|$$
$$\leq |\phi(Tx) - \phi(Tx_i)| + |\phi(Tx_i) - \phi_j(Tx_i)| + |\phi_j(Tx_i) - \phi_j(Tx)|$$
$$\leq \|Tx - Tx_i\| + \|S\phi - S\phi_j\| + \|Tx_i - Tx\| < \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon\ .$$

Thus $\|T^*\phi - T^*\phi_j\| \leq \epsilon$, so $T^*(\mathcal{B}_{\mathcal{Y}^*})$ is totally bounded. 👍

While the proof of XX.13.2.6. is not difficult, there is an alternate approach (of comparable difficulty) on a Hilbert space which is worth noting:

XX.13.2.7. THEOREM. If $\mathcal{Y}$ is a Hilbert space, then $\mathcal{K}(\mathcal{X},\mathcal{Y})$ is the closure of the finite-rank bounded operators from $\mathcal{X}$ to $\mathcal{Y}$.

PROOF: We again use a total boundedness argument. Let $T \in \mathcal{K}(\mathcal{X},\mathcal{Y})$, and $\epsilon > 0$. Choose $x_1,\cdots,x_n \in \mathcal{B}_\mathcal{X}$ such that, for any $x \in \mathcal{B}_\mathcal{X}$, there is an i with $\|Tx - Tx_i\| < \epsilon/2$. Let $P \in \mathcal{B}(\mathcal{Y})$ be the orthogonal projection onto the subspace spanned by $\{Tx_1,\cdots,Tx_n\}$. Then P is finite-rank, $PTx_i = Tx_i$ for all i, and $\|P\| = 1$. If $x \in \mathcal{B}_\mathcal{X}$, and a corresponding x_i is chosen, then

$$\|PTx - Tx\| \leq \|PTx - PTx_i\| + \|Tx_i - Tx\| < \epsilon/2 + \epsilon/2 = \epsilon\ .$$

Thus $\|PT - T\| \leq \epsilon$, and PT has finite rank. 👍

XX.13.2.8. If $T \in \mathcal{K}(\mathcal{X},\mathcal{Y})$ and $\mathcal{Y}$ is a Hilbert space, an alternate proof of XX.13.2.6. is then obtained by taking (T_n) of finite rank and $T_n \to T$; then $T_n^* \to T^*$, and T_n^* is finite-rank by XX.13.1.9.. Similarly, if $T \in \mathcal{K}(\mathcal{X},\mathcal{Y})$ and $\mathcal{X}$ is a Hilbert space, apply XX.13.2.7. to $T^* \in \mathcal{K}(\mathcal{Y}^*,\mathcal{X})$ (using XX.13.2.6.) to conclude that $\mathcal{K}(\mathcal{X},\mathcal{Y})$ is also the closure of the finite-rank bounded operators from $\mathcal{X}$ to $\mathcal{Y}$ (there is also a simple direct proof.)

XX.13.2.9. DEFINITION. A Banach space $\mathcal{X}$ has the *approximation property* if every compact operator on $\mathcal{X}$ is a (norm-)limit of finite-rank (bounded) operators on $\mathcal{X}$.

XX.13.2.10. XX.13.2.7. shows that Hilbert spaces have the approximation property. All standard Banach spaces have the approximation property. P. ENFLO [Enf73] constructed the first example of a Banach space without the approximation property; see [Jam82] for a discussion of this and related matters, and [Gow95] for more recent developments.

Strictly Singular Operators

There is a slightly larger class of operators which to some extent behave like compact operators:

XX.13.2.11. DEFINITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, with $\mathcal{X}$ infinite-dimensional. An operator $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ is *strictly singular* if T is not bounded below on any infinite-dimensional subspace of $\mathcal{X}$.

XX.13.2.12. It follows from () that every compact operator is strictly singular. Every strictly singular operator on a Hilbert space, or many standard Banach spaces, is compact; but there are operators on some Banach spaces which are strictly singular but not compact (XX.13.6.2.).

XX.13.3. Fredholm Operators

XX.13.3.1. DEFINITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$. T is a *Fredholm operator* if $\mathcal{N}(T)$ is finite-dimensional and $\mathcal{R}(T)$ has finite codimension. The *index* of T is $dim\ \mathcal{N}(T) - codim\ \mathcal{R}(T)$. Denote the set of Fredholm operators from $\mathcal{X}$ to $\mathcal{Y}$ by $\mathcal{F}(\mathcal{X}, \mathcal{Y})$, and write $\mathcal{F}(\mathcal{X})$ for $\mathcal{F}(\mathcal{X}, \mathcal{X})$.

XX.13.3.2. EXAMPLES.

(i) If $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional, then every operator from $\mathcal{X}$ to $\mathcal{Y}$ is Fredholm with index $dim\ \mathcal{X} - dim\ \mathcal{Y}$.

(ii) If T is invertible, then T is Fredholm with index 0.

(iii) Let S be the unilateral shift on the Hilbert space l^2 of square-summable sequences, defined by $S(\alpha_1, \alpha_2, \dots) = (0, \alpha_1, \alpha_2, \dots)$. Then S is Fredholm with index -1. S^n is Fredholm with index $-n$, and $(S^*)^n$ is Fredholm with index n.

XX.13.3.3. Definition XX.13.3.1. is not the most common one, which in addition requires $\mathcal{R}(T)$ to be closed. However, the two definitions are equivalent, since a Fredholm operator in the sense of XX.13.3.1. always has closed range. Although strictly speaking we will not need this fact for our development of the theory, we give a proof for completeness (and to show that we are not redefining the notion of Fredholm operator), and also because the argument gives some important structural properties of Fredholm operators.

XX.13.3.4. PROPOSITION. Let $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then $\mathcal{R}(T)$ is closed.

PROOF: Let $\mathcal{W}$ be a complementary subspace for $\mathcal{R}(T)$ in $\mathcal{Y}$; then $\mathcal{W}$ is finite-dimensional and hence closed. Let $\mathcal{Z}$ be a closed subspace of $\mathcal{X}$ complementary to $\mathcal{N}(T)$ (XX.13.1.2.). Then $\mathcal{R}(T|_{\mathcal{Z}}) = \mathcal{R}(T)$. Define a bounded operator S from the Banach space $\mathcal{W} \oplus \mathcal{Z}$ to $\mathcal{Y}$ by $S(w, z) = w + Tz$. Then S is a bijection, hence a homeomorphism by the Open Mapping Theorem, and $S(\{0\} \oplus \mathcal{Z}) = \mathcal{R}(T)$. 👍

A corollary of the proof gives a simple picture of a Fredholm operator:

XX.13.3.5. COROLLARY. Let $\mathcal{X}$, $\mathcal{Y}$ be Banach spaces, and $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then there are closed subspaces $\mathcal{Z}$, $\mathcal{W}$ of $\mathcal{X}$ and $\mathcal{Y}$ complementary to $\mathcal{N}(T)$ and $\mathcal{R}(T)$, respectively, such that $\mathcal{X} \cong \mathcal{N}(T) \oplus \mathcal{Z}$, $\mathcal{Y} \cong \mathcal{W} \oplus \mathcal{R}(T)$, and $T \cong 0 \oplus U$, where U is a linear isomorphism from $\mathcal{Z}$ onto $\mathcal{R}(T)$ (i.e. U is bounded below.)

XX.13.3.6. COROLLARY. Let $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then there is an $S \in \mathcal{F}(\mathcal{Y}, \mathcal{X})$, with $index(S) = -index(T)$, such that $ST = I_{\mathcal{X}} - P$ and $TS = I_{\mathcal{Y}} - Q$, where P and Q are finite-rank bounded idempotent operators.

PROOF: Under the above identifications, S may be taken as $0 \oplus U^{-1} : \mathcal{W} \oplus \mathcal{R}(T) \to \mathcal{N}(T) \oplus \mathcal{Z}$. 👍

XX.13.3.7. The operator S is a *quasi-inverse* for T: $T = TST$. (If S is as defined in the proof, T is also a quasi-inverse for S.)

We obtain another characterization of Fredholm operators from XX.13.3.5.:

XX.13.3.8. COROLLARY. Let $\mathcal{X}$, $\mathcal{Y}$ be Banach spaces, and $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$. Then $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ if and only if there is a closed subspace $\mathcal{Z}$ of $\mathcal{X}$, of finite codimension, with $T(\mathcal{Z})$ of finite codimension in $\mathcal{Y}$, such that T is bounded below on $\mathcal{Z}$. If there is such a $\mathcal{Z}$, then $index(T) = codim(\mathcal{Z}) - codim(T(\mathcal{Z}))$.

XX.13.3.9. COROLLARY. Let $\mathcal{X}$, $\mathcal{Y}$ be Banach spaces, and $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then there is an $\epsilon > 0$ such that $T + S \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ and $index(T + S) = index(T)$ for all $S \in \mathcal{B}(X, Y)$, $\|S\| < \epsilon$.

PROOF: Let $\mathcal{Z}$ be a closed subspace of finite codimension in $\mathcal{X}$ such that T is bounded below on $\mathcal{Z}$. Then, if $\|S\|$ is sufficiently small, $T + S$ is bounded below on $\mathcal{Z}$ and $(T + S)(\mathcal{Z})$ is a small enough perturbation of $T(\mathcal{Z})$ to have the same codimension (). 👍

Another important consequence of XX.13.3.5. is:

XX.13.3.10. Corollary. Let $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$. Then $T^* \in \mathcal{F}(\mathcal{Y}^*, \mathcal{X}^*)$, $dim(\mathcal{N}(T^*)) = codim(\mathcal{R}(T))$, $codim(\mathcal{R}(T^*)) = dim(\mathcal{N}(T))$, and $index(T^*) = -index(T)$.

Proof: If $T = 0 \oplus U : \mathcal{N}(T) \oplus \mathcal{Z} \to \mathcal{W} \oplus \mathcal{R}(T)$, then $T^* = 0 \oplus U^* : \mathcal{W}^* \oplus \mathcal{R}(T)^* \to \mathcal{N}(T)^* \oplus \mathcal{Z}^*$, and U^* is invertible; thus $\mathcal{N}(T^*) = \mathcal{W}^* \oplus \{0\}$ and $\mathcal{R}(T^*) = \{0\} \oplus \mathcal{Z}^*$. We have $\mathcal{W}^* \cong \mathcal{W}$ and $\mathcal{N}(T)^* \cong \mathcal{N}(T)$ since they are finite-dimensional. 👍

XX.13.3.11. Actually, if $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$, then $\mathcal{N}(T^*) = \mathcal{R}(T)^\perp = \{\phi \in \mathcal{Y}^* : \phi|_{\mathcal{R}(T)} = 0\} \cong (\mathcal{Y}/\overline{\mathcal{R}(T)})^*$. Thus $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ is Fredholm if and only if $\mathcal{R}(T)$ is closed and $\mathcal{N}(T)$ and $\mathcal{N}(T^*)$ are finite-dimensional, and $index(T) = dim\ \mathcal{N}(T) - dim\ \mathcal{N}(T^*)$.

It is useful not to require that the range of a Fredholm operator be closed as part of the definition. The next proof is an example.

XX.13.3.12. Proposition. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, $T \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$, $S \in \mathcal{B}(\mathcal{Y}, \mathcal{X})$. If ST and TS are Fredholm, then S and T are Fredholm.

Proof: We show T is Fredholm. $\mathcal{N}(T) \subseteq \mathcal{N}(ST)$, which is finite-dimensional, and $\mathcal{R}(T) \supseteq \mathcal{R}(TS)$, which has finite codimension. 👍

Conversely:

XX.13.3.13. Theorem. Let $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ be Banach spaces, and $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ and $S \in \mathcal{F}(\mathcal{Y}, \mathcal{Z})$. Then $ST \in \mathcal{F}(\mathcal{X}, \mathcal{Z})$ and $index(ST) = index(S) + index(T)$.

Proof: Let $\mathcal{U}$ be a closed subspace of $\mathcal{X}$ complementary to $\mathcal{N}(T)$ on which T is bounded below. Then $U = T|_{\mathcal{U}}$ is an isomorphism from $\mathcal{U}$ onto $\mathcal{R}(T)$. Set $\mathcal{Y}_2 = \mathcal{R}(T) \cap \mathcal{N}(S)$ and $\mathcal{X}_2 = U^{-1}(\mathcal{Y}_2)$. Then $\mathcal{X}_2$ is a finite-dimensional subspace of $\mathcal{U}$. Let $\mathcal{X}_1$ be a closed subspace of $\mathcal{U}$ complementary to $\mathcal{X}_2$, and let $\mathcal{Y}_1 = U(\mathcal{X}_1) = T(\mathcal{X}_1)$. Then $\mathcal{Y}_1$ is a closed subspace of $\mathcal{Y}$ with zero intersection with $\mathcal{N}(S)$, and $\mathcal{Y}_1$ and $\mathcal{Y}_2$ span $\mathcal{R}(T)$, and since $\mathcal{Y}_2$ is finite-dimensional and $\mathcal{R}(T)$ has finite codimension in $\mathcal{Y}$, $\mathcal{Y}_1$ has finite codimension in $\mathcal{Y}$. Let $\mathcal{V}$ be a closed subspace of $\mathcal{Y}$ containing $\mathcal{Y}_1$ which is complementary to $\mathcal{N}(S)$ (XX.13.1.2.), and $\mathcal{Y}_3$ a subspace of $\mathcal{V}$ complementary to $\mathcal{Y}_1$. Then $\mathcal{Y}_3$ is finite-dimensional, and we have

$$\mathcal{Y} \cong \mathcal{Y}_1 \oplus \mathcal{Y}_2 \oplus \mathcal{Y}_3 \oplus \mathcal{Y}_4$$

where $\mathcal{Y}_4$ is a subspace of $\mathcal{N}(S)$ complementary to $\mathcal{Y}_2$. Then $\mathcal{Y}_2, \mathcal{Y}_3, \mathcal{Y}_4$ are finite-dimensional, $\mathcal{Y}_2$ and $\mathcal{Y}_4$ span $\mathcal{N}(S)$, and $\mathcal{Y}_1$ and $\mathcal{Y}_2$ span $\mathcal{R}(T)$. Then S is bounded below on $\mathcal{V}$ (cf. the proof of XX.13.3.4.), and $V = S|_{\mathcal{V}}$ maps $\mathcal{V}$ isomorphically onto $\mathcal{R}(S)$. Also, $UV = ST|_{\mathcal{X}_1}$ maps $\mathcal{X}_1$ isomorphically onto $\mathcal{Z}_1 = V(\mathcal{Y}_1) = S(Y_1)$. Since $\mathcal{Y}_1$ has finite codimension in $\mathcal{Y}$ and $\mathcal{R}(S)$ has finite codimension in $\mathcal{Z}$, $\mathcal{Z}_1$ has finite codimension in $\mathcal{Z}$ and ST is Fredholm.

We have

$$\mathcal{X} \cong \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \mathcal{X}_3$$

$$\mathcal{Z} \cong \mathcal{Z}_1 \oplus \mathcal{Z}_2 \oplus \mathcal{Z}_3$$

where $\mathcal{X}_3 = \mathcal{N}(T)$, $\mathcal{Z}_3 = V(\mathcal{Y}_3) = S(\mathcal{Y}_3)$, and $\mathcal{Z}_2$ is a subspace of $\mathcal{Z}$ complementary to $\mathcal{R}(S)$. Set $a_k = dim(\mathcal{X}_k)$, $b_k = dim(\mathcal{Y}_k)$, and $c_k = dim(\mathcal{Z}_k)$. We then have

$$index(T) = a_3 - (b_3 + b_4)$$

$$index(S) = (b_2 + b_4) - c_2$$

$$index(ST) = (a_2 + a_3) - (c_2 + c_3) \ .$$

But $b_2 = a_2$ and $c_3 = b_3$ since U and V are isomorphisms, so we have

$$index(ST) = b_2 + a_3 - b_3 - c_2 = index(S) + index(T) \ .$$



XX.13.4. Fredholm Operators of Index 0 and the Fredholm Alternative

We first note an immediate consequence of XX.13.3.5.:

XX.13.4.1. PROPOSITION. Let $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ have index 0. Then there are $V, F \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ with V invertible and F finite-rank, with $T = V + F$.

The converse is also true (XX.13.5.12.).

PROOF: Under the identifications $\mathcal{X} \cong \mathcal{N}(T) \oplus \mathcal{Z}$, $\mathcal{Y} \cong \mathcal{W} \oplus \mathcal{R}(T)$, $T \cong 0 \oplus U$, we have that $dim(\mathcal{N}(T)) = dim(\mathcal{W}) < \infty$, so there is a (necessarily bounded) linear isomorphism $S : \mathcal{N}(T) \to \mathcal{W}$. Set $V = S \oplus U$, $F = -S \oplus 0$. 

The next observation is called the *Fredholm Alternative*. The last statement follows from XX.13.3.10..

XX.13.4.2. PROPOSITION. Let $T \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ have index 0. Then T is injective if and only if T is surjective. So either the equation $Tx = y$ has a unique solution for each $y \in \mathcal{Y}$, or the homogeneous equation $Tx = 0$ has a nontrivial (but finite-dimensional) space of solutions.

Furthermore, the equation $T^*\phi = \psi$ has a unique solution for each ψ if and only if $Tx = y$ has a unique solution for each y, and the solution set of $T^*\phi = 0$ has the same (finite) dimension as the solution set of $Tx = 0$.

XX.13.4.3. The result is sometimes paraphrased: "If the solution to $Tx = y$ is unique, then it exists."

XX.13.5. Relation between Compact and Fredholm Operators

The next theorem is the main result of this section.

XX.13.5.1. THEOREM. Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $I + T$ is Fredholm, of index 0.

The proof that $I + T$ is Fredholm consists of the next three lemmas.

XX.13.5.2. LEMMA. Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $\mathcal{N}(I+T)$ is finite-dimensional.

PROOF: $T = -I$ on $\mathcal{N}(I+T)$, and thus the unit ball of $\mathcal{N}(I+T)$ is compact. 👍

XX.13.5.3. LEMMA. Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $\mathcal{R}(I+T)$ is closed.

PROOF: Let $\mathcal{Z}$ be a closed subspace of $\mathcal{X}$ complementary to $\mathcal{N}(I+T)$ (XX.13.1.2.). It suffices to show that $I+T$ is bounded below on $\mathcal{Z}$; then $\mathcal{R}(I+T) = \mathcal{R}((I+T)|_{\mathcal{Z}})$ is closed. Suppose (x_n) is a sequence of unit vectors in $\mathcal{Z}$ with $\|(I+T)x_n\| \to 0$. Passing to a subsequence we may assume that $Tx_n \to y$ for some y, since T is compact. But $(I+T)x_n \to 0$, so $x_n \to -y$, and thus $\|y\| = 1$ and $Ty = -y$, i.e. $(I+T)y = 0$, $y \in \mathcal{N}(I+T)$. But $y \in \mathcal{Z}$ since $-x_n \to y$ and $\mathcal{Z}$ is closed; hence $y \in \mathcal{Z} \cap \mathcal{N}(I+T) = \{0\}$, a contradiction. 👍

XX.13.5.4. LEMMA. Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $\mathcal{R}(I+T)$ has finite codimension.

PROOF: Consider the Banach space $\mathcal{Y} = \mathcal{X}/\mathcal{R}(I+T)$. We show that $\mathcal{Y}$ is finite-dimensional by showing that the dual space $\mathcal{Y}^*$ is finite-dimensional. $\mathcal{Y}^*$ can be identified with the subspace of $\mathcal{X}^*$ consisting of functionals vanishing on $\mathcal{R}(I+T)$. But $\mathcal{Y}^*$ is precisely $\mathcal{N}((I+T)^*) = \mathcal{N}(I+T^*)$. $T^* \in \mathcal{K}(\mathcal{X}^*)$ by XX.13.2.6., so $\mathcal{N}(I+T^*)$ is finite-dimensional by XX.13.5.2.. 👍

It remains to be shown that $index(I+T) = 0$. Our proof is partly based in spirit on [RSN90, §70] (repeated without attribution in the rather dreadful paper [Ram01]). We first obtain another important property of $I+T$ which is a simple (but nonobvious) consequence of the compactness of T.

XX.13.5.5. PROPOSITION. Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then $I+T$ has finite ascent and finite descent.

PROOF: If $I+T$ does not have finite ascent, by XX.13.1.1. choose for each n a vector $x_n \in \mathcal{N}((I+T)^n)$ with $\|x_n\| \le 2$ and $\|x_n - y\| \ge 1$ for all $y \in \mathcal{N}((I+T)^{n-1})$. Then, for $m < n$,

$$\|Tx_m - Tx_n\| = \|x_n - (I+T)x_n + (I+T)x_m - x_m\| \ge 1$$

since $x_m - (I+T)x_m + (I+T)x_n \in \mathcal{N}((I+T)^{n-1})$. Thus (Tx_n) has no convergent subsequence, contradicting compactness of T.

The proof that $I+T$ has finite descent is virtually identical. 👍

XX.13.5.6. LEMMA. If $\mathcal{X}$ is a Banach space and $T \in \mathcal{K}(\mathcal{X})$, then $index(I+T) = 0$.

PROOF: Let n be the maximum of the ascent and descent of $I+T$, so $\mathcal{N}((I+T)^n) = \mathcal{N}((I+T)^{n+k})$ and $\mathcal{R}((I+T)^n) = \mathcal{R}((I+T)^{n+k})$ for all k. Note that $(I+T)^n = I+S$ for $S \in \mathcal{K}(\mathcal{X})$, so $(I+T)^n$ is Fredholm and hence $\mathcal{R} = \mathcal{R}((I+T)^n)$ is closed and $\mathcal{N} = \mathcal{N}((I+T)^n)$ is finite-dimensional. Then $\mathcal{N} \cap \mathcal{R} = \{0\}$, for if $0 \neq y = (I+T)^n x \in \mathcal{N} \cap \mathcal{R}$, then $x \in \mathcal{N}((I+T)^{2n}) \setminus \mathcal{N}((I+T)^n)$, a contradiction. Thus $(I+T)^n$ maps $\mathcal{R}$ isomorphically onto $\mathcal{R}$ (since $\mathcal{R} = \mathcal{R}((I+T)^{2n})$.) If $\mathcal{Z}$ is a closed subspace of $\mathcal{X}$ containing $\mathcal{R}$ which is complementary to $\mathcal{N}$, then $(I+T)^n$ also maps $\mathcal{Z}$ isomorphically onto $\mathcal{R}$, and it follows that $\mathcal{Z} = \mathcal{R}$, i.e. $\mathcal{N}$ and $\mathcal{R}$ are complementary subspaces of $\mathcal{X}$. Since both $\mathcal{N}$ and $\mathcal{R}$ are invariant under $I+T$ and $(I+T)|_{\mathcal{R}}$ is an isomorphism, $index(I+T) = index((I+T)|_{\mathcal{N}})$. But $\mathcal{N}$ is finite-dimensional, so $index((I+T)|_{\mathcal{N}}) = 0$. 👍

This completes the proof of XX.13.5.1.. 💪

The proof of XX.13.5.6. actually shows that $I+T$ has a simple form:

XX.13.5.7. COROLLARY. Let $\mathcal{X}$ be a Banach space, $T \in \mathcal{K}(\mathcal{X})$. Then there are complementary closed subspaces $\mathcal{N}$ and $\mathcal{R}$ of $\mathcal{X}$, each invariant under T (and hence under $I+T$), with $\mathcal{N}$ finite-dimensional, such that $(I+T)|_{\mathcal{R}}$ is invertible and $(I+T)|_{\mathcal{N}}$ is nilpotent.

It is well known (and useful) that every operator on a finite-dimensional space has this property.

The next theorem [Atk53] gives a useful alternate characterization of Fredholm operators.

XX.13.5.8. THEOREM. [ATKINSON'S THEOREM] Let $\mathcal{X}$ be an infinite-dimensional Banach space and $\pi : \mathcal{B}(\mathcal{X}) \to \mathcal{Q}(\mathcal{X})$ the quotient map onto the Calkin algebra $\mathcal{Q}(\mathcal{X}) = \mathcal{B}(\mathcal{X})/\mathcal{K}(\mathcal{X})$ of $\mathcal{X}$. Then $T \in \mathcal{B}(\mathcal{X})$ is Fredholm if and only if $\pi(T)$ is invertible in $\mathcal{Q}(\mathcal{X})$.

PROOF: If T is Fredholm, let S be a quasi-inverse for T as in XX.13.3.6.; then $\pi(S)$ is an inverse for $\pi(T)$.

Conversely, suppose $\pi(T)$ is invertible in $\mathcal{Q}(\mathcal{X})$, and let $S \in \mathcal{B}(\mathcal{X})$ such that $\pi(S) = \pi(T)^{-1}$. Then $ST = I+K$, $TS = I+L$ for $K, L \in \mathcal{K}(\mathcal{X})$. $I+K$ and $I+L$ are Fredholm by XX.13.5.1., so S and T are Fredholm by XX.13.3.12.. 👍

XX.13.5.9. COROLLARY.

(i) If $\mathcal{X}$ is a Banach space, $S, T \in \mathcal{F}(\mathcal{X})$, and $K \in \mathcal{K}(\mathcal{X})$, then ST and $T+K$ are Fredholm.

(ii) The set $\mathcal{F}(\mathcal{X})$ of Fredholm operators on $\mathcal{X}$ is open in $\mathcal{B}(\mathcal{X})$.

Part (ii) is an immediate consequence of XX.13.3.9., or follows directly from (). It also follows from XX.13.3.9. that *index* is locally constant on $\mathcal{F}(\mathcal{X})$, hence constant on connected components. In particular, we obtain:

XX.13.5.10. Corollary. Let $\mathcal{X}$ be a Banach space, and $T \in \mathcal{F}(\mathcal{X})$. Then $index(T) = index(T+K)$ for any $K \in \mathcal{K}(\mathcal{X})$.

Proof: T and $T+K$ are connected in $\mathcal{F}(\mathcal{X})$ by the path $T+tK$, $0 \le t \le 1$. 👍

XX.13.5.11. The *index* map from the connected components of $\mathcal{F}(\mathcal{X})$ to $\mathbb{Z}$ is thus well defined, but it is not either injective or surjective in general: it can fail to be injective since the invertible elements of $\mathcal{B}(\mathcal{X})$ are not necessarily connected [**?**], and we can even have $index(T) = 0$ for every $T \in \mathcal{F}(\mathcal{X})$ (e.g. for the examples of [**?**]; cf. [**?**].) If $\mathcal{H}$ is a Hilbert space, the *index* map is a bijection from the connected components of $\mathcal{F}(\mathcal{H})$ onto $\mathbb{Z}$.

XX.13.5.12. Theorem. Let $T \in \mathcal{F}(\mathcal{X})$. Then the following are equivalent:

(i) $index(T) = 0$.

(ii) There are $U, K \in \mathcal{B}(\mathcal{X})$ with U invertible, K compact, and $T = U + K$.

(iii) There are $V, F \in \mathcal{B}(\mathcal{X})$ with V invertible, F finite-rank, and $T = V + F$.

Proof: (i) ⇒ (iii) is XX.13.4.1. and (iii) ⇒ (ii) is trivial. For (ii) ⇒ (i), if $T = U+K$, then $U^{-1}T = I+U^{-1}K$ is Fredholm with index 0 by XX.13.5.1.. It is obvious that if S is Fredholm and W is invertible, then WS is Fredholm and $index(WS) = index(S)$; thus $T = U(U^{-1}T)$ is Fredholm and $index(T) = 0$. 👍

The argument of [RSN90, §70] gives a direct proof of (iii) ⇒ (i).

XX.13.5.13. Note that if $\mathcal{X}$ has the approximation property, then there is a simple proof of (ii) ⇒ (iii): if $T = U + K$ with U invertible and K compact, approximate K closely enough by a finite rank F that $V = U + (K - F)$ is invertible (using the fact that the set of invertible elements of $\mathcal{B}(\mathcal{X})$ is open); then $T = V + F$.

Finally, we state the Fredholm Alternative in its usual form:

XX.13.5.14. Corollary. [Fredholm Alternative] If $\mathcal{X}$ is a Banach space and $T \in \mathcal{K}(\mathcal{X})$, then $I + T$ is injective if and only if it is surjective. So either the equation $(I+T)x = y$ has a unique solution for every y, or the homogeneous equation $(I+T)x = 0$ has a nontrivial (but finite-dimensional) space of solutions.

Furthermore, the equation $(I+T^*)\phi = \psi$ has a unique solution for each ψ if and only if $(I+T)x = y$ has a unique solution for each y, and the solution set of $(I+T^*)\phi = 0$ has the same (finite) dimension as the solution set of $(I+T)x = 0$.

XX.13.5.15. Note also that if $T \in \mathcal{K}(\mathcal{X})$, then XX.13.5.1. and its consequences, notably XX.13.5.14., are valid not only for $I + T$ but also for $T - \lambda I = -\lambda(I + (-\lambda^{-1}T))$ for any $\lambda \neq 0$. This is useful in describing the eigenvalues of T. The results of Fredholm theory, especially XX.13.5.5., are used to prove that the spectrum of a compact operator on an infinite-dimensional Banach space consists of 0 and at most countably many isolated nonzero eigenvalues of finite multiplicity.

XX.13.5.16. The results of this section can be extended: if T is Fredholm and S is strictly singular, then $T+S$ is Fredholm and $index(T+S) = index(T)$ (cf. ()). Thus the Fredholm Alternative holds more generally for strictly singular operators.

XX.13.6. Exercises

XX.13.6.1. (a) Use the Spectral Theorem () to show that every normal operator on a Hilbert space has finite ascent 1.

(b) Consider the multiplication operator on ℓ^2 given by

$$T(x_1, x_2, \dots) = \left(x_1, \frac{x_2}{2}, \dots, \frac{x_n}{n}, \dots\right) .$$

Show that T is self-adjoint, hence normal, and compact, but does not have finite descent.

The difference between ascent and descent (at least in this case) is that the null space of an operator is closed but the range is not necessarily closed. In fact, if T is a normal operator on a Hilbert space, then the closure of the range of T^n is the same as the closure of $\mathcal{R}(T)$ for all n (it is the orthogonal complement of $\mathcal{N}(T) = \mathcal{N}(T^n)$).

(c) Define $S : \ell^2 \to \ell^2$ by

$$S(x_1, x_2, \dots) = \left(\frac{x_2}{2}, \dots, \frac{x_n}{n}, \dots\right) .$$

Then S is a (nonnormal) compact operator which does not have finite ascent or descent.

XX.13.6.2. (a) Show that every strictly singular operator on a Hilbert space is compact.

(b) Show that the inclusion map from ℓ^1 to ℓ^2 is not compact. [Consider the standard basis.]

(c) Show that the inclusion map from ℓ^1 to ℓ^2 is strictly singular. [Use that the weak topology on ℓ^1 has the same convergent sequences as as the norm topology ().]

(d) Let $\mathcal{X} = \ell^1 \oplus \ell^2$. Define $T \in \mathcal{B}(\mathcal{X})$ by $T(x, y) = (0, x)$. Show that T is strictly singular but not compact.

(e) Show that if $p < q$, the inclusion of ℓ^p into ℓ^q is strictly singular but not compact. So if $1 < p < q < \infty$, there is an operator on the reflexive (even uniformly convex, if the 2-norm is used on the direct sum) Banach space $\ell^p \oplus \ell^q$ which is strictly singular but not compact.

XX.13.6.3. Let $\mathcal{X}$ be a Banach space. Show that the set of strictly singular operators on $\mathcal{X}$ is a closed two-sided ideal in $\mathcal{B}(\mathcal{X})$.

XX.13.6.4. Let X be a Banach space, and S a strictly singular operator on $\mathcal{X}$. Modify the proofs for compact operators to show the following:

(a) If $T \in \mathcal{F}(\mathcal{X})$, then $T+S \in \mathcal{F}(\mathcal{X})$ and $index(T+S) = index(T)$. In particular, if T is invertible, then $T+S$ is a Fredholm operator of index 0.

(b) $\sigma(S)$ consists of 0 and at most countably many nonzero numbers, each an eigenvalue of finite multiplicity, with no nonzero accumulation point.

XX.14. The Spectral Theorem

The Spectral Theorem, in its various forms, gives a precise description of the structure of self-adjoint and, more generally, normal operators on a Hilbert space. It roughly says that such an operator is, up to unitary equivalence, a "multiplication operator with multiplicity" with a simple explicit description. The name "Spectral Theorem" comes from the fact that the description is expressed in terms of the spectrum of the operator.

XX.14.1. The Finite-Dimensional Spectral Theorem

We first discuss the Spectral Theorem on a finite-dimensional Hilbert space. As might be expected, this situation is pure Linear Algebra with no essential use of functional analysis.

XX.14.1.1. Theorem. [Finite-Dimensional Spectral Theorem] Let $\mathcal{H}$ be a finite-dimensional Hilbert space, and T a normal operator on $\mathcal{H}$. Then there is an orthonormal basis for $\mathcal{H}$ consisting of eigenvectors of T.

XX.14.1.2. Corollary. Let $A \in \mathbb{M}_n = M_n(\mathbb{C})$ be a normal matrix. Then A is unitarily equivalent to a diagonal matrix.

XX.14.1.3. The converse is true too: if $A \in \mathbb{M}_n$ is unitarily equivalent to a diagonal matrix, then A is normal.

We now give the proof of Theorem XX.14.1.1..

Proof: The proof of XX.14.1.1. is a simple complete induction (infinite regress), based on () and the elementary facts that if T is normal, then so is $T - \lambda I$ for any $\lambda \in \mathbb{C}$ and that any operator on a finite-dimensional complex vector space has at least one eigenvalue ().

Suppose the result holds for Hilbert spaces of dimension $< n$, and that $\mathcal{H}$ is n-dimensional. Let λ be an eigenvalue of T, and $\mathcal{E}_\lambda$ the corresponding eigenspace, i.e.

$$\mathcal{E}_\lambda = \{x \in \mathcal{H} : Tx = \lambda x\} \ .$$

Then $\mathcal{E}_\lambda = \mathcal{N}(T - \lambda I)$ is a nontrivial subspace of $\mathcal{H}$, so $\mathcal{E}_\lambda^\perp = \mathcal{R}(T - \lambda I)$ () is a subspace of $\mathcal{H}$ of dimension $< n$ (recall that every subspace of $\mathcal{H}$ is closed). If $x \in \mathcal{E}_\lambda^\perp$, then $x = (T - \lambda I)y$ for some y, and hence

$$Tx = T(T - \lambda I)y = (T - \lambda I)(Ty) \in \mathcal{R}(T - \lambda I) = \mathcal{E}_\lambda^\perp \ .$$

So T maps $\mathcal{E}_\lambda^\perp$ into itself, and so $S = T|_{\mathcal{E}_\lambda^\perp}$ is an operator on $\mathcal{E}_\lambda^\perp$. It is easily checked that the adjoint of S is $T^*|_{\mathcal{E}_\lambda^\perp}$, since $\mathcal{E}_\lambda = \mathcal{N}((T - \lambda I)^*)$ and $\mathcal{E}_\lambda^\perp = \mathcal{R}((T - \lambda I)^*)$ because $T - \lambda I$ is normal, and therefore $\mathcal{E}_\lambda$ and $\mathcal{E}_\lambda^\perp$ are invariant under T^* by the same argument as for T. Thus S is normal, so by the inductive assumption there is an orthonormal basis for $\mathcal{E}_\lambda^\perp$ consisting of eigenvectors of S, and putting this basis together with any orthonormal basis of $\mathcal{E}_\lambda$ gives an orthonormal basis for $\mathcal{H}$ consisting of eigenvectors of T. 👍

XX.14.1.4. The orthonormal basis in XX.14.1.1. is not unique. However, the decomposition of $\mathcal{H}$ into eigenspaces of T is unique, and is a more natural way to phrase the Spectral Theorem:

XX.14.1.5. Corollary. Let $\mathcal{H}$ be a finite-dimensional Hilbert space, and T a normal operator on $\mathcal{H}$. Let $\lambda_1, \ldots, \lambda_m$ be the distinct eigenvalues of T, and $\mathcal{E}_{\lambda_1}, \ldots, \mathcal{E}_{\lambda_m}$ the corresponding eigenspaces. Then

(i) $\mathcal{E}_{\lambda_j} \perp \mathcal{E}_{\lambda_k}$ for $j \neq k$.

(ii) The subspaces $\{\mathcal{E}_{\lambda_1}, \ldots, \mathcal{E}_{\lambda_m}\}$ span $\mathcal{H}$.

Thus $\mathcal{H} \cong \mathcal{E}_{\lambda_1} \oplus \cdots \oplus \mathcal{E}_{\lambda_m}$.

There is an alternate phrasing in terms of projection operators. For each k, let E_k be the orthogonal projection of $\mathcal{H}$ onto $\mathcal{E}_{\lambda_k}$. Then:

XX.14.1.6. Corollary. Let $\mathcal{H}$ be a finite-dimensional Hilbert space, and T a normal operator on $\mathcal{H}$. Let $\lambda_1, \ldots, \lambda_m$ be the distinct eigenvalues of T. Then there are unique projections $E_1, \ldots, E_m \in \mathcal{B}(\mathcal{H})$ such that

(i) $E_j \perp E_k$ for $j \neq k$.

(ii) $E_1 + \cdots + E_m = I$.

(iii) $E_k T = T E_k = E_k T E_k = \lambda_k E_k$ for each k.

(iv) $\lambda_1 E_1 + \cdots + \lambda_m E_m = T$.

Furthermore, the E_k are nonzero. The expression in (iv) is called the *spectral resolution* of T.

XX.14.1.7. This representation can be extended. If f is a polynomial with complex coefficients, and $f(T)$ is defined in the standard way (), then an easy calculation shows that

$$f(T) = f(\lambda_1)E_1 + \cdots + f(\lambda_m)E_m$$

where $E_1, \ldots, E_m$ are as in XX.14.1.6.. Since $f(T)$ is also normal, and the eigenvalues of $f(T)$ are $f(\lambda_1), \ldots, f(\lambda_m)$, then this expression is almost the spectral resolution of $f(T)$, except that the $f(\lambda_k)$ need not be distinct and if not some terms in this sum could be combined.

XX.14.1.8. In particular, if f_k is a polynomial with $f_k(\lambda_k) = 1$ and $f_k(\lambda_j) = 0$ for $j \neq k$ (cf. V.11.1.3.), then $f_k(T) = E_k$. Thus the E_k are expressible as polynomials in T.

XX.14.1.9. Conversely, an operator on a finite-dimensional $\mathcal{H}$ with an orthonormal basis of eigenvectors is unitarily equivalent to a diagonal operator, hence normal. A nonnormal operator on $\mathbb{C}^n$ may or may not have n linearly independent eigenvectors (), and even if it does they cannot be orthonormal.

XX.15. L^p-Spaces

In this section, we define and prove the basic properties of the L^p-spaces of a measure space. These are very important examples of Banach spaces, especially L^2-spaces, which are Hilbert spaces.

XX.15.1. Conjugate Exponents

This subsection simply sets up some simple notation and terminology which will prove very useful.

XX.15.1.1. DEFINITION. Let p be a real number greater than 1. The number $q = \frac{p}{p-1}$ is called the *conjugate exponent* to p. The pair (p, q) is a *conjugate pair of exponents*.

If $p > 1$, then $q > 1$ also. This definition is symmetric in p and q, as is immediate from the next proposition, which gives several equivalent versions of the defining equality. The proof is a simple exercise in high-school algebra.

XX.15.1.2. PROPOSITION. Let $p, q > 1$. The following are equivalent:

(i) q is the conjugate exponent to p, i.e. $q = \frac{p}{p-1}$.

(ii) p is the conjugate exponent to q, i.e. $p = \frac{q}{q-1}$.

(iii) $\frac{1}{p} + \frac{1}{q} = 1$.

(iv) $pq - q = p$.

(v) $p - \frac{p}{q} = 1$.

(vi) $\frac{p}{q} = p - 1$.

(vii) $(p-1)(q-1) = 1$.

Note that if $1 < p < 2$, then $q > 2$, and if $p = 2$ then $q = 2$ (i.e. 2 is "self-conjugate"). As p increases, q decreases; as $p \to 1$, $q \to \infty$ (and vice versa). We extend the definition of a conjugate pair to include the pairs $(1, \infty)$ and $(\infty, 1)$, i.e. the conjugate exponent to 1 is ∞ and vice versa. These pairs behave somewhat differently than the pairs for $1 < p < \infty$, however, and some results do not apply to them.

XX.15.2. $\mathcal{L}^p$-Spaces, Norms, and Inequalities

XX.15.2.1. DEFINITION. Let $(X, \mathcal{A}, \mu)$ be a measure space and $0 < p < \infty$. If f is an $\mathcal{A}$-measurable function from X to $\mathbb{C}$, set

$$J_p(f) = \int_X |f|^p \, d\mu \in [0, \infty]$$

$$\|f\|_p = [J_p(f)]^{1/p} = \left[\int_X |f|^p \, d\mu\right]^{1/p}$$

(where $\infty^{1/p} = \infty$).

We let $\mathcal{L}^p_{\mathbb{C}}(X,\mathcal{A},\mu)$ (or $\mathcal{L}^p_{\mathbb{C}}(\mu)$ if X and $\mathcal{A}$ are understood) be the set of measurable functions from X to $\mathbb{C}$ such that $J_p(f) < \infty$. $\mathcal{L}^p_{\mathbb{R}}(X,\mathcal{A},\mu)$ is defined similarly. We let $\mathcal{L}^p(X,\mathcal{A},\mu)$ stand for either $\mathcal{L}^p_{\mathbb{C}}(X,\mathcal{A},\mu)$ or $\mathcal{L}^p_{\mathbb{R}}(X,\mathcal{A},\mu)$ when the base field is unimportant (it must of course always be specified and kept in mind).

The space $\mathcal{L}^p(\mu)$ is really only interesting if μ is σ-finite or at least semifinite, since it is rather degenerate otherwise (XX.15.5.7., Exercise XX.15.9.4.–XX.15.9.5.). However, the definition and basic properties need no assumptions about the measure space.

The next proposition records some obvious properties of the $\mathcal{L}^p$-spaces.

XX.15.2.2. PROPOSITION. Let $(X,\mathcal{A},\mu)$ be a measure space and $0 < p < \infty$. Then

(i) $\mathcal{L}^p_{\mathbb{R}}(X,\mathcal{A},\mu) \subseteq \mathcal{L}^p_{\mathbb{C}}(X,\mathcal{A},\mu)$.

(ii) If $f \in \mathcal{L}^p(X,\mathcal{A},\mu)$, g is $\mathcal{A}$-measurable, and $|g| \le |f|$ a.e., then $g \in \mathcal{L}^p(X,\mathcal{A},\mu)$. In particular, if g is $\mathcal{A}$-measurable, then $g \in \mathcal{L}^p(X,\mathcal{A},\mu)$ if and only if $|g| \in \mathcal{L}^p(X,\mathcal{A},\mu)$.

(iii) If $r > 0$ and $f \in \mathcal{L}^p(X,\mathcal{A},\mu)$, then $|f|^{p/r} \in \mathcal{L}^r(X,\mathcal{A},\mu)$ and $\||f|^{p/r}\|_r = \|f\|_p^{p/r}$. In particular, if $p > 1$ and q is the conjugate exponent to p, and $f \in \mathcal{L}^p(X,\mathcal{A},\mu)$, then $|f|^{p-1} \in \mathcal{L}^q(X,\mathcal{A},\mu)$ and $\||f|^{p-1}\|_q = \|f\|_p^{p-1}$ (XX.15.1.2.(vi)).

(iv) If $f : X \to \mathbb{C}$ is $\mathcal{A}$-measurable, then $f \in \mathcal{L}^p_{\mathbb{C}}(X,\mathcal{A},\mu)$ if and only if the real and imaginary parts of f are in $\mathcal{L}^p_{\mathbb{R}}(X,\mathcal{A},\mu)$.

We will eventually show that $\|\cdot\|_p$ is a seminorm on $\mathcal{L}^p(\mu)$ if $p \ge 1$. The triangle inequality (which only holds for $p \ge 1$) will take some work. We first get a weak form of the triangle inequality which will be enough to show that $\mathcal{L}^p(\mu)$ is a vector space, even for $0 < p < 1$:

XX.15.2.3. PROPOSITION. Let $(X,\mathcal{A},\mu)$ be a measure space, $0 < p < \infty$, and $f,g \in \mathcal{L}^p_{\mathbb{C}}(X,\mathcal{A},\mu)$, $\alpha \in \mathbb{C}$. Then

$$J_p(f+g) \le 2^p[J_p(f) + J_p(g)]$$
$$J_p(\alpha f) = |\alpha|^p J_p(f)\ .$$

In particular, $f+g$ and αf are in $\mathcal{L}^p_{\mathbb{C}}(X,\mathcal{A},\mu)$, and $\mathcal{L}^p_{\mathbb{C}}(X,\mathcal{A},\mu)$ is a complex vector space (hence $\mathcal{L}^p_{\mathbb{R}}(X,\mathcal{A},\mu)$ is a real vector space).

PROOF: The statement for αf is obvious. For the first statement, we have

$$\begin{aligned} J_p(f+g) &= \int_X |f+g|^p\,d\mu \le \int_X (|f|+|g|)^p\,d\mu \le \int_X (2\max(|f|,|g|))^p\,d\mu \\ &= 2^p \int (\max(|f|^p,|g|^p)\,d\mu \le 2^p \int_X (|f|^p + |g|^p)\,d\mu \\ &= 2^p \left[\int_X |f|^p\,d\mu + \int_X |g|^p\,d\mu\right] = 2^p[J_p(f) + J_p(g)]\ . \end{aligned}$$

In order to prove the triangle inequality for $\|\cdot\|_p$ $(1 < p < \infty)$, we need to establish some inequalities: Young's Inequality for products, Hölder's Inequality, and finally Minkowski's Inequality. We begin with Young's Inequality for products after observing a simple preliminary lemma:

XX.15.2.4. LEMMA. Let $x, y \in (0, \infty)$ and $0 \leq t \leq 1$. Then

$$t \log x + (1-t) \log y \leq \log(tx + (1-t)y) \ .$$

We have equality if and only if $x = y$ or $t = 0$ or 1.

The proof is essentially just the observation that the graph of $\log x$ is (strictly) concave downward.

XX.15.2.5. PROPOSITION. [YOUNG'S INEQUALITY FOR PRODUCTS] Let $\alpha, \beta \in [0, \infty)$ and (p, q) a conjugate pair of exponents with $1 < p < \infty$. Then

$$\alpha\beta \leq \frac{1}{p}\alpha^p + \frac{1}{q}\beta^q$$

with equality if and only if $\alpha^p = \beta^q$.

PROOF: If α or β is 0, the result is trivial. Otherwise, we have

$$\log(\alpha\beta) = \log\alpha + \log\beta = \frac{1}{p}\log(\alpha^p) + \frac{1}{q}\log(\beta^q) \leq \log\left(\frac{1}{p}\alpha^p + \frac{1}{q}\beta^q\right)$$

where the inequality comes from Lemma XX.15.2.4. with $x = \alpha^p$, $y = \beta^q$, $t = \frac{1}{p}$. Since the log function is strictly increasing, the result follows. 👍

See Exercise XX.15.9.1. for an equivalent statement and another proof of Young's Inequality for products, and Exercise XX.15.9.2. for a more general version. There is another inequality due to YOUNG for convolutions ().

We now obtain Hölder's Inequality, which is a generalization of the CBS Inequality (XX.11.2.1.). This inequality will not only be used in this section, but will be important in other applications.

XX.15.2.6. PROPOSITION. [HÖLDER'S INEQUALITY] Let $(X, \mathcal{A}, \mu)$ be a measure space, (p, q) a conjugate pair with $1 < p < \infty$, $f \in \mathcal{L}^p(\mu)$, and $g \in \mathcal{L}^q(\mu)$. Then $fg \in \mathcal{L}^1(\mu)$ and

$$\int_X |fg| \, d\mu \leq \|f\|_p \|g\|_q \ .$$

We have equality if and only if $|g|^q = \alpha|f|^p$ a.e. for some $\alpha > 0$, or if $f = 0$ a.e. or $g = 0$ a.e.

PROOF: If $f = 0$ a.e. or $g = 0$ a.e. both sides are 0, so equality holds. Otherwise, let $F = \frac{1}{\|f\|_p} f$, $G = \frac{1}{\|g\|_q} g$. Then $\|F\|_p = \|G\|_q = 1$, and we have, using Young's Inequality for products,

$$\frac{1}{\|f\|_p \|g\|_q} \int_X |fg| \, d\mu = \int_X |FG| \, d\mu \leq \frac{1}{p}\int_X |F|^p \, d\mu + \frac{1}{q}\int_X |G|^q \, d\mu = \frac{1}{p} + \frac{1}{q} = 1 \ .$$

We have equality if and only if $|F|^p = |G|^q$ a.e. Note that multiplying f or g by a nonzero scalar does not change $|F|$ and $|G|$. 👍

We can now prove the triangle inequality for $\| \cdot \|_p$ for $1 \leq p < \infty$. This inequality is usually called *Minkowski's Inequality* (see XX.15.9.11. for a more general version).

XX.15.2.7. THEOREM. [MINKOWSKI'S INEQUALITY] Let $(X, \mathcal{A}, \mu)$ be a measure space, and $1 \leq p < \infty$. If $f, g \in \mathcal{L}^p(X, \mathcal{A}, \mu)$, then $f + g \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ and

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p \ .$$

PROOF: The case $p = 1$ is essentially just the usual triangle inequality for numbers, and is left to the reader. So we assume $p > 1$.

We have already noted that $f + g \in \mathcal{L}^p(\mu)$ (XX.15.2.3.). Let q be the conjugate exponent to p. Then by XX.15.2.2.(iii), $|f + g|^{p-1} \in \mathcal{L}^q(\mu)$ and

$$\||f + g|^{p-1}\|_q = \|f + g\|_p^{p/q} = \|f + g\|_p^{p-1} \ .$$

Next note that

$$|f + g|^p = |f + g||f + g|^{p-1} \leq (|f| + |g|)|f + g|^{p-1} = |f||f + g|^{p-1} + |g||f + g|^{p-1} \ .$$

Thus, by Hölder's Inequality,

$$\|f + g\|_p^p = \int_X |f + g|^p \, d\mu \leq \int_X |f||f + g|^{p-1} \, d\mu + \int_X |g||f + g|^{p-1} \, d\mu$$

$$\leq \|f\|_p \||f + g|^{p-1}\|_q + \|g\|_p \||f + g|^{p-1}\|_q = (\|f\|_p + \|g\|_p)\|f + g\|_p^{p-1} \ .$$

If $\|f + g\|_p = 0$, the result is trivial; otherwise, divide by $\|f + g\|_p^{p-1}$. 👍

XX.15.2.8. COROLLARY. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $1 \leq p < \infty$. Then $\| \cdot \|_p$ is a seminorm on $\mathcal{L}^p(X, \mathcal{A}, \mu)$.

PROOF: It is obvious that $\|f\|_p \in [0, \infty)$ if $f \in \mathcal{L}^p(\mu)$, and it follows from XX.15.2.3. that if $f \in \mathcal{L}^p(\mu)$ and α is a scalar, then $\|\alpha f\|_p = |\alpha| \|f\|_p$ (these statements are also true if $0 < p < 1$). Minkowski's Inequality gives the triangle inequality (for $1 \leq p < \infty$). 👍

This seminorm is usually not a norm: $\|f\|_p = 0$ if (and only if) $f = 0$ a.e. We will remedy this in the next section.

The function $\| \cdot \|_p$ does not satisfy the triangle inequality for $0 < p < 1$ in general:

XX.15.2.9. EXAMPLE. Let $X = \{a, b\}$ be a two-point space, and μ counting measure on $(X, \mathcal{P}(X))$. Let $f(a) = 1$, $f(b) = 0$, $g(a) = 0$, $g(b) = 1$. For any p, $0 < p < \infty$, we have $\|f\|_p = \|g\|_p = 1$, $\|f + g\|_p = 2^{1/p}$. So if $p < 1$, $\|f + g\|_p > 2$ and $\| \cdot \|_p$ does not satisfy the triangle inequality.

XX.15.3. L^p-Spaces

Definition as a quotient, completeness, density of simple functions, density of continuous functions for Borel measures

Let $(X, \mathcal{A}, \mu)$ be a measure space. Since $\|\cdot\|_p$ is usually only a seminorm and not a norm on $\mathcal{L}^p(X, \mathcal{A}, \mu)$, by the usual process () we reduce to the quotient by the subspace of vectors of seminorm 0, which in this case is the set of functions which are 0 a.e. This reduction effectively means that we identify functions which are equal a.e.

XX.15.3.1. Definition. Let $0 \le p < \infty$. $L^p(X, \mathcal{A}, \mu)$ (denoted just $L^p(\mu)$ when X and $\mathcal{A}$ are understood) is the quotient of $\mathcal{L}^p(X, \mathcal{A}, \mu)$ by the subspace

$$\{f \in \mathcal{L}^p(X, \mathcal{A}, \mu) : \|f\|_p = 0\} = \{f : f \text{ measurable}, f = 0 \text{ a.e.}\} .$$

(As with $\mathcal{L}^p$, L^p can stand for either $L^p_{\mathbb{R}}$ or $L^p_{\mathbb{C}}$; the base field should always be specified in a specific context.)

If $f \in \mathcal{L}^p(\mu)$, we denote by $[f]$ the image of f in $L^p(\mu)$:

$$[f] = \{g \in \mathcal{L}^p(\mu) : f = g \text{ a.e.}\} .$$

As in (), the seminorm $\|\cdot\|_p$ on $\mathcal{L}^p(\mu)$ drops to a norm on $L^p(\mu)$, also denoted $\|\cdot\|_p$, by $\|[f]\|_p = \|f\|_p$.

Many authors are rather sloppy in distinguishing between $\mathcal{L}^p(\mu)$ and $L^p(\mu)$, and functional analysts tend to be rather casual about the difference; we will eventually follow this convention of thinking of elements of $L^p(\mu)$ as functions rather than equivalence classes of functions, but in this section we will make a careful distinction in establishing the basic properties of these spaces (which can be used to carefully justify the casual convention).

The most important fact about L^p-spaces is that they are Banach spaces:

XX.15.3.2. Theorem. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $1 \le p < \infty$. If (f_k) is a Cauchy sequence in $\mathcal{L}^p(X, \mathcal{A}, \mu)$, then there is a subsequence of (f_k) converging pointwise a.e. to an $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ such that $f_k \to f$ in $\mathcal{L}^p(X, \mathcal{A}, \mu)$. In particular, $L^p(X, \mathcal{A}, \mu)$ is complete.

Proof: A Cauchy sequence with a convergent subsequence converges to the same limit (). Thus, passing to a subsequence, we may assume $\|f_{k+1} - f_k\|_p \le 2^{-k}$ for each k, and we will show that (f_k) converges pointwise a.e. to a function f, and $\|f_k - f\|_p \to 0$.

For each n, let $g_n = f_1 + \sum_{k=1}^n |f_{k+1} - f_k|$. Then (g_n) is an increasing sequence; let $g = \lim g_n = \sup g_n$. Then g is an extended nonnegative real-valued measurable function on $(X, \mathcal{A})$. For each n we have

$$\int_X g_n^p \, d\mu = \|g_n\|_p^p \le \left[\|f_1\|_p + \sum_{k=1}^n \|f_{k+1} - f_k\|_p \right]^p \le (1 + \|f_1\|_p)^p$$

and so, by the Monotone Convergence Theorem,

$$\int_X g^p \, d\mu = \sup_n \int_X g_n^p \, d\mu \le (1 + \|f_1\|_p)^p < \infty .$$

Thus, although g may technically not be in $\mathcal{L}^p(X, \mathcal{A}, \mu)$ since it may take the value $+\infty$, it is finite a.e. and equal a.e. to a function $h \in \mathcal{L}^p(X, \mathcal{A}, \mu)$. For any $x \in X$ with $g(x) < \infty$ the series $f_{k+1}(x) = f_1(x) + \sum_{k=1}^{\infty}[f_{k+1}(x) - f_k(x)]$ of scalars converges absolutely, hence converges; call the sum $f(x)$. Extend f to be 0 on the x's for which $g(x) = +\infty$. Then f is an $\mathcal{A}$-measurable function from X to $\mathbb{C}$, $f_k \to f$ pointwise a.e., and $|f| \leq |h|$ a.e., so $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$. We also have

$$|f - f_k|^p \leq |h|^p \text{ a.e.}$$

for all k, and $|h|^p \in \mathcal{L}^1(X, \mathcal{A}, \mu)$; since $|f - f_k|^p \to 0$ pointwise a.e., by the Dominated Convergence Theorem we have

$$\lim_{k\to\infty} \|f - f_k\|_p^p = \lim_{k\to\infty} \left[\int_X |f - f_k|^p \ d\mu\right] = 0 \ .$$

It is interesting to note that, although it is somewhat easier to prove that $\|\cdot\|_p$ is a seminorm (i.e. satisfies the triangle inequality) for $p = 2$ than for general p (cf. ()), and considerably easier for $p = 1$, it is no easier to prove completeness of L^p for $p = 1$ or $p = 2$ than for general p.

XX.15.3.3. An important observation from XX.15.3.2., stated a bit imprecisely, is that if $f_n \to f$ in $L^p(X, \mathcal{A}, \mu)$, then there is a subsequence of (f_n) converging pointwise a.e. to f. It is necessary to pass to subsequences in general if $p < \infty$ (Exercise XX.15.9.12.). If $f_n \to f$ in $L^p(X, \mathcal{A}, \mu)$, then $f_n \to f$ in measure ().

It is usually important to identify some dense subsets of "nice" elements in any Banach space, and this is particularly important in the L^p-spaces. For example, we will show () that in L^p-spaces arising from Borel measures on good topological spaces, the continuous functions are dense in L^p for any p. One general and very useful fact is true in any L^p-space:

XX.15.3.4. THEOREM. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $1 \leq p < \infty$. Then the (classes of) simple functions in $L^p(X, \mathcal{A}, \mu)$ are dense in $L^p(X, \mathcal{A}, \mu)$.

PROOF: By working with real and imaginary parts, it suffices to prove the result for $L^p_{\mathbb{R}}(\mu)$. If $f \in \mathcal{L}^p_{\mathbb{R}}(\mu)$, then by () there is a sequence of real-valued simple functions (s_n) converging pointwise to f, with $|s_n| \leq |f|$ for all n. Thus each $s_n \in \mathcal{L}^p_{\mathbb{R}}(\mu)$. We have $|s_n - f|^p \leq |2f|^p$ for all n, and $|2f|^p \in \mathcal{L}^1(\mu)$. Since $|s_n - f|^p \to 0$ pointwise, by the Dominated Convergence Theorem we have

$$\lim_{n\to\infty} \|s_n - f\|_p^p = \lim_{n\to\infty} \left[\int_X |s_n - f|^p \, d\mu\right] = 0 \ .$$

XX.15.4. $\mathcal{L}^\infty$- and L^∞-Spaces

XX.15.5. Comparison Between $\mathcal{L}^p$-Spaces and L^p-Spaces

Let $(X, \mathcal{A}, \mu)$ be a measure space. There is no containment between the $\mathcal{L}^p$-spaces or the L^p-spaces of $(X, \mathcal{A}, \mu)$ as p varies, except in certain special cases. There are two competing constraints which can cause a (measurable) function f to fail to be in $\mathcal{L}^p(X, \mathcal{A}, \mu)$:

(i) f can go to zero "too slowly" (or not at all) at "infinity."

(ii) f can go to infinity "too rapidly" on a small portion of X.

The first problem gets worse as p decreases, and the second problem gets worse as p increases. We will make these problems precise in this subsection, and give examples.

Consider the first constraint.

XX.15.5.1. Example. Suppose $(X, \mathcal{A}, \mu)$ contains measurable subsets of arbitrarily large finite measure (e.g. μ is σ-finite but not finite). Choose an increasing sequence (B_n) in $\mathcal{A}$ with $\mu(B_1) \geq 1$ and $\mu(B_n) + 1 \leq \mu(B_{n+1}) < \infty$ for all $n > 1$. Then, by taking successive differences, there is a sequence (A_n) of pairwise disjoint sets in $\mathcal{A}$ with $1 \leq \mu(A_n) < \infty$ for all n. Fix p, $0 < p < \infty$, and define a function $f : X \to [0, 1]$ by setting

$$f(x) = \frac{1}{[n\mu(A_n)]^{1/p}} \text{ if } x \in A_n$$

$$f(x) = 0 \text{ if } x \notin \cup_n A_n \ .$$

Then f is measurable, and we have

$$J_p(f) = \sum_{n=1}^{\infty} \frac{1}{n\mu(A_n)} \mu(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

so $f \notin \mathcal{L}^p(X, \mathcal{A}, \mu)$. However, if $p < r < \infty$, we have

$$J_r(f) = \sum_{n=1}^{\infty} \frac{1}{n^{r/p}\mu(A_n)^{r/p}} \mu(A_n) = \sum_{n=1}^{\infty} \frac{1}{n^{r/p}\mu(A_n)^{r/p-1}} \leq \sum_{n=1}^{\infty} \frac{1}{n^{r/p}} < \infty$$

so $f \in \cap_{r>p} \mathcal{L}^r(X, \mathcal{A}, \mu)$. (Note that f is bounded, so $f \in \mathcal{L}^\infty(X, \mathcal{A}, \mu)$ also.) In particular, $\mathcal{L}^r(X, \mathcal{A}, \mu)$ is not contained in $\mathcal{L}^p(X, \mathcal{A}, \mu)$ for $p < r \leq \infty$.

In the absence of such subsets, we have a containment:

XX.15.5.2. Proposition. Let $(X, \mathcal{A}, \mu)$ be a finite measure space, and let $0 < p \leq r \leq \infty$. Then $\mathcal{L}^r(X, \mathcal{A}, \mu)$ is contained in $\mathcal{L}^p(X, \mathcal{A}, \mu)$.

Proof: If $r = \infty$, the result holds since any bounded measurable function on a finite measure space is integrable. So assume $r < \infty$ and let $f \in \mathcal{L}^r(X, \mathcal{A}, \mu)$. Set

$$A = \{x \in X : |f(x)| \geq 1\}$$

$$B = \{x \in X : |f(x)| < 1\} \ .$$

Then $|f|^p \le |f|^r$ on A, so we have

$$J_p(f) = \int_X |f|^p \, d\mu = \int_A |f|^p \, d\mu + \int_B |f|^p \, d\mu \le \int_A |f|^r \, d\mu + \int_B |f|^p \, d\mu$$

$$\le J_r(f) + \mu(B) < \infty$$

so $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$. 👍

Now we turn to the second constraint.

XX.15.5.3. EXAMPLE. Let $(X, \mathcal{A}, \mu)$ be a measure space which contains measurable sets of arbitrarily small positive measure. Let $B_1 \in \mathcal{A}$ with $0 < \mu(B_1) \le 1$, and inductively let $B_n \in \mathcal{A}$ with $0 < \mu(B_n) \le \frac{1}{3}\mu(B_{n-1})$ for $n > 1$. Set $A_n = B_n \setminus (\cup_{k=n+1}^{\infty} B_k)$. Then the A_n are pairwise disjoint, and since

$$\mu(\cup_{k=n+1}^{\infty} B_k) \le \sum_{k=n+1}^{\infty} \mu(B_k) \le \sum_{m=1}^{\infty} 3^{-m}\mu(B_n) = \frac{\mu(B_n)}{2}$$

we have $0 < \frac{1}{2}\mu(B_n) \le \mu(A_n)$. Also, $\mu(A_n) \le \mu(B_n) \le 3^{-n+1}$.

Fix p, $0 < p < \infty$. Define a function $f : X \to [0, \infty)$ by

$$f(x) = [\mu(A_n)]^{-1/p} \text{ if } x \in A_n$$

$$f(x) = 0 \text{ if } x \notin \cup_n A_n \ .$$

Then we have

$$J_p(f) = \sum_{n=1}^{\infty} [\mu(A_n)]^{-1} \mu(A_n) = \infty$$

so $f \notin \mathcal{L}^p(X, \mathcal{A}, \mu)$. But if $0 < r < p$, we have

$$J_r(f) = \sum_{n=1}^{\infty} [\mu(A_n)]^{-r/p} \mu(A_n) = \sum_{n=1}^{\infty} [\mu(A_n)]^{1-r/p} \le \sum_{n=1}^{\infty} [3^{1-r/p}]^{-n+1} < \infty$$

so $f \in \cap_{r<p} \mathcal{L}^r(X, \mathcal{A}, \mu)$. In particular, $\mathcal{L}^r(X, \mathcal{A}, \mu)$ is not contained in $\mathcal{L}^p(X, \mathcal{A}, \mu)$ for any $r < p$. Also, f is not essentially bounded, so $f \notin \mathcal{L}^\infty(X, \mathcal{A}, \mu)$, i.e. $\mathcal{L}^p(X, \mathcal{A}, \infty)$ is not contained in $\mathcal{L}^\infty(X, \mathcal{A}, \mu)$.

We again have a containment in the absence of this situation:

XX.15.5.4. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space. Suppose there is an $\epsilon > 0$ such that, for every $A \in \mathcal{A}$, either $\mu(A) = 0$ or $\mu(A) \geq \epsilon$. Let $0 < r \leq p \leq \infty$. Then $\mathcal{L}^r(X, \mathcal{A}, \mu)$ is contained in $\mathcal{L}^p(X, \mathcal{A}, \mu)$.

PROOF: Let $f \in \mathcal{L}^r(\mu)$. We first show that f is essentially bounded. We may assume $\|f\|_r > 0$, since otherwise $f = 0$ a.e. Set $M = \frac{\|f\|_r}{\epsilon^{1/r}}$ and let $C = \{x \in X : |f(x)| > M\}$. Since

$$M^r \mu(C) = \frac{\|f\|_r^r}{\epsilon} \mu(C) < \int_C |f|^r \, d\mu \leq \int_X |f|^r \, d\mu = \|f\|_r^r$$

we have $\mu(C) < \epsilon$, so $\mu(C) = 0$, i.e. $|f| \leq M$ a.e. and $f \in \mathcal{L}^\infty(X, \mathcal{A}, \mu)$. If $p < \infty$, let $A = \{x : |f(x)| \geq 1\}$ and $B = \{x : |f(x)| < 1\}$. Then $|f|^p < |f|^r$ on B, and

$$\mu(A) \leq \int_A |f|^r \, d\mu \leq \int_X |f|^r \, d\mu = J_r(f) < \infty$$

and we have

$$J_p(f) = \int_X |f|^p \, d\mu = \int_A |f|^p \, d\mu + \int_B |f|^p \, d\mu$$

$$< \int_A |f|^p \, d\mu + \int_B |f|^r \, d\mu \leq M^p \mu(A) + J_r(f) < \infty$$

so $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$. 👍

XX.15.5.5. EXAMPLES. (a) A set with counting measure satisfies the hypotheses of XX.15.5.4. (with $\epsilon = 1$). Thus $\ell^r(X) \subseteq \ell^p(X)$ for any X if $r < p$.

(b) If $(X, \mathcal{A}, \mu)$ is a measure space containing both sets of arbitrarily large and arbitrarily small finite positive measure (e.g. $(\mathbb{R}, \mathcal{M}, \lambda)$), then there is no containment either way between $\mathcal{L}^p(X, \mathcal{A}, \mu)$ and $\mathcal{L}^r(X, \mathcal{A}, \mu)$ for $p \neq r$.

There is, however, a large subspace which is common to $\mathcal{L}^p(\mu)$ for all p. A measurable function f has *support of finite measure* if $\{x : f(x) \neq 0\}$ has finite measure. (Some authors say that such a function has *finite support*, but this term, while more efficient, is ambiguous.) We can similarly define functions with support of σ-finite measure.

If $(X, \mathcal{A}, \mu)$ is a measure space, let $\mathcal{L}^{\text{oo}}(X, \mathcal{A}, \mu)$ (distinguish between oo and ∞!) be the set of essentially bounded measurable functions with support of finite measure. The characteristic function of a set of finite measure is in $\mathcal{L}^{\text{oo}}(X, \mathcal{A}, \mu)$, hence so is any linear combination of such functions (simple functions with support of finite measure).

XX.15.5.6. PROPOSITION. $\mathcal{L}^{\text{oo}}(X, \mathcal{A}, \mu) \subseteq \mathcal{L}^p(X, \mathcal{A}, \mu)$ for $0 < p \leq \infty$, and $\mathcal{L}^{\text{oo}}(X, \mathcal{A}, \mu)$ is dense in $\mathcal{L}^p(X, \mathcal{A}, \mu)$ if $p < \infty$.

The density for $p < \infty$ is immediate from the proof of XX.15.3.4., since a simple function which is in $\mathcal{L}^p$ for some p must be in $\mathcal{L}^{\text{oo}}$ (Exercise ()).

$\mathcal{L}^{\infty\circ}(X, \mathcal{A}, \mu)$ is not dense in $\mathcal{L}^{\infty}(X, \mathcal{A}, \mu)$ unless μ is finite, in which case $\mathcal{L}^{\infty\circ}(X, \mathcal{A}, \mu) = \mathcal{L}^{\infty}(X, \mathcal{A}, \mu)$. We say a function f has *support of almost finite measure* if $\{x : |f(x)| > \epsilon\}$ has finite measure for all $\epsilon > 0$. This implies, but is not equivalent to, that f has support of σ-finite measure (consider the constant function 1 on $(\mathbb{R}, \mathcal{L}, \lambda)$). It is easily shown (Exercise ()) that the closure of $\mathcal{L}^{\infty\circ}(X, \mathcal{A}, \mu)$ in $\mathcal{L}^{\infty}(X, \mathcal{A}, \mu)$ is the set of essentially bounded functions with support of almost finite measure.

It is worth noting the next result:

XX.15.5.7. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space and $0 < p < \infty$. Then any function in $\mathcal{L}^p(X, \mathcal{A}, \mu)$ has support of almost finite measure (in particular, of σ-finite measure).

This proposition is false for $\mathcal{L}^{\infty}(X, \mathcal{A}, \mu)$ if μ is not finite.

XX.15.5.8. REMARK. For a fixed measure space $(X, \mathcal{A}, \mu)$, $L^p(X, \mathcal{A}, \mu)$ is constructed from $\mathcal{L}^p(X, \mathcal{A}, \mu)$ by dividing out by the same subspace of functions for every p. Thus if $\mathcal{L}^r(X, \mathcal{A}, \mu)$ is contained in $\mathcal{L}^p(X, \mathcal{A}, \mu)$ for some p and r, it is technically strictly correct also that $L^r(X, \mathcal{A}, \mu)$ is contained in $L^p(X, \mathcal{A}, \mu)$. The converse also holds, as is easy to check. Thus, in all the above containment results of this section, $\mathcal{L}^p$ and $\mathcal{L}^r$ can be replaced by L^p and L^r to give strictly correct containment statements.

XX.15.5.9. Now suppose $(X, \mathcal{A}, \mu)$ is a measure space and $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$. Set $\nu = \mu|_{\mathcal{B}}$. It is then obvious that $\mathcal{L}^p(X, \mathcal{B}, \nu)$ is a subspace of $\mathcal{L}^p(X, \mathcal{A}, \mu)$ for any p, and the p-seminorms agree.

However, it is technically not strictly correct in most cases to say that $L^p(X, \mathcal{B}, \nu)$ is a subspace of $L^p(X, \mathcal{A}, \mu)$, since the elements of these spaces are different: if $f \in \mathcal{L}^p(X, \mathcal{B}, \nu)$, the corresponding elements are

$$[f]_{\mathcal{A}} = \{g : g \ \mathcal{A}\text{-measurable}, \ g = f \text{ a.e.}\}$$

$$[f]_{\mathcal{B}} = \{g : g \ \mathcal{B}\text{-measurable}, \ g = f \text{ a.e.}\}$$

which are not the same sets in general. However, the map $[f]_{\mathcal{B}} \mapsto [f]_{\mathcal{A}}$ is a linear isometry from $L^p(X, \mathcal{B}, \nu)$ onto a (complete, hence closed) subspace of $L^p(X, \mathcal{A}, \mu)$, usually called the *inclusion map*, and by making this identification we usually, if rather casually, think of $L^p(X, \mathcal{B}, \nu)$ as a closed subspace of $L^p(X, \mathcal{A}, \mu)$ for any p.

Normally, the inclusion map sends $L^p(X, \mathcal{B}, \nu)$ onto a proper subspace of $L^p(X, \mathcal{A}, \mu)$. But there is one important case where it should be noted that the inclusion map is surjective, hence an isometric isomorphism:

XX.15.5.10. PROPOSITION. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $(X, \mathcal{M}, \bar{\mu})$ its completion. Then, for any p, the inclusion map from $L^p(X, \mathcal{A}, \mu)$ to $L^p(X, \mathcal{M}, \bar{\mu})$ is an isomorphism.

The proof is immediate from the fact that every $\mathcal{M}$-measurable function is equal a.e. to an $\mathcal{A}$-measurable function (). Note that $\mathcal{L}^p(X, \mathcal{M}, \bar{\mu})$ is considerably larger than $\mathcal{L}^p(X, \mathcal{A}, \mu)$ in general.

In particular, $L^p(\mathbb{R}, \mathcal{L}, \lambda)$ is naturally isomorphic to $L^p(\mathbb{R}, \mathcal{B}, \lambda)$ for any p.

Embeddings of Quotient Spaces

XX.15.5.11. In light of these subtleties, it is desirable to have a careful and systematic way of saying one quotient space is "embedded" in another. For another example, we would like to say that $C([0,1]) \subseteq L^p([0,1],\mathcal{L},\lambda)$ for any p, but this does not make sense; what we mean is that $C([0,1]) \subseteq \mathcal{L}^p([0,1],\mathcal{L},\lambda)$ and this embedding passes to the quotient in an obvious way. It will be especially important to have such a notion of embedding for Sobolev spaces (), where the "embeddings" are not generally obvious and sometimes rather tricky.

XX.15.5.12. Instead of working in the utmost generality, we will stick to the case of normed vector spaces made as a quotient from a vector space with a seminorm. So we will notationally use script letters to denote a vector space (over a fixed base field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) with a seminorm and the same Roman letter to denote the normed quotient, for example let $\mathcal{V}$ be a vector space with seminorm $\|\cdot\|_{\mathcal{V}}$, $\mathcal{N}_{\mathcal{V}} = \{v \in \mathcal{V} : \|v\|_{\mathcal{V}} = 0\}$, $V = \mathcal{V}/\mathcal{N}_{\mathcal{V}}$. The prototype example is $\mathcal{V} = \mathcal{L}^p(X,\mathcal{A},\mu)$, $\mathcal{N}_{\mathcal{V}} = \{f \in \mathcal{V} : f = 0 \ \mu\text{-a.e.}\}$, $V = L^p(X,\mathcal{A},\mu)$. And as a special case, if $(V,\|\cdot\|_V)$ is a normed vector space, we can take $\mathcal{V} = V$, $\|\cdot\|_{\mathcal{V}} = \|\cdot\|_V$, $\mathcal{N}_{\mathcal{V}} = \{0\}$.

XX.15.5.13. Definition. Let $(\mathcal{V},\|\cdot\|_{\mathcal{V}})$ and $(\mathcal{W},\|\cdot\|_{\mathcal{W}})$ be vector spaces with seminorms, and $V = \mathcal{V}/\mathcal{N}_{\mathcal{V}}$ and $W = \mathcal{W}/\mathcal{N}_{\mathcal{W}}$ the quotient normed spaces. We say V is *embedded in* W, written $V \hookrightarrow W$, if

(i) $\mathcal{V} \subseteq \mathcal{W}$.

(ii) $\mathcal{N}_{\mathcal{V}} = \mathcal{V} \cap \mathcal{N}_{\mathcal{W}}$ (so there is an injective induced map $\iota : V \to W$).

(iii) There is a constant K such that $\|v\|_{\mathcal{W}} \leq K\|v\|_{\mathcal{V}}$ for all $v \in \mathcal{V}$ (i.e. ι is bounded [continuous]).

XX.15.5.14. As an example, $C([0,1]) \hookrightarrow L^p([0,1],\mathcal{L},\lambda)$ for $1 \leq p \leq \infty$. And if $(X,\mathcal{A},\mu)$ is a measure space, $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{A}$, and $\nu = \mu|_{\mathcal{B}}$, then $L^p(X,\mathcal{B},\nu) \hookrightarrow L^p(X,\mathcal{A},\mu)$ for $1 \leq p \leq \infty$ (in both cases we may take $K = 1$). And:

XX.15.5.15. Proposition. Let $(X,\mathcal{A},\mu)$ be a finite measure space, and let $1 \leq p \leq r \leq \infty$. Then $L^r(X,,\mu) \hookrightarrow L^p(X,\mathcal{A},\mu)$.

Proof: By XX.15.5.2., $\mathcal{L}^r(X,\mathcal{A},\mu) \subseteq \mathcal{L}^p(X,\mathcal{A},\mu)$, and the null spaces coincide. We need only show that there is a K such that $\|f\|_p \leq K\|f\|_r$ for all $f \in \mathcal{L}^r(X,\mathcal{A},\mu)$. It suffices to show this for $\|f\|_r = 1$. So suppose f is such that

$$\|f\|_r = \left[\int_X |f|^r \, d\mu\right]^{1/r} = 1 = \int_X |f|^r \, d\mu \ .$$

Let $A = \{x \in X : |f(x)| \leq 1\}$, $B = X \setminus A = \{x \in X : |f(x)| > 1\}$. Then

$$\int_X |f|^p \, d\mu = \int_A |f|^p \, d\mu + \int_B |f|^p \, d\mu \leq \mu(A) + \int_B |f|^r \, d\mu$$

$$\leq \mu(X) + \int_X |f|^r \, d\mu = \mu(X) + 1$$

so $\|f\|_p \leq (\mu(X)+1)^{1/p}$. 👍

XX.15.5.16. More generally, at least in the case where both $\mathcal{V}$ and $\mathcal{W}$ are subspaces of a larger vector space (e.g. if they are both spaces of $\mathbb{F}$-valued functions on a set X), we will say that $V \hookrightarrow W$ if, writing π for the quotient map from $\mathcal{V}$ to V,

(i) $\pi(\mathcal{V} \cap \mathcal{W}) = V$.

(ii) $\mathcal{N}_{\mathcal{V}} \cap \mathcal{W} = \mathcal{V} \cap \mathcal{N}_{\mathcal{W}}$ (so there is an injective induced map $\iota : V \to W$).

(iii) There is a constant K such that $\|v\|_{\mathcal{W}} \leq K\|v\|_{\mathcal{V}}$ for all $v \in \mathcal{V} \cap \mathcal{W}$ (i.e. ι is bounded [continuous]).

This will be useful, for example, in saying that a space of equivalence classes "consists of" continuous functions, meaning that each equivalence class contains a (unique) continuous function.

XX.15.5.17. Note that if $V \hookrightarrow W$, so that V can be considered to be a subspace of W, the embedding map ι is not generally a topological embedding, i.e. the relative W-topology on V may be strictly weaker than the V-topology. In fact, $\iota(V)$ is frequently dense in W in our applications. And in many applications, ι is actually compact (). In this case, we say V is *compactly embedded* in W, written $V \hookrightarrow_c W$.

XX.15.6. L^p-spaces for Borel Measures

Density of continuous functions, separability, Egorov's Theorem

XX.15.7. The Dual of L^p-Spaces

The dual of (most) L^p-spaces can be nicely described as an L^q-space, where q is the conjugate exponent. The next result is called the *Riesz Representation Theorem*. There is another result also called the "Riesz Representation Theorem" (), so when this name is used readers should be cautioned to note which theorem is being referenced.

XX.15.7.1. Theorem. [Riesz Representation Theorem for L^p] Let $1 \leq p \leq \infty$, and let q be the conjugate exponent to p. Let $(X, \mathcal{A}, \mu)$ be a measure space.

(i) Let $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$. Then g defines a bounded linear functional ϕ_g on $L^p(X, \mathcal{A}, \mu)$ by

$$\phi_g(f) = \int_X fg \, d\mu \ .$$

We have $\|\phi_g\| \leq \|g\|_q$.

(ii) If $p > 1$ or if μ is σ-finite, then $\|\phi_g\| = \|g\|$ for all $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$. If $g, h \in \mathcal{L}^q(X, \mathcal{A}, \mu)$ and $\phi_g = \phi_h$, then $g = h$ a.e.

(iii) If $1 < p < \infty$, or if $p = 1$ and μ is σ-finite, then every bounded linear functional on $L^p(X, \mathcal{A}, \mu)$ is of the form ϕ_g for some $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$.

XX.15.7.2. Corollary. Let $1 \leq p < \infty$. Let $(X, \mathcal{A}, \mu)$ be a measure space, with μ σ-finite if $p = 1$. Then $L^p(X, \mathcal{A}, \mu)^*$ is isometrically isomorphic to $L^q(X, \mathcal{A}, \mu)$ under the correspondence $\phi_g \leftrightarrow [g]$.

XX.15.7.3. We give the proof of Theorem XX.15.7.1.. Part (i) is an almost immediate corollary of Hölder's Inequality XX.15.2.6.: since $fg \in \mathcal{L}^1(X, \mathcal{A}, \mu)$ for any $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$, the linear functional ϕ_g is well defined on $L^p(X, \mathcal{A}, \mu)$, and

$$\|\phi_g\| = \sup\{|\phi_g(f)| : f \in \mathcal{L}^p(X, \mathcal{A}, \mu\}, \|f\|_p = 1\} = \sup\left\{\left|\int_X fg\, d\mu\right| : f \in \mathcal{L}^p(X, \mathcal{A}, \mu), \|f\|_p = 1\right\}$$

$$\leq \sup\left\{\int_X |fg|\, d\mu : f \in \mathcal{L}^p(X, \mathcal{A}, \mu), \|f\|_p = 1\right\} \leq \|g\|_q \ .$$

XX.15.7.4. For (ii), we only need to show $\|\phi_g\| \geq \|g\|_q$, and we may assume $\|g\|_q \neq 0$. If $1 < p < \infty$ set

$$f(x) = \begin{cases} \frac{|g(x)|^q}{g(x)} & \text{if} \quad g(x) \neq 0 \\ 0 & \text{if} \quad g(x) = 0 \end{cases} \ .$$

Then

$$|f|^p = (|g|^{q-1})^p = |g|^q$$

so $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ and $\|f\|_p = \|g\|_q^{q/p}$; and $fg = |g|^q$, so $\phi_g(f) = \|g\|_q^q$; thus

$$\|\phi_g\| \geq \frac{\|g\|_q^q}{\|g\|_q^{q/p}} = \|g\|_q^{q-q/p} = \|g\|_q \ .$$

If $p = \infty$, so $q = 1$, let $g \in \mathcal{L}^1(X, \mathcal{A}, \mu)$. Define the *signum function* σ_g of g by

$$\sigma_g(x) = \begin{cases} \frac{g(x)}{|g(x)|} & \text{if} \quad g(x) \neq 0 \\ 0 & \text{if} \quad g(x) = 0 \end{cases}$$

Then σ_g takes only values of absolute value 1, perhaps along with 0, so $\sigma_g \in \mathcal{L}^\infty(X, \mathcal{A}'\mu)$ and $\|\sigma_g\|_\infty = 1$ (unless $g = 0$ a.e.) We have $\overline{\sigma_g} g = |g|$, so $\phi_g(\overline{\sigma_g}) = \|g\|_1 \leq \|\phi_g\|$.

XX.15.7.5. Part (ii) is trickier to prove for $p = 1$, and we need the hypothesis that μ is σ-finite. Fix $n \in \mathbb{N}$, large enough that $\|g\|_\infty > \frac{1}{n}$, and set

$$A = \left\{x \in X : \|g\|_\infty - \frac{1}{n} \leq |g(x)| \leq \|g\|_\infty\right\} \ .$$

Then $\mu(A) > 0$. Let (A_k) be an increasing sequence of sets in A with $A = \cup_k A_k$ and $0 < \mu(A_k) < \infty$. Set

$$f_k(x) = \begin{cases} \frac{|g(x)|}{g(x)} & \text{if} \quad x \in A_k \\ 0 & \text{if} \quad x \notin A_k \end{cases} \ .$$

Then $|f_k| = \chi_{A_k}$, so $f_k \in \mathcal{L}^1(X, \mathcal{A}, \mu)$ and $\|f_k\|_1 = \mu(A_k)$. We have $f_k g = |g|\chi_{A_k}$, so

$$\phi_g(f_k) = \int_X |g|\chi_{A_k}\, d\mu = \int_{A_k} |g|\, d\mu \geq \left(\|g\|_\infty - \frac{1}{n}\right)\mu(A_k) \ .$$

Thus

$$\|\phi_g\| \geq \frac{\left(\|g\|_\infty - \frac{1}{n}\right)\mu(A_k)}{\mu(A_k)} = \|g\|_\infty - \frac{1}{n} \ .$$

Since $n \in \mathbb{N}$ is arbitrarily large, $\|\phi_g\| \geq \|g\|_\infty$.
For the last statement of (ii), if $\phi_g = \phi_h$, then $0 = \phi_g - \phi_h = \phi_{g-h}$, so $0 = \|\phi_{g-h}\| = \|g - h\|_q$, $g - h = 0$ a.e.

XX.15.7.6. We now turn to (iii). We will prove this in three stages: first for μ finite, then σ-finite, and then for general μ (if $p > 1$). First assume μ is finite, and let ϕ be a bounded linear functional on $L^p(X, \mathcal{A}, \mu)$. For $A \in \mathcal{A}$ define

$$\nu(A) = \phi([\chi_A])$$

(note that $\chi_A \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ since μ is finite). Then ν is a finitely additive complex measure on $(X, \mathcal{A})$. If (A_k) is a sequence of pairwise disjoint sets in $\mathcal{A}$ with union A, set $B_n = \cup_{k=1}^n A_k$; then $(\chi_A - \chi_{B_n})^p$ decreases to 0, so by the MCT

$$\|\chi_A - \chi_{B_n}\|_p^p = \int_X (\chi_A - \chi_{B_n})^p \, d\mu \to 0$$

and thus $[\chi_{B_n}] \to [\chi_A]$ in $L^p(X, \mathcal{A}, \mu)$. Since ϕ is continuous,

$$\nu(B_n) = \sum_{k=1}^n \nu(A_k) = \sum_{k=1}^n \phi(\chi_{A_k}) = \phi\left(\sum_{k=1}^n \chi_{A_k}\right) = \phi(\chi_{B_n}) \to \phi(\chi_A) = \nu(A)$$

so ν is countably additive, i.e. a complex measure on (X, ν). Then

$$|\nu(A)| = |\phi(\chi_A)| \le \|\phi\| \|\chi_A\|_p = \|\phi\| \mu(A)^{1/p}$$

for all $A \in \mathcal{A}$.

We have $\nu \ll \mu$, so by the Radon-Nikodým Theorem there is a $g = \frac{d\nu}{d\mu} \in \mathcal{L}^1(X, \mathcal{A}, \mu)$ such that $\nu(A) = \int_A g \, d\mu$ for all $A \in \mathcal{A}$. Then $g \in \mathcal{L}^1(X, \mathcal{A}, \mu)$. If $s = \sum_{k=1}^n \alpha_k \chi_{A_k}$ is any simple function, then by linearity we have

$$\phi(s) = \sum_{k=1}^n \alpha_k \phi(\chi_{A_k}) = \sum_{k=1}^n \alpha_k \int_{A_k} g \, d\mu = \sum_{k=1}^n \int_{A_k} sg \, d\mu = \int_X sg \, d\mu \ .$$

Since $s \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ and ϕ is bounded, we have

$$\left| \int_X sg \, d\mu \right| = |\phi(s)| \le \|\phi\| \|s\|_p \ .$$

Now let $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$. Let (s_n) be an increasing sequence of simple functions converging pointwise to $|f|$. Then $h_n = (|f| - s_n)^p$ decreases to 0, so by the MCT

$$\int_X h_n \, d\mu = \int_X (|f| - s_n)^p \, d\mu \to 0$$

and thus $s_n \to |f|$ in $L^p(X, \mathcal{A}, \mu)$.

We show $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$. First consider the case $p = 1$, so $q = \infty$. Let $\alpha > \|\phi\|$, and set

$$A = \{x \in X : |g(x)| \ge \alpha\} \ .$$

Set

$$f = \begin{cases} \frac{|g(x)|}{g(x)} & \text{if} \quad x \in A \\ 0 & \text{if} \quad x \notin A \end{cases} \ .$$

Then f is bounded, hence in $\mathcal{L}^p(X, \mathcal{A}, \mu)$; $|f| = \chi_A$, so $\|f\|_p = \mu(A)$. Let (s_n) be a sequence of simple functions converging pointwise to f, with $|s_n| \leq 1$ for all n. Then $s_n g \to fg = |g|\chi_A$, and $|s_n g| \leq |g|$ for all n. Thus by the DCT

$$\int_X |g|\chi_A \, d\mu = \int_X fg \, d\mu = \lim_{n\to\infty} \int_X s_n g \, d\mu = \lim_{n\to\infty} \phi(s_n) = \phi(f)$$

so we have

$$\int_X |g|\chi_A \, d\mu \leq \|\phi\|\|f\|_1 = \|\phi\|\mu(A) \ .$$

On the other hand, $|g| > \alpha$ on A, so

$$\int_X |g|\chi_A \, d\mu = \int_A |g| \, d\mu \geq \alpha\mu(A) \ .$$

Since $\alpha\mu(A) \leq \|\phi\|\mu(A)$ and $\alpha > \|\phi\|$, we conclude $\mu(A) = 0$. This is true for every $\alpha > \|\phi\|$, so $\|g\|_\infty \leq \|\phi\|$, and in particular $g \in \mathcal{L}^\infty(X, \mathcal{A}, \mu)$.

Now suppose $p > 1$. Write $\phi = \phi_{re} + i\phi_{im}$; the corresponding g is $g_{re} + ig_{im}$, where g_{re} and g_{im} are the functions corresponding to ϕ_{re} and ϕ_{im} respectively. To show $\phi \in \mathcal{L}^q(X, \mathcal{A}, \mu)$, it suffices to show that g_{re} and g_{im} are in $\mathcal{L}^q(X, \mathcal{A}, \mu)$, i.e. we may assume g is real-valued. Then σ_g is a simple function. Let (s_n) be an increasing sequence of nonnegative simple functions converging pointwise to $|g|$. Then $r_n = \sigma_g s_n^{q-1}$ is a simple function, $|r_n| = s_n^{q-1} \leq |g|^{q-1}$, $s_n^q \leq r_n g$, and $(r_n g)$ is an increasing sequence of nonnegative functions converging to $|g|^q$. For each n we have

$$\int_X s_n^q \, d\mu \leq \int_X r_n g \, d\mu = \phi(r_n) \leq \|\phi\|\|r_n\|_p$$

$$\int_X s_n^q \, d\mu = \int_X s_n^{p(q-1)} \, d\mu = \int_X (s_n^{q-1})^p \, d\mu = \int_X |r_n|^p \, d\mu = \|r_n\|_p^p$$

$$\int_X s_n^q \, d\mu \leq \|\phi\|\|r_n\|_p = \|\phi\| \left[\int_X s_n^q \, d\mu\right]^{1/p}$$

$$\left[\int_X s_n^q \, d\mu\right]^{1-1/p} = \left[\int_X s_n^q \, d\mu\right]^{1/q} \leq \|\phi\| \ .$$

By the MCT,

$$\left[\int_X |g|^q \, d\mu\right]^{1/q} = \lim_{n\to\infty} \left[\int_X s_n^q \, d\mu\right]^{1/q} \leq \|\phi\|$$

so $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$.

To conclude the proof of (iii) in the finite case, let $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$. Let (s_n) be a sequence of simple functions converging to f, with $|s_n| \leq |f|$ for all n. Then $[s_n] \to [f]$ in $L^p(X, \mathcal{A}, \mu)$ (). Since $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$, $|fg| \in \mathcal{L}^1(X, \mathcal{A}, \mu)$ by Hölder's Inequality, and $|s_n g| \leq |fg|$ for all n. Since $s_n g \to fg$, by the DCT we have

$$\int_X fg \, d\mu = \lim_{n\to\infty} \int_X s_n g \, d\mu = \lim_{n\to\infty} \phi([s_n]) = \phi([f])$$

since ϕ is continuous. Thus $\phi = \phi_g$.

XX.15.7.7. Now assume μ is σ-finite. Let (X_n) be an increasing sequence of sets in $\mathcal{A}$ with $X = \cup_n X_n$ and $\mu(X_n) < \infty$. Let $\mu_n = \mu|_{X_n}$. By the first part of the proof, for each n there is a $g_n \in \mathcal{L}^q(X, \mathcal{A}, \mu)$ with $g_n = 0$ on $X \setminus X_n$ and $\phi(f) = \int_{X_n} f g_n \, d\mu_n = \int_X f g_n \, d\mu$ for any $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ supported on X_n. By uniqueness of g_n, we must have $g_{n+1}|_{X_n} = g_n$ a.e., so modifying the g_n on null sets we may assume $g_{n+1}|_{X_n} = g_n$ for all n. Let g be the function on X which is equal to g_n on X_n for each n. If $p = 1$, we clearly have $\|g\|_\infty = \sup_n \|g_n\|_\infty$, and since $\|g_n\|_\infty \leq \|\phi\|$ for all n, we have $\|g\|_\infty \leq \|\phi\|$, so $g \in \mathcal{L}^\infty(X, \mathcal{A}, \mu)$.

Suppose $p > 1$, so q is finite. We have that $|g_n|$ increases to $|g|$, hence $|g_n|^q$ increases to $|g|^q$, and by the MCT we have

$$\int_X |g|^q \, d\mu = \lim_{n\to\infty} \int_X |g_n|^q \, d\mu \leq \|\phi\|^q$$

so $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$.

If $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$, then $|fg| \in \mathcal{L}^1(X, \mathcal{A}, \mu)$ by Hölder's Inequality. If $f_n = f\chi_{X_n}$, then $f_n g_n = f_n g \to fg$, so by the DCT

$$\int_X fg \, d\mu = \lim_{n\to\infty} \int_X f_n g_n \, d\mu = \lim_{n\to\infty} \phi(f_n) = \phi(f)$$

by continuity of ϕ. Thus $\phi = \phi_g$.

XX.15.7.8. Finally, suppose μ is arbitrary and $p > 1$. For each $A \in \mathcal{A}$ of σ-finite measure, let $\mu_A = \mu|_A$; then there is a $g_A \in \mathcal{L}^q(X, \mathcal{A}, \mu)$, identically 0 on A^c, so that

$$\phi([f]) = \int_A f g_A \, d\mu_A = \int_X f g_A \, d\mu$$

for all $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ supported on A. If $A \subseteq B$, then $g_B|_A = g_A$ a.e. For A σ-finite, set $\rho(A) = \int_X |g_A|^q \, d\mu$. We have $\rho(A) \leq \|\phi\|^q$ for every A. Set

$$M = \sup\{\rho(A) : A \in \mathcal{A} \text{ of } \sigma\text{-finite measure}\} \, .$$

Then $M \leq \|\phi\|^q$. Let (A_n) be a sequence of σ-finite sets in $\mathcal{A}$ with $\rho(A_n) \to M$; we may assume the sets A_n are increasing since a finite or countable union of sets of σ-finite measure has σ-finite measure. Set $A = \cup_n A_n$. Then $A \in \mathcal{A}$ has σ-finite measure. If $B \in \mathcal{A}$ has σ-finite measure and $C = B \setminus A$, then $\int_C |g_C|^q \, d\mu = 0$, so $g_C = 0$ a.e. and $g_B = g_A \chi_B$ a.e. If $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$, then the support B of f is σ-finite, so

$$\phi(f) = \int_B f g_B \, d\mu = \int_X f g_A \, d\mu$$

so $\phi = \phi_{g_A}$.

This completes the proof of Theorem XX.15.7.1.. 

XX.15.7.9. Theorem XX.15.7.1.(ii) holds for $p = 1$ more generally if μ is decomposable, but (iii) can fail (XX.15.9.10.). Part (ii) can fail for $p = 1$ for a bad measure space (XX.15.9.4.). Parts (i) and (ii) hold for $p = \infty$, but (iii) rarely does, even if μ is finite (XX.15.7.12.).

XX.15.7.10. Thus, if $1 < p < \infty$, we have $L^p(X, \mathcal{A}, \mu)^{**} = L^p(X, \mathcal{A}, \mu)$ (more precisely, the natural map from $L^p(X, \mathcal{A}, \mu)$ to $L^p(X, \mathcal{A}, \mu)^{**}$ is an isometric isomorphism) for any measure space $(X, \mathcal{A}, \mu)$, i.e. $L^p(X, \mathcal{A}, \mu)$ is *reflexive* for $1 < p < \infty$. $L^1(X, \mathcal{A}, \mu)$ and $L^\infty(X, \mathcal{A}, \mu)$ are rarely reflexive (only if they are finite-dimensional!)

XX.15.7.11. XX.15.7.1.(i) can be rephrased in a symmetric form: if $(X, .\mathcal{A}, \mu)$ is a measure space, and (p, q) is a conjugate pair (allowing $(1, \infty)$ and $(\infty, 1)$), then there is a pairing $\beta : L^p(X, \mathcal{A}, \mu) \times L^q(X, \mathcal{A}, \mu) \to \mathbb{F}$ given by

$$\beta(f, g) = \int_X fg \, d\mu$$

which is a bilinear form (XX.11.1.1.), and which is bounded (XX.11.12.1.) with $\|\beta\| \leq 1$ by Hölder's Inequality. By XX.15.7.1.(ii) we have $\|\beta\| = 1$ if $1 < p, q < \infty$ or if μ is σ-finite, and in fact we always have $\|\beta\| = 1$ if L^p and L^q are nonzero (cf. XX.15.9.4.).

The Dual of L^∞

Each element of $L^1(X, \mathcal{A}, \mu)$ defines a bounded linear functional on $L^\infty(X, \mathcal{A}, \mu)$, so there is a bounded linear map $T : L^1(X, \mathcal{A}, \mu) \to L^\infty(X, \mathcal{A}, \mu)^*$ given by $T([g]) = \phi_g$. This map is isometric (XX.15.7.1.(ii)), but is rarely surjective (in fact, if and only if $L^\infty(X, \mathcal{A}, \mu)$ is finite-dimensional)..

XX.15.7.12. Examples. (i) Let c be the closed subspace of $\ell^\infty = L^\infty(\mathbb{N}, \mathcal{P}(\mathbb{N}), \gamma)$ consisting of convergent sequences. If $x = (\xi_1, \xi_2, \dots) \in c$, define

$$\psi(x) = \lim_{n \to \infty} \xi_n \ .$$

Then ψ is a bounded linear functional on c, and $\|\psi\| = 1$. By the Hahn-Banach Theorem, ψ extends to a bounded linear functional ϕ on ℓ^∞ with $\|\phi\| = 1$ (such a ϕ is called a *Banach limit* on ℓ^∞). But there cannot be a $y = (\eta_1, \eta_2, \dots) \in \ell^1$ with

$$\phi(x) = \int_{\mathbb{N}} xy \, d\gamma = \sum_{n=1}^{\infty} \xi_n \eta_n$$

for every $x = (\xi_1, \xi_2, \dots) \in \ell^\infty$: if there were such a y, for each n we would have

$$0 = \psi(\boldsymbol{e}_n) = \phi(\boldsymbol{e}_n) = \int_{\mathbb{N}} \boldsymbol{e}_n y \, d\gamma = \eta_n$$

so $y = 0$, a contradiction.

(ii) $L^\infty([0, 1], \mathcal{L}, \lambda)$ is a unital commutative C*-algebra (), thus is of the form $C(X)$ for a (rather bizarre) compact Hausdorff space X (). If $x \in X$, define a linear functional ϕ on $C(X)$ by $\phi(f) = f(x)$. This ϕ does not come from integration with an element of $\mathcal{L}^1([0, 1], \mathcal{L}, \lambda)$. In fact, $L^1([0, 1], \mathcal{L}, \lambda)$ is identified with the set of *normal* bounded linear functionals on $L^\infty([0, 1], \mathcal{L}, \lambda)$ ().

Example (i) can be put in the same context: $\ell^\infty \cong C(\beta\mathbb{N})$ (), so by the other Riesz Representation Theorem () $(\ell^\infty)^*$ can be identified with the set of complex Radon measures on $\beta\mathbb{N}$. The subspace ℓ^1 corresponds to the Radon measures supported on $\mathbb{N}$.

A Converse to Hölder's Inequality

We have the following partial converse to XX.15.2.6., giving a criterion for when a function is p'th power integrable:

XX.15.7.13. THEOREM. Let $(X, \mathcal{A}, \mu)$ be a semifinite measure space, and f a real- or complex-valued $\mathcal{A}$-measurable function. Let $1 \leq p \leq \infty$, and let q be the conjugate exponent. If $fg \in \mathcal{L}^1(X, \mathcal{A}, \mu)$ for all $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$, then $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$.

PROOF: Since both the hypothesis and conclusion are true for f if and only if they are true for $|f|$, it suffices to assume f is nonnegative.

The case $p = 1$ is trivial, with no restriction on μ: take g to be the constant function 1.

Now suppose $1 < p < \infty$. First assume μ is σ-finite. Let (s_n) be an increasing sequence of nonnegative simple functions converging pointwise to f, with the support of each s_n of finite measure. Then $f_n \in \mathcal{L}^p(X, \mathcal{A}, \mu)$, so defines a bounded linear functional ϕ_n on $L^q(X, \mathcal{A}, \mu)$ by

$$\phi_n([g]) = \int_X f_n g \, d\mu \ .$$

Fix $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$. Then, for every n,

$$|\phi_n([g])| = \left| \int_X f_n g \, d\mu \right| \leq \int_X |f_n g| \, d\mu \leq \int_X |fg| \, d\mu$$

we have that the ϕ_n are pointwise bounded. Thus by the Uniform Boundedness Theorem (), they are uniformly bounded, i.e. there is an M such that $|\phi_n([g])| \leq M\|g\|_q$ for all n and all $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$. Define a linear functional ϕ on $L^q(X, \mathcal{A}, \mu)$ by

$$\phi([g]) = \int_X fg \, d\mu \ .$$

If $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$, then $f_n g \to fg$, and $|f_n g| \leq |fg|$, so by the DCT

$$|\phi(g)| = \left| \int_X fg \, d\mu \right| = \lim_{n\to\infty} \left| \int_X f_n g \, d\mu \right| = \lim_{n\to\infty} |\phi_n(g)| \leq M\|g\|_q$$

so ϕ is bounded and $\|\phi\| \leq M$. Thus there is an $h \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ with $\phi = \phi_h$. We show that $f = h$ a.e. If $A \in \mathcal{A}$ with $\mu(A) < \infty$, set $\gamma_A = \sigma(f - h)\chi_A$. Then $\gamma_A \in \mathcal{L}^q(X, \mathcal{A}, \mu)$ and $\overline{\gamma_A}(f - h) = |f - h|\chi_A$. We have

$$0 = \int_X f\gamma_A \, d\mu - \int_X h\gamma_A \, d\mu = \int_X (f - h)\gamma_A \, d\mu = \int_A |f - h| \, d\mu$$

so $f - h = 0$ a.e. on A. Since A is arbitrary and μ is σ-finite, $f = h$ a.e. on X.

Now suppose μ is just semifinite. If the support of f is not σ-finite, then there is an m such that $B = \left\{x : f(x) \geq \frac{1}{m}\right\}$ is not σ-finite, hence has infinite measure. Since μ is semifinite, B contains measurable sets of arbitrarily large finite measure (XVII.1.4.14.); hence there is a $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$, supported on B, with $g \notin \mathcal{L}^1(X, \mathcal{A}, \mu)$ (XX.15.5.1.). Since $fg \geq \frac{1}{m}g$, fg is not integrable, a contradiction.

Thus $E = \{x : f(x) > 0\}$ has σ-finite measure. Reducing to the measure space $(E, \mathcal{A}_E, \mu|_E)$, which is σ-finite, we conclude from the first part of the proof that $f|_E \in \mathcal{L}^p(E, \mathcal{A}_E, \mu|_E)$, so $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$.

Now let $p = \infty$. Suppose f is not essentially bounded. If $f = \infty$ on a set of positive measure, then there is an $A \in \mathcal{A}$ with $0 < \mu(A) < \infty$ and $f = \infty$ on A. Then $\chi_A \in \mathcal{L}^1(X, \mathcal{A}, \mu)$, but $f\chi_A$ is not integrable, a contradiction. If f is finite a.e., then there is a sequence (A_n) of pairwise disjoint sets in $\mathcal{A}$ with $0 < \mu(A_n) < \infty$ and $|f| \geq 2^n$ on A_n. Set $g_n = \frac{1}{2^n \mu(A_n)} \chi_{A_n}$. Then

$$\sum_{n=1}^{\infty} g_n$$

converges pointwise and in $\|\cdot\|_1$ to a (nonnegative) function $g \in \mathcal{L}^1(X, \mathcal{A}, \mu)$. But

$$\int_{A_n} fg \, d\mu \geq 1$$

for all n, so fg cannot be integrable, again a contradiction. 👍

XX.15.7.14. As noted, the semifinite hypothesis is unnecessary if $p = 1$. If $1 < p < \infty$ and μ is not semifinite, the result fails: if $A \in \mathcal{A}$ with $\mu(A) = \infty$ contains no subsets of finite positive measure, then any integrable function equals 0 a.e. on A. Thus $\chi_A g = 0$ a.e., hence is integrable, for any $g \in \mathcal{L}^q(X, \mathcal{A}, \mu)$. But $\chi_A \notin \mathcal{L}^p(X, \mathcal{A}, \mu)$. The case $p = \infty$ can also fail if μ is not semifinite: let X be an infinite set, and take $\mu(A) = \infty$ for all nonempty $A \subseteq X$.

XX.15.8. Conditional Expectations on L^p-Spaces

We continue the discussion of conditional expectations from XVIII.6.5.16.. Let $(X, \mathcal{A}, \mu)$ be a probability measure space, and let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{A}$. Fix a field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. We will suppress the subscript $\mathbb{F}$ on $L^1_{\mathbb{F}}(X, \mathcal{A}, \mu)$, etc., but keep in mind that $\mathbb{F}$ has been specified.

XX.15.8.1. First note that $L^1(X, \mathcal{A}, \mu)$ is a Banach space under $\|\cdot\|_1$, and that $L^1(X, \mathcal{B}, \mu|_{\mathcal{B}})$ is a closed subspace. The conditional expectation map $\mathbf{E}$ is a linear transformation from $L^1(X, \mathcal{A}, \mu)$ onto the closed subspace $L^1(X, \mathcal{B}, \mu|_{\mathcal{B}})$; call this map $\mathbf{E}_1$. $L^\infty(X, \mathcal{A}, \mu)$ is a subspace of $L^1(X, \mathcal{A}, \mu)$ which is dense for the norm $\|\cdot\|_1$ (all simple functions are in L^∞), and is itself a Banach space under the different norm $\|\cdot\|_\infty$. Similarly, $L^\infty(X, \mathcal{B}, \mu|_{\mathcal{B}})$ is a $\|\cdot\|_1$-dense subspace of $L^1(X, \mathcal{B}, \mu|_{\mathcal{B}})$ which is an $\|\cdot\|_\infty$-closed subspace of $L^\infty(X, \mathcal{A}, \mu)$. $\mathbf{E}$ maps $L^\infty(X, \mathcal{A}, \mu)$ onto the closed subspace $L^\infty(X, \mathcal{B}, \mu|_{\mathcal{B}})$; call this map $\mathbf{E}_\infty$.

XX.15.8.2. The first observation is that $\mathbf{E}_\infty$ is bounded and has norm 1 (the argument that $\|\mathbf{E}_\infty\| \leq 1$ is described in XVIII.6.5.16., and equality holds since $\mathbf{E}_\infty(1) = 1$). We similarly have:

XX.15.8.3. Proposition. The map $\mathbf{E}_1 : L^1(X, \mathcal{A}, \mu) \to L^1(X, \mathcal{B}, \mu|_{\mathcal{B}})$ is bounded and has norm 1.

Proof: We use the fact that $L^\infty(X, \mathcal{B}, \mu|_{\mathcal{B}})$ is the dual of $L^1(X, \mathcal{B}, \mu|_{\mathcal{B}})$. By the Hahn-Banach Theorem, if $f \in \mathcal{L}^1(X, \mathcal{A}, \mu)$, then

$$\|\mathbf{E}_1([f])\|_1 = \sup \left\{ \left| \int_X g\mathbf{E}(f|\mathcal{B}) \, d\mu \right| : g \in \mathcal{L}^\infty(X, \mathcal{B}, \mu|_{\mathcal{B}}), \|g\|_\infty = 1 \right\} .$$

But for each such g, by XVIII.6.5.9. we have

$$\left|\int_X g\mathbf{E}(f|\mathcal{B})\,d\mu\right| = \left|\int_X fg\,d\mu\right| \leq \int_X |fg|\,d\mu \leq \int_X |f|\,d\mu = \|[f]\|_1 \ .$$

Thus $\mathbf{E}_1$ is bounded and $\|\mathbf{E}_1\| \leq 1$. Since $\mathbf{E}_1(1) = 1$, equality holds. 👍

XX.15.8.4. We can extend this argument. If $1 < p < \infty$, we have

$$L^\infty(X, \mathcal{A}, \mu) \subseteq L^p(X, \mathcal{A}, \mu) \subseteq L^1(X, \mathcal{A}, \mu)$$

since μ is finite, and similarly for $L^p(X, \mathcal{B}, \mu|_\mathcal{B})$, which is a $\|\cdot\|_p$-closed subspace of $L^p(X, \mathcal{A}, \mu)$. We have:

XX.15.8.5. THEOREM. Let $(X, \mathcal{A}, \mu)$ be a probability measure space, and $\mathcal{B}$ a sub-σ-algebra of $\mathcal{A}$. Let $\mathbf{E} : L^1(X, \mathcal{A}, \mu) \to L^1(X, \mathcal{B}, \mu|_\mathcal{B})$ be the conditional expectation map. Let $1 \leq p \leq \infty$. Then
(i) If $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$, then $\mathbf{E}(f|\mathcal{B}) \in \mathcal{L}^p(X, \mathcal{B}, \mu|_\mathcal{B})$.
(ii) If $\mathbf{E}_p : L^p(X, \mathcal{A}, \mu) \to L^p(X, \mathcal{B}, \mu|_\mathcal{B})$ is the induced map, then $\mathbf{E}_p$ is bounded and $\|\mathbf{E}_p\| = 1$.

PROOF: The result has already been proved for $p = 1$ and $p = \infty$. Suppose $1 < p < \infty$, and let q be the conjugate exponent. Let $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$. For each n, set

$$f_n(x) = \begin{cases} f(x) & \text{if} \quad |f(x)| \leq n \\ 0 & \text{if} \quad |f(x)| > n \end{cases}$$

Then $|f_n| \in \mathcal{L}^\infty(X, \mathcal{A}, \mu)$, so $E(|f_n||\mathcal{B}) \in \mathcal{L}^\infty(X, \mathcal{B}, \mu|_\mathcal{B}) \subseteq \mathcal{L}^p(X, \mathcal{B}, \mu|_\mathcal{B})$. If $g \in \mathcal{L}^q(X, \mathcal{B}, \mu|_\mathcal{B})$, then by XX.15.2.6.

$$\left|\int_X g\mathbf{E}(|f_n||\mathcal{B})\,d\mu\right| \leq \int_X |g\mathbf{E}(|f_n||\mathcal{B})|\,d\mu = \int_X |\mathbf{E}(|f_n|g|\mathcal{B})|\,d\mu \leq \int_X \mathbf{E}(|f_n g||\mathcal{B})\,d\mu$$

$$= \int_X |f_n g|\,d\mu \leq \|f_n\|_p \|g\|_q \leq \|f\|_p \|g\|_q$$

for each n. Since $L^q(X, \mathcal{B}, \mu|_\mathcal{B})$ is the dual of $L^p(X, \mathcal{B}, \mu|_\mathcal{B})$, by the Hahn-Banach Theorem we have, for each n,

$$\|\mathbf{E}(|f_n||\mathcal{B})\|_p = \sup\left\{\left|\int_X g\mathbf{E}(|f_n||\mathcal{B})\,d\mu\right| : g \in \mathcal{L}^q(X, \mathcal{B}, \mu|_\mathcal{B}), \|g\|_q = 1\right\} \leq \|f\|_p \ .$$

Alternately, to avoid using the Hahn-Banach Theorem, set $g_n(x) = \mathbf{E}(|f_n||\mathcal{B})(x)^{p-1}$. Then $\mathbf{E}(|f_n||\mathcal{B})g_n = \mathbf{E}(|f_n||\mathcal{B})^p = |g_n|^q$ so $g_n \in \mathcal{L}^q(X, \mathcal{B}, |_\mathcal{B})$ and

$$\|g_n\|_q = \|\mathbf{E}(|f_n||\mathcal{B})\|_p^{p/q} = \|\mathbf{E}(|f_n||\mathcal{B})\|_p^{p-1}$$

and therefore

$$\|\mathbf{E}(|f_n||\mathcal{B})\|_p^p = \int_X \mathbf{E}(|f_n|\mathcal{B})^p\,d\mu = \int_X g_n\mathbf{E}(|f_n|\mathcal{B})\,d\mu \leq \|f\|_p\|g_n\|_q = \|f\|_p\|\mathbf{E}(|f_n||\mathcal{B})\|_p^{p-1}$$

and we again conclude that $\|\mathbf{E}(|f_n||\mathcal{B})\|_p \le \|f\|_p$ for each n.
Since $|f_n|$ increases to $|f|$ a.e., $\mathbf{E}(|f_n||\mathcal{B})$ increases to $\mathbf{E}(|f||\mathcal{B})$ a.e., so by the MCT,

$$\int_X |\mathbf{E}(f|\mathcal{B})|^p \, d\mu \le \int_X \mathbf{E}(|f||\mathcal{B})^p \, d\mu = \lim_{n\to\infty} \int_X \mathbf{E}(|f_n||\mathcal{B})^p \, d\mu \le \|f\|_p^p$$

and thus $\mathbf{E}(f|\mathcal{B}) \in \mathcal{L}^p(X, \mathcal{B}, \mu|_{\mathcal{B}})$ and $\|\mathbf{E}(f|\mathcal{B})\|_p \le \|f\|_p$. So $\mathbf{E}_p$ is defined and bounded, and $\|\mathbf{E}_p\| \le 1$. Equality holds since $\mathbf{E}_p(1) = 1$. 👍

XX.15.8.6. Thus $L^p(X, \mathcal{B}, \mu|_{\mathcal{B}})$ is a very special subspace of $L^p(X, \mathcal{A}, \mu)$: there is a projection of norm 1 onto the subspace, which is unusual for general closed subspaces of Banach spaces (). If $p = 2$, $L^2(X, \mathcal{A}, \mu)$ is a Hilbert space, so $\mathbf{E}_2$ is necessarily the orthogonal projection onto the subspace $L^2(X, \mathcal{B}, \mu|_{\mathcal{B}})$ (this can be verified directly).

XX.15.9. Exercises

XX.15.9.1. Let $0 < r < 1$.
(a) Let $f(x) = x^r - rx$ for $x \ge 0$. Show that f is increasing on $[0, 1]$ and decreasing on $[1, \infty)$, and has a unique maximum $1 - r$ at $x = 1$.
(b) If $x \in [0, \infty)$, show that $x^r \le rx + (1 - r)$, with equality if and only if $x = 1$.
(c) If $a, b > 0$, show that $a^r b^{1-r} \le ra + (1 - r)b$, with equality if and only if $a = b$. [Apply (b) with $x = \frac{a}{b}$.]
(d) Show that the inequality in (c) is equivalent to Young's Inequality for products. [Let $r = \frac{1}{p}$, $\alpha = a^r$, $\beta = b^{1-r}$.]

XX.15.9.2. Here is the general version of Young's Inequality for products:
THEOREM. [YOUNG'S INEQUALITY FOR PRODUCTS, GENERAL VERSION] Let ϕ be a strictly increasing continuous function from $[0, \infty)$ onto $[0, \infty)$ (that is, $\phi(0) = 0$ and $\lim_{x\to+\infty} \phi(x) = +\infty$), and let ψ be the inverse function of ϕ. For $x \ge 0$, set

$$\Phi(x) = \int_0^x \phi(t) \, dt$$

$$\Psi(x) = \int_0^x \psi(t) \, dt \ .$$

If $\alpha, \beta \in [0, \infty)$, then

$$\alpha\beta \le \Phi(\alpha) + \Psi(\beta)$$

with equality if and only if $\beta = \phi(\alpha)$.

See Figure XX.1 for a geometric interpretation in terms of areas. The red area is $\Phi(\alpha)$ and the blue area is $\Psi(\beta)$, and $\alpha\beta$ is the area of the rectangle bounded by the dotted lines and the axes.

(a) Use () to show that, for $\beta \ge 0$,

$$\Psi(\beta) = \int_0^\beta \int_0^{\psi(y)} dx \, dy = \int_0^{\psi(\beta)} \int_{\phi(x)}^\beta dy \, dx \ .$$

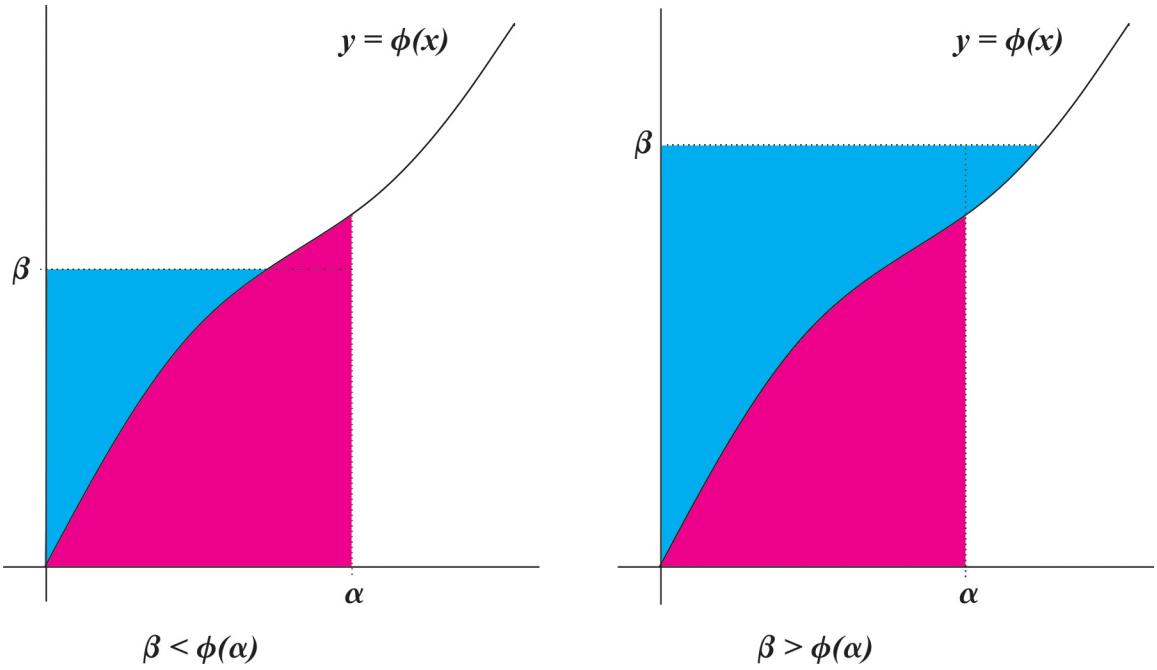


Figure XX.1: Geometric Interpretation of Young's Inequality for Products

(b) If $\beta \geq \phi(\alpha)$, then

$$\alpha\beta = \int_0^\alpha \int_0^\beta dy\,dx = \int_0^\alpha \int_0^{\phi(x)} dy\,dx + \int_0^\alpha \int_{\phi(x)}^\beta dy\,dx = \Phi(\alpha) + \int_0^\alpha \int_{\phi(x)}^\beta dy\,dx\ .$$

(c) Use (a) and (b) to prove the Theorem if $\beta \geq \phi(\alpha)$. Give a similar proof if $\beta \leq \phi(\alpha)$.
(d) Show that the Young's Inequality for products of XX.15.2.5. is the special case $\phi(x) = x^{p-1}$.

XX.15.9.3. Prove the following generalized versions of Hölder's Inequality:

THEOREM. [HÖLDER'S INEQUALITY, GENERAL VERSION] Let $(X, \mathcal{A}, \mu)$ be a measure space.
(i) Let $p, q > 1$, $r \geq 1$, with

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{r}$$

and let $f \in \mathcal{L}^p(\mu)$ and $g \in \mathcal{L}^q(\mu)$. Then $fg \in \mathcal{L}^r(\mu)$ and

$$\|fg\|_r \leq \|f\|_p \|g\|_q\ .$$

(ii) Let $p_1, \ldots, p_n > 1$, $r \geq 1$, with

$$\frac{1}{p_1} + \cdots + \frac{1}{p_n} = \frac{1}{r}$$

and let $f_k \in \mathcal{L}^{p_k}(\mu)$ for $1 \le k \le n$. Then $f_1 f_2 \cdots f_n \in \mathcal{L}^r(\mu)$ and

$$\|f_1 f_2 \cdots f_n\|_r \le \|f_1\|_{p_1} \|f_2\|_{p_2} \cdots \|f_n\|_{p_n} \ .$$

[To prove (i), note that

$$\frac{1}{p/r} + \frac{1}{q/r} = 1$$

and apply Hölder's inequality to $|f|^r$ and $|g|^r$. Prove (ii) by induction on n. For the inductive step set

$$\frac{1}{p} = \frac{1}{p_1} + \cdots + \frac{1}{p_n}$$

and $q = p_{n+1}$ to obtain $f_1 f_2 \cdots f_n \in \mathcal{L}^p$ and $\|f_1 \cdots f_{n+1}\|_r \le \|f_1 \cdots f_n\|_p \|f_{n+1}\|_{p_{n+1}}$.]

XX.15.9.4. Let $X = \{x\}$ be a one-point space, and let $\mu(\{x\}) = \infty$. Show that $\mathcal{L}^p(X, \mathcal{P}(X), \mu) = \{0\}$ for $1 \le p < \infty$, and that $\mathcal{L}^\infty(X, \mathcal{P}(X), \mu)$ is one-dimensional (the set of all scalar-valued functions). In particular, $L^\infty(X, \mathcal{P}(X), \mu)$ is larger than $L^1(X, \mathcal{P}(X), \mu)^*$.

XX.15.9.5. Let X be an uncountable set, and for $A \subseteq X$ let $\mu(A) = 0$ if A is countable and $\mu(A) = \infty$ if A is uncountable.

(a) If $1 \le p < \infty$, show that $\mathcal{L}^p(X, \mathcal{P}(X), \mu)$ is the set of functions vanishing except on a countable set, and $L^p(X, \mathcal{P}(X), \mu) = \{0\}$.

(b) Show that $\mathcal{L}^\infty(X, \mathcal{P}(X), \mu)$ is the set of all functions which are bounded on a cocountable set, and $L^\infty(X, \mathcal{P}(X), \mu)$ is infinite-dimensional and even nonseparable.

XX.15.9.6. (a) Even in a finite continuous measure space the L^p-spaces can be small simply because of the lack of sufficiently many measurable functions. Let X be an uncountable set, $\mathcal{A}$ the σ-algebra of countable-cocountable sets, and μ the measure on $(X, \mathcal{A})$ with $\mu(A) = 0$ if A is countable and $\mu(A) = 1$ if A is cocountable. A scalar-valued measurable function on A must be constant except on a countable set (). Show that $L^p(X, \mathcal{A}, \mu)$ is one-dimensional for all p, $1 \le p \le \infty$ (every measurable function is equal a.e. to a constant function).

(b) Show that the result of (a) holds true in any measure space in which every measurable set has measure either 0 or 1. Thus, if X is a set of measurable cardinality () and μ is a full $\{0,1\}$-valued measure on $(X, \mathcal{P}(X))$, then $L^p(X, \mathcal{P}(X), \mu)$ is one-dimensional for $1 \le p \le \infty$ even though there are plenty of measurable functions; every scalar-valued function is still equal a.e. to a constant function.

XX.15.9.7. (a) Show that, in the situation of Example XX.15.5.1., for any p there is an $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ which is not in $\mathcal{L}^r(X, \mathcal{A}, \mu)$ for any $r < p$. [For each n, construct a function f_n as in Example XX.15.5.1. which is in $\mathcal{L}^p(\mu)$ but not in $\mathcal{L}^{p-1/n}(\mu)$, and consider $\sum_{n=1}^\infty \frac{2^{-n}}{\|f_n\|_p} f_n$. Modify the argument slightly for $p = \infty$.]

(b) In the situation of XX.15.5.3., use a similar argument to show that if $0 < p < \infty$, there is an $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ which is not in $\mathcal{L}^r(X, \mathcal{A}, \mu)$ for any $r > p$.

(c) If $(X, \mathcal{A}, \mu)$ contains sets of both arbitrarily large and arbitrarily small finite positive measure, and $0 < s \le r \le \infty$, show that

(i) There is an f such that $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ if and only if $s \le p \le r$ (in particular, if $s = r$, there is an f such that $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ if and only if $p = r$).

(ii) If $s < r$, there is an f such that $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ if and only if $s < p \le r$.

(iii) If $s < r$, there is an f such that $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ if and only if $s \le p < r$.

(iv) If $s < r$, there is an f such that $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ if and only if $s < p < r$.

XX.15.9.8. Let $(X, \mathcal{A}, \mu)$ be a measure space, and $0 < s < p < r \le \infty$.

(a) Using the arguments of the proofs of XX.15.5.2. and XX.15.5.4., show that every $f \in \mathcal{L}^p(X, \mathcal{A}, \mu)$ can be written as $f = g + h$ with $g \in \mathcal{L}^s(X, \mathcal{A}, \mu)$, $h \in \mathcal{L}^r(X, \mathcal{A}, \mu)$, $gh = 0$, and $J_p(f) \le J_s(g) + J_r(h)$.

(b) Show that the function

$$\|f\| = \inf\{\|g\|_s + \|h\|_r : g \in \mathcal{L}^s(\mu), h \in \mathcal{L}^r(\mu), f = g + h\}$$

is a seminorm on the subspace $\mathcal{L}^s(\mu) + \mathcal{L}^r(\mu)$ of the space of $\mathcal{A}$-measurable functions, that $\|f\| = 0$ if and only if $f \in Z$, and that the quotient, which we denote $L^s(\mu) + L^r(\mu)$, is complete under the corresponding norm. Show that the inclusion map from $L^p(\mu)$ into $L^s(\mu) + L^r(\mu)$ is bounded.

(c) Show that $\mathcal{L}^s(\mu) \cap \mathcal{L}^r(\mu) \subseteq \mathcal{L}^p(\mu)$, that $\|f\| = \|f\|_s + \|f\|_r$ is a seminorm on $\mathcal{L}^s(\mu) \cap \mathcal{L}^r(\mu)$, $\|f\| = 0$ if and only if $f \in Z$, and that the quotient, which we denote $L^s(\mu) \cap L^r(\mu)$, is complete under the corresponding norm. Show that the inclusion map from $L^s(\mu) \cap L^r(\mu)$ into $L^p(\mu)$ is bounded.

XX.15.9.9. (a) Prove directly that if $p > 1$ and q is conjugate to p, the map $a \mapsto \phi_a$ is an isometric isomorphism from ℓ^q to $(\ell^p)^*$, where

$$\phi_a(x) = \sum_{k=1}^{\infty} \alpha_k \xi_k$$

for $a = (\alpha_1, \alpha_2, \dots)$ and $x = (\xi_1, \xi_2, \dots)$. [Use Hölder's inequality and the fact that $\{\boldsymbol{e}_k\}$ is a Schauder basis for ℓ^p ().]

(b) Give a similar proof that $(c_0)^* \cong \ell^1$ and that $(\ell^1)^* \cong \ell^\infty$.

This gives an alternate proof that ℓ^p is complete for $1 \le p \le \infty$.

XX.15.9.10. Show that if $(X, \mathcal{A}, \mu)$ is a decomposable measure space, show that $L^1(X, \mathcal{A}, \mu)^*$ is naturally isometrically isomorphic to $L^\infty(X, \underline{\mathcal{A}}, \underline{\mu})$, where $(X, \underline{\mathcal{A}}, \underline{\mu})$ is the saturation () of $(X, \mathcal{A}, \mu)$. Thus XX.15.7.1.(iii) holds for $p = 1$ for a decomposable measure space if and only if it is saturated. Show that XX.15.7.1.(ii) for $p = 1$ always holds for a decomposable measure space.

XX.15.9.11. [LL97] (a) Use Tonelli's Theorem to prove the following version of Minkowski's Inequality:

THEOREM. [MINKOWSKI'S INEQUALITY, GENERAL VERSION] Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$ be σ-finite measure spaces, and $1 \le p < \infty$. If f is a nonnegative $(\mu \times \nu)$-measurable function on $X \times Y$, then

$$\int_Y \left[\int_X f(x,y)^p \, d\mu(x)\right]^{1/p} d\nu(y) \ge \left(\int_X \left[\int_Y f(x,y) \, d\nu(y)\right]^p d\mu(x)\right)^{1/p}.$$

If $1 < p < \infty$, equality and finiteness imply there are nonnegative measurable functions g on X and h on Y such that $f(x,y) = g(x)h(y)$ a.e.

[Set $H(x) = \int_Y f(x,y)\,d\nu(y)$ and bound $\int_X H^p\,d\mu = \int_X H(x)\cdot H(x)^{p-1}\,d\mu(x)$ using Hölder's Inequality.]

(b) Use (a) to give an alternate proof of the usual Minkowski Inequality XX.15.2.7. in the σ-finite case. [Reduce to the case where f and g are nonnegative and take Y to be a two-point space with counting measure.]

(c) Give a counterexample to the last statement in the theorem if $p = 1$.

XX.15.9.12. If $n \in \mathbb{N}$, write $n = 2^m + k$ for some $m \geq 0$ and $0 \leq k < 2^m$. Let f_n be the characteristic (indicator) function of the interval $\left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]$.

(a) Show that $f_n \to 0$ in $\mathcal{L}^p([0,1], \mathcal{L}, \lambda)$ for $1 \leq p < \infty$.

(b) Show that $f_n \to 0$ in measure.

(c) Show there is no $x \in [0,1]$ such that $f_n(x) \to 0$ (in fact, $(f_n(x))$ does not converge for any $x \in [0,1]$).

XX.15.9.13. Let $1 < p < \infty$, and q its conjugate exponent, and let $(X, \mathcal{A}, \mu)$ be a measure space.

(a) Prove the *Clarkson inequalities* [Cla36]:
If $p \leq 2$ and $x, y \in L^p(X, \mathcal{A}, \mu)$, then

$$\left\|\frac{x+y}{2}\right\|_p^p + \left\|\frac{x-y}{2}\right\|_p^p \geq \frac{1}{2}\|x\|_p^p + \frac{1}{2}\|y\|_p^p$$

$$\left\|\frac{x+y}{2}\right\|_p^q + \left\|\frac{x-y}{2}\right\|_p^q \leq \left[\frac{1}{2}\|x\|_p^p + \frac{1}{2}\|y\|_p^p\right]^{q-1}$$

and the reversed inequalities hold if $p \geq 2$. [Note that if $p = 2$ the inequalities reduce to the Parallelogram Law (XX.11.2.6.).]

(b) Use the Clarkson inequalities to prove that $L^p(X, \mathcal{A}, \mu)$ is uniformly convex (XX.11.11.15.) for $1 < p < \infty$.

XX.16. C*-Algebras

C*-algebras are the nicest kinds of Banach algebras, and have a beautiful theory with profound applications. We will touch on only some of the basics of this extensive theory here; see [Bla06] for a much more comprehensive treatment.

XX.16.1. Banach *-Algebras and C*-Algebras

XX.16.1.1. DEFINITION. An *involution* on a Banach algebra A is a conjugate-linear isometric antiautomorphism of order two, usually denoted $x \mapsto x^*$. In other words, $(x+y)^* = x^* + y^*$, $(xy)^* = y^*x^*$, $(\lambda x)^* = \bar{\lambda}x^*$, $(x^*)^* = x$, $\|x^*\| = \|x\|$ for all $x, y \in A$, $\lambda \in \mathbb{C}$. A *Banach *-algebra* is a Banach algebra with an involution.

An *(abstract) C*-algebra* is a Banach *-algebra A satisfying the *C*-axiom*:

$$\|x^*x\| = \|x\|^2 \text{ for all } x \in A$$

The deceptively simple and innocuous C*-axiom turns out to be extremely powerful, forcing rigid structure on a C*-algebra. For example, it follows that the norm is completely determined by the algebraic structure and is thus unique (XX.16.2.34.), that *-homomorphisms of C*-algebras are automatically contractive (XX.16.1.11.), and that every C*-algebra can be isometrically represented as a concrete C*-algebra of operators (XX.16.4.35.). One obvious, but useful, consequence is that in a C*-algebra, $x^*x = 0 \Rightarrow x = 0$. [In fact, in a C*-algebra, $\sum x_j^* x_j = 0$ implies that each $x_j = 0$ (XX.16.3.2.(i), XX.16.3.3.).]

XX.16.1.2. In many older references, abstract C*-algebras were called *B*-algebras*, with the name "C*-algebra" reserved for concrete C*-algebras. The term "C*-algebra," first introduced in [Seg47] (for concrete C*-algebras, but viewed in a somewhat abstract manner), did not become universal until well after the publication of [Dix69] in 1964 (there were occasional references to "B*-algebras" in the literature at least as late as 1980; see even [Lax02]). According to [DB86, p. 6], the "C" in "C*-algebra" originally meant "closed", and not, as commonly believed, "continuous", although the interpretation as standing for "continuous" is nicely in line with the modern point of view of C*-algebra theory as "noncommutative topology."

The issue of terminology is clouded by the fact that several different (but ultimately equivalent) axiom schemes have been used for C*-algebras over the years. For example, it is easily seen that the C*-axiom, along with submultiplicativity of the norm, implies that the involution is isometric, so it is unnecessary to include isometry of the involution as an axiom. The C*-axiom has sometimes been replaced by the apparently weaker axiom that $\|x^*x\| = \|x^*\|\|x\|$ for all x. It turns out that this weaker axiom, along with the Banach algebra axioms, also implies isometry of the involution (a much harder result), so the weakened axiom is equivalent to the C*-axiom. See also XX.16.3.4.. See [DB86] for details about the C*-algebra axioms.

XX.16.1.3. EXAMPLES.

(i) $\mathcal{B}(\mathcal{H})$ with the usual algebraic operations, involution, and operator norm, is a C*-algebra for any Hilbert space $\mathcal{H}$ (). If $\mathcal{H}$ is n-dimensional, we obtain that the $n \times n$ matrices $\mathbb{M}_n = \mathcal{B}(\mathbb{C}^n)$ form a C*-algebra with the usual involution (conjugate transpose) and operator norm. More generally, any norm-closed *-subalgebra of $\mathcal{B}(\mathcal{H})$ is a C*-algebra; such C*-algebras are called *concrete C*-algebras*. For example, the algebra $\mathcal{K}(\mathcal{H})$ of compact operators () on a Hilbert space $\mathcal{H}$ is a concrete C*-algebra. If $\mathcal{H}$ is infinite-dimensional, $\mathcal{K}(\mathcal{H})$ is nonunital; thus a C*-algebra need not be unital. We denote by $\mathbb{K}$ the C*-algebra of compact operators on a separable, infinite-dimensional Hilbert space.

(ii) Let X be a locally compact Hausdorff space, and $C_0(X)$ the complex-valued continuous functions on X vanishing at infinity. Give $C_0(X)$ its usual pointwise operations and supremum norm. Define an involution by $f^*(x) = \overline{f(x)}$. Then $C_0(X)$ is a commutative C*-algebra. In fact, every commutative C*-algebra is of this form (XX.16.2.5.). $C_0(X)$ has a unit (identity) if and only if X is compact; in this case, we usually write $C(X)$. More generally, if B is a C*-algebra, then the set $C_0(X, B)$ of (norm-)continuous functions from X to B vanishing at infinity, with pointwise operations and supremum norm, is a C*-algebra. In particular, $C_0(X, \mathbb{M}_n) \cong M_n(C_0(X))$ is a C*-algebra. (In fact, a matrix algebra over any C*-algebra is a C*-algebra (**??**); this is a special case of the tensor product for C*-algebras (**??**)).

(iii) Let G be a locally compact topological group with (left) Haar measure μ. Then $L^1(G, \mu)$ becomes a Banach *-algebra under convolution. It is not a C*-algebra unless G is trivial. This example will be treated in more detail in **??**.

XX.16.1.4. If A is a Banach algebra and J is a closed ideal ("ideal" will always mean "two-sided ideal" unless otherwise specified) in A, then the quotient norm makes A/J into a Banach algebra. If A is a Banach *-algebra and J is a *-ideal (i.e. closed under *), then A/J is a Banach *-algebra. It turns out that if A is a C*-algebra and J is a closed ideal in A, then J is automatically a *-ideal and A/J is a C*-algebra in the quotient norm (XX.16.3.14.).

XX.16.1.5. A Banach algebra is separable if and only if it is countably generated as a Banach algebra (a countably generated algebra over $\mathbb{Q}(i)$ is countable). A C*-algebra is separable if and only if it is countably generated as a C*-algebra.

XX.16.1..1 Unitization

XX.16.1.1. A Banach algebra, even a C*-algebra, need not be unital (e.g. $\mathcal{K}(\mathcal{H})$, $C_0(X)$ for X non-compact). However, every nonunital Banach algebra A can be embedded in a unital Banach algebra $A^\dagger$ (XX.12.1.6.). If A is a Banach *-algebra, for $(a, \lambda) \in A^\dagger$ define $(a, \lambda)^* = (a^*, \bar{\lambda})$; this is an involution making $A^\dagger$ a (unital) Banach *-algebra. With the norm of XX.12.1.6. $A^\dagger$ is not a C*-algebra, but:

XX.16.1.2. Proposition. If A is a C*-algebra, the operator norm on $A^\dagger$ as left multiplication operators on A, i.e.

$$\|(a, \lambda)\| = \sup\{\|ab + \lambda b\| : \|b\| = 1\}$$

is a C*-norm on $A^\dagger$ extending the norm of A.

Proof: $A^\dagger$ is a subalgebra of $\mathcal{B}(A)$ with the operator norm; thus it is a normed algebra, and in particular the norm is submultiplicative. If $(a, \lambda) \in A^\dagger$ and $b \in A$, $\|b\| = 1$, then $(a, \lambda)^*(a, \lambda) = (a^*a + \lambda a^* + \bar{\lambda}a, \lambda\bar{\lambda})$, so

$$\|ab + \lambda b\|^2 = \|(ab + \lambda b)^*(ab + \lambda b)\| = \|(b^*a^* + \bar{\lambda}b^*)(ab + \lambda b)\|$$
$$= \|b^*(a^*ab + \lambda a^*b + \bar{\lambda}ab + \lambda\bar{\lambda}b)\| \leq \|a^*ab + \lambda a^*b + \bar{\lambda}ab + \lambda\bar{\lambda}b\| \leq \|(a, \lambda)^*(a, \lambda)\|$$

and taking the supremum over all b we obtain $\|(a, \lambda)\|^2 \leq \|(a, \lambda)^*(a, \lambda)\|$, and the opposite inequality follows from submultiplicativity (first use this inequality and submultiplicativity to conclude that $\|(a, \lambda)^*\| =$

$\|(a,\lambda)\|$). A embeds in $A^\dagger$ by $a \mapsto (a,0)$, so $\|(a,0)\| = \sup_{\|b\|=1} \|ab\| \leq \|a\|$, with equality (for $a \neq 0$) if $b = \frac{a^*}{\|a\|}$. $A^\dagger$ is complete since A is complete and $A^\dagger/A$ is one-dimensional (XX.2.1.2.). 👍

The algebra $A^\dagger$ contains A as a (closed) ideal, and $A^\dagger/A \cong \mathbb{C}$. If A is unital, $A^\dagger \cong A \oplus \mathbb{C}$ as C*-algebras (under the map $(a,\lambda) \mapsto (a - \lambda 1, \lambda)$).

XX.16.1.3. As an example, if X is a locally compact noncompact Hausdorff space, it is easy to see that $C_0(X)^\dagger \cong C(X^\dagger)$, where $X^\dagger$ is the one-point compactification of X.

XX.16.1.4. A bounded homomorphism between the Banach algebras A and B extends uniquely to a bounded unital homomorphism from $A^\dagger$ to $B^\dagger$. A *-homomorphism between Banach *-algebras extends to a unital *-homomorphism between the unitizations.

Special Elements in a C*-Algebra

In a C*-algebra, certain types of elements have standard names arising from operator theory, reflecting the types of operators they become when the C*-algebra is represented as a concrete C*-algebra of operators:

XX.16.1.5. Definition. Let A be a C*-algebra and $x \in A$. Then x is

self-adjoint if $x = x^*$.

normal if $x^*x = xx^*$.

a *projection* if $x = x^* = x^2$.

a *partial isometry* if x^*x is a projection.

If A is unital, then x is

an *isometry* if $x^*x = 1$.

a *coisometry* if $xx^* = 1$.

unitary if $x^*x = xx^* = 1$.

XX.16.1.6. The self-adjoint elements of A form a closed real vector subspace A_{sa} of A. (Note, however, that A_{sa} is not closed under multiplication unless A is commutative.) If A is unital, then the set of unitaries of A forms a group $\mathcal{U}(A)$. Every self-adjoint or unitary element is normal. If x is self-adjoint, then e^{ix} is unitary. 0 and 1 are projections; isometries (and, in particular, unitaries) are partial isometries. Coisometries are too; in fact, x is a partial isometry if and only if $x = xx^*x$ [one direction is obvious; for the other, if x^*x is a projection and $y = x - xx^*x$, then $y^*y = 0$, so $y = 0$]. Thus if x is a partial isometry, then so is x^*.

XX.16.1.7. If x is any element, then $a = (x + x^*)/2$ and $b = (x - x^*)/2i$ are self-adjoint, and $x = a + ib$; thus $A_{sa} + iA_{sa} = A$. The elements a and b are called the *real* and *imaginary* parts of x. It is obvious from their definitions that $\|a\|, \|b\| \leq \|x\|$.

Norm and Spectrum

XX.16.1.8. It follows from the C*-axiom that every nonzero projection, and hence every nonzero partial isometry (in particular, every unitary), has norm 1. Thus, if x is unitary, then $\sigma_A(x) \subseteq \{\lambda : |\lambda| \leq 1\}$. Since $x^{-1} = x^*$ is also unitary and hence $\sigma_A(x^{-1}) = \{\lambda^{-1} : \lambda \in \sigma_A(x)\}$ is also contained in the unit disk, $\sigma_A(x)$ is actually contained in the unit circle.

The next two results give important properties of the spectrum and norm for self-adjoint and, more generally, normal elements of a C*-algebra.

XX.16.1.9. PROPOSITION. If x is self-adjoint, then $\sigma_A(x) \subseteq \mathbb{R}$.

PROOF: If x is self-adjoint, then e^{ix} is unitary, so

$$\sigma_A(e^{ix}) = \{e^{i\lambda} : \lambda \in \sigma_A(x)\}$$

(XX.12.4.3.(i)) is contained in the unit circle. 👍

XX.16.1.10. PROPOSITION. If A is a C*-algebra and $x \in A$ is normal (e.g. self-adjoint), then $r_\sigma(x) = \|x\|$, i.e. there is a $\lambda \in \sigma_A(x)$ with $|\lambda| = \|x\|$.

PROOF: If x is self-adjoint, then from the C*-axiom $\|x^2\| = \|x\|^2$, and by iteration $\|x^{2^n}\| = \|x\|^{2^n}$ for all n. Thus $r_\sigma(x) = \lim_{n\to\infty} \|x^{2^n}\|^{2^{-n}} = \|x\|$. More generally, if y is normal, then

$$r_\sigma(y) = \lim_{n\to\infty} \|y^{2^n}\|^{2^{-n}} = \lim_{n\to\infty} \|(y^*)^{2^n} y^{2^n}\|^{2^{-n-1}}$$

$$= \lim_{n\to\infty} \|(y^*y)^{2^n}\|^{2^{-n-1}} = [r_\sigma(y^*y)]^{1/2} = \|y^*y\|^{1/2} = \|y\| \ .$$

👍

As a consequence, *-homomorphisms between C*-algebras are automatically contractive:

XX.16.1.11. COROLLARY. If A is a Banach *-algebra, B a C*-algebra, and $\phi : A \to B$ a *-homomorphism, then $\|\phi\| \leq 1$.

PROOF: If $x \in A$, then $\sigma_B(\phi(x)) \cup \{0\} \subseteq \sigma_A(x) \cup \{0\}$. Thus

$$\|\phi(x)\|^2 = \|\phi(x^*x)\| = r_\sigma(\phi(x^*x)) \leq r_\sigma(x^*x) \leq \|x^*x\| \leq \|x\|^2 \ .$$

👍

If A is a unital Banach algebra and B is a unital Banach subalgebra, and $x \in B$, then it is not necessarily true that $\sigma_B(x) = \sigma_A(x)$ (XX.12.3.11.). But if A is a C*-algebra, the spectra are the same:

XX.16.1.12. Corollary. If B is a unital C*-subalgebra of a unital C*-algebra A and $x \in B$, then $\sigma_B(x) = \sigma_A(x)$. If B is a general C*-subalgebra of a general C*-algebra A, and $x \in B$, then $\sigma_B(x) \cup \{0\} = \sigma_A(x) \cup \{0\}$.

Proof: This is true if $x = x^*$ by XX.12.3.12. and XX.16.1.9.. For general x, if $x - \lambda 1$ is invertible in A, then so are $(x - \lambda 1)^*(x - \lambda 1)$ and $(x - \lambda 1)(x - \lambda 1)^*$. Thus they are invertible in B, so $x - \lambda 1$ is left and right invertible in B. 👍

XX.16.2. Commutative C*-Algebras and Continuous Functional Calculus

The main result of this section is *Gelfand's Theorem* (sometimes called the *Gelfand-Naimark Theorem*, although this name is more commonly used for a different theorem, cf. XX.16.4.35.):

XX.16.2.1. Theorem. (i) Let A be a commutative unital C*-algebra. Then there is a compact Hausdorff space X such that A is isometrically *-isomorphic to $C(X)$.

(ii) X can be recovered from $C(X)$ either as $\mathrm{Prim}(C(X)) = \mathrm{Maxspec}(C(X))$ (XIII.7.10.20., XX.16.4.39.) with the hull-kernel topology, or as the set $\widehat{C(X)}$ of unital *-homomorphisms from $C(X)$ to $\mathbb{C}$ with the topology of elementwise convergence (weak-* topology). In particular, if $C(X) \cong C(Y)$, then X is homeomorphic to Y.

(iii) If X and Y are compact Hausdorff spaces and $\phi : X \to Y$ is a continuous function, then $\check{\phi} : C(Y) \to C(X)$ defined by $\check{\phi}(f) = f \circ \phi$ is a unital *-homomorphism, and every unital *-homomorphism arises in this manner.

(iv) The correspondences $X \leftrightarrow C(X)$ and $\phi \leftrightarrow \check{\phi}$ give a contravariant category equivalence between the category of compact Hausdorff spaces and continuous functions and the category of commutative unital C*-algebras and unital *-homomorphisms.

XX.16.2.2. As a result, all the topology of compact Hausdorff spaces is encoded in the corresponding C*-algebras, and in principle all of topology (of compact Hausdorff spaces) can be done on the C*-algebra level. In practice, this is certainly not the way to do most topology! However, the point of view that a general C*-algebra is a "noncommutative topological space" is very powerful and underlies much of the modern theory and applications of C*-algebras. Interaction with topology has gone both ways: ideas and techniques from topology, particularly K-theory, have proved fundamentally useful in operator algebra theory, and on the other hand operator algebra methods have been of significant use in some parts of topology.

XX.16.2.3. There is much more fine structure associated with the category equivalence. For example, if X and Y are compact Hausdorff spaces and $\phi : X \to Y$ is continuous, the *-homomorphism $\check{\phi} : C(Y) \to C(X)$ is surjective (i.e. a quotient map, cf. XX.16.3.15.) if and only if ϕ is injective, and $\check{\phi}$ is injective if and only if ϕ is surjective (i.e. a quotient map, cf. XIII.8.3.13.). The closed ideals of $C(X)$ are in natural one-one correspondence with the closed sets in X: the ideal of $C(X)$ corresponding to a closed subset Z of X is the set J_Z of continuous functions from X to $\mathbb{C}$ vanishing on Z. The quotient C*-algebra $C(X)/J_Z$ is naturally isomorphic to $C(Z)$ via restriction of functions (the Tietze Extension Theorem () for compact Hausdorff spaces is included in this fact, since it implies that every element of $C(Z)$ can be extended to an element of $C(X)$, cf. XX.10.3.12.).

The Nonunital Case

XX.16.2.4. There is a nonunital version of XX.16.2.1., but it cannot be stated quite so cleanly. The commutative C*-algebras are easy to describe: they are all of the form $C_0(X)$, the complex-valued continuous functions vanishing at infinity on a locally compact Hausdorff space X. But the *-homomorphisms from $C_0(X)$ to $C_0(Y)$ do not generally correspond to continuous functions from Y to X. For example, a constant function from $\mathbb{R}$ to $\mathbb{R}$ does not give a *-homomorphism from $C_0(\mathbb{R})$ to $C_0(\mathbb{R})$, and conversely the zero homomorphism from $C_0(\mathbb{R})$ to $C_0(\mathbb{R})$ does not come from a continuous function from $\mathbb{R}$ to $\mathbb{R}$ (there are many less trivial examples too). Instead, the *-homomorphisms from $C_0(X)$ to $C_0(Y)$ can be described as corresponding to *proper* () continuous functions from open subsets of Y to X.

Alternatively, it is much cleaner to consider pointed compact Hausdorff spaces $(Z, *)$, where Z is a compact Hausdorff space and $*$ is a fixed basepoint in Z. There is a natural correspondence with locally compact Hausdorff spaces: if X is locally compact, associate X with $(X^\dagger, \infty)$, where $X^\dagger$ is the one-point compactification of X (). Conversely, if $(Z, *)$ is a pointed compact Hausdorff space, associate the locally compact Hausdorff space $Z \setminus \{*\}$. The pointed space $(Z, *)$ corresponds to a compact Hausdorff space if and only if $*$ is an isolated point in Z. Then, if X is locally compact, we have that $C_0(X)$ is naturally isomorphic to the ideal of $C(X^\dagger)$ consisting of functions which are zero at ∞, i.e. if $(Z, *)$ is a pointed compact Hausdorff space, the associated commutative C*-algebra is the algebra of complex-valued continuous functions on Z vanishing at $*$. It is then easy to see that the *-homomorphisms from $C_0(X)$ to $C_0(Y)$ exactly correspond to continuous functions from $Y^\dagger$ to $X^\dagger$ sending ∞ to ∞, i.e. basepoint-preserving continuous maps. (For example, the extreme case of the zero homomorphism corresponds to the constant function from $Y^\dagger$ to $X^\dagger$ sending everything to ∞.) Thus, as a slight extension of XX.16.2.1., we obtain:

XX.16.2.5. THEOREM. (i) Let A be a commutative C*-algebra. Then there is a compact Hausdorff space Z and point $* \in Z$ such that A is isometrically *-isomorphic to $C^*(Z, *) := C_0(X)$, where $X = Z \setminus \{*\}$ is locally compact.

(ii) X can be recovered from $C_0(X)$ as $\mathrm{Prim}(C_0(X)) = \mathrm{Maxspec}(C_0(X))$ with the hull-kernel topology. Z can be recovered from $C^*(Z, *)$ as the set $\widehat{C^*(Z, *)}$ of *-homomorphisms from $C^*(Z, *)$ to $\mathbb{C}$ with the topology of elementwise convergence (weak-* topology), with $*$ corresponding to the zero homomorphism. In particular, if $C_0(X) \cong C_0(Y)$ for locally compact Hausdorff spaces X and Y, then X is homeomorphic to Y.

(iii) If $(Z, *_Z)$ and $(W, *_W)$ are pointed compact Hausdorff spaces and $\phi : Z \to W$ is a pointed continuous function (i.e. ϕ is continuous and $\phi(*_Z) = *_W$), then $\check{\phi} : C^*(W, *_W) \to C^*(Z, *_Z)$ defined by $\check{\phi}(f) = f \circ \phi$ is a *-homomorphism, and every *-homomorphism arises in this manner.

(iv) The correspondences $(Z, *) \leftrightarrow C^*(Z, *)$ and $\phi \leftrightarrow \check{\phi}$ give a contravariant category equivalence between the category of pointed compact Hausdorff spaces and pointed continuous functions and the category of commutative C*-algebras and *-homomorphisms.

XX.16.2.6. The fine structure described in XX.16.2.3. persists in this setting, but care must be exercised in phrasing it. Details are left to the interested reader.

Proof of the Theorem

We now prove Gelfand's Theorem. We will only prove the unital version XX.16.2.1.; the proof of the nonunital version XX.16.2.5. is a straightforward variation, but the additional technicalities detract from

an intuitive appreciation of the argument. We first prove (ii) and (iii), which are fairly elementary (given Urysohn's Lemma), and then return to (i) which requires some nontrivial Banach algebra theory. The proof will actually give some fine structure beyond the statement of the theorem.

We begin with a characterization of the maximal ideals of a $C(X)$.

XX.16.2.7. LEMMA. Let X be a compact Hausdorff space. For each $p \in X$, the set M_p of all functions in $C(X)$ which vanish at p is a maximal ideal in $C(X)$, and every maximal ideal of $C(X)$ is of this form. If $p, q \in X$, $p \neq q$, then $M_p \neq M_q$.

PROOF: Let $p \in X$. It is obvious that $C(X)/M_p$ is isomorphic to $\mathbb{C}$ via the correspondence $f \mapsto f(p)$. Since the quotient is a field, M_p is a maximal ideal in $C(X)$. Conversely, suppose M is a maximal ideal in $C(X)$, and suppose that there is no point of X for which all functions in M vanish. For each $p \in X$, there is an $f_p \in M$ with $f_p(p) \neq 0$. The set $U_p = \{x \in X : f_p(x) \neq 0\}$ is an open set in X containing p, and $\{U_p : p \in X\}$ is an open cover of X. By compactness, there is a finite subcover $\{U_{p_1}, \ldots, U_{p_n}\}$. Set

$$g = f_{p_1}\bar{f}_{p_1} + \cdots + f_{p_n}\bar{f}_{p_n} \in M$$

and note that $g(x) > 0$ for every $x \in X$. Thus g is invertible in $C(X)$, a contradiction since a proper ideal cannot contain an invertible element. Thus there is some $p \in X$ at which all functions in M vanish, so $M \subseteq M_p$, and $M = M_p$ by maximality. Since functions in $C(X)$ separate points of X by Urysohn's Lemma, and constant functions are in $C(X)$, we have $M_p \neq M_q$ if $p \neq q$. 👍

XX.16.2.8. Thus $p \leftrightarrow M_p$ is a one-one correspondence between X and $\mathrm{Maxspec}(C(X))$. If ϕ is a unital *-homomorphism from $C(X)$ to $\mathbb{C}$, then the kernel of ϕ is a maximal ideal of $C(X)$, hence of the form M_p for some $p \in X$, and ϕ must just be evaluation $\hat{p}$ at p, i.e. $\phi(f) = \hat{p}(f) = f(p)$. Thus there is also a natural one-one correspondence between X and $\widehat{C(X)}$, so as sets X, $\mathrm{Maxspec}(C(X))$, and $\widehat{C(X)}$ can be naturally identified.

XX.16.2.9. We now show how the topology of X can be recovered as the hull-kernel topology on $\mathrm{Maxspec}(C(X))$. If E is a subset of $X \cong \mathrm{Maxspec}(C(X))$, recall that the closure $\tilde{E}$ of E in the hull-kernel topology consists of all M_p such that $J := \cap_{q \in E} M_q \subseteq M_p$. The ideal J clearly consists of all f such that $f(q) = 0$ for all $q \in E$. If $\bar{E}$ is the closure in X of E regarded as a subset of X, and $p \in \bar{E}$, then every $f \in J$ vanishes at p by continuity, so $M_p \in \tilde{E}$. But if $p \notin \bar{E}$, by Urysohn's Lemma there is an $f \in C(X)$ such that $f = 0$ on $\bar{E}$ and $f(p) = 1$; this f is in J but not M_p, so $M_p \notin \tilde{E}$. Thus $\tilde{E} = \bar{E}$ and the correspondence $p \leftrightarrow M_p$ is a homeomorphism.

XX.16.2.10. Now consider the topology of pointwise convergence on $\widehat{C(X)}$: $\hat{p}_i \to \hat{p}$ if and only if $\hat{p}_i(f) \to \hat{p}(f)$ for every $f \in C(X)$, i.e. $f(p_i) \to f(p)$ for all $f \in C(X)$. If $p_i \to p$ in X, this condition is clearly satisfied. If $p_i \not\to p$ in X, there is an open neighborhood U of p such that $p_i \notin U$ for a cofinal set of i. By Urysohn's Lemma there is a function $f \in C(X)$ with $f(p) = 1$ and $f = 0$ on U^c, and $f(p_i) \not\to f(p)$ for this f, i.e. $\hat{p}_i \not\to \hat{p}$. Thus $p \leftrightarrow \hat{p}$ is a homeomorphism.

XX.16.2.11. Finally, we observe that every primitive ideal (XX.16.4.39.) of $C(X)$ is of the form M_p. By definition, primitive ideals are the kernels of irreducible representations. Every irreducible representation of a commutative C*-algebra is one-dimensional, hence a unital *-homomorphism to $\mathbb{C}$, hence $\hat{p}$ for some $p \in X$, and its kernel is M_p. Alternatively, every primitive ideal is a closed prime ideal (XX.16.4.11.), so it suffices to prove the following fact:

XX.16.2.12. Proposition. Let X be a compact Hausdorff space, and J a prime ideal of $C(X)$. Then there is a unique $p \in X$ such that $O_p \subseteq J \subseteq M_p$, where

$$O_p = \{f \in C(X) : f = 0 \text{ in a neighborhood of } p\} .$$

The straightforward proof is left to the reader (cf. [GJ76, 7.15]). Since O_p is dense in M_p, it follows that every closed prime ideal of $C(X)$ is maximal.

XX.16.2.13. This completes the proof of (ii). We now turn to (iii). Let X and Y be compact Hausdorff spaces. If $\phi : X \to Y$ is continuous, it is obvious that $\check{\phi} : C(Y) \to C(X)$, defined by $\check{\phi}(f) = f \circ \phi$, is well defined and is a unital *-homomorphism. Conversely, let α be a unital *-homomorphism from $C(Y)$ to $C(X)$. For any $p \in X$, $\hat{p} \circ \alpha$ is a unital *-homomorphism from $C(Y)$ to $\mathbb{C}$, hence $\hat{p} \circ \alpha = \hat{q}$ for a unique $q \in Y$. Set $\phi(p) = q$. This defines a function $\phi : X \to Y$, and $\alpha(f) = f \circ \phi$ for all $f \in C(X)$, and in particular $f \circ \phi \in C(X)$ for all $f \in C(Y)$. It remains only to show that ϕ is continuous. The argument is similar to the one in XX.16.2.10.. If ϕ is discontinuous at $p \in X$, then there is a net (p_i) in X with $p_i \to p$ in X but $\phi(p_i) \not\to \phi(p)$ in Y. There is an open neighborhood U of $\phi(p)$ in Y with $\phi(p_i) \notin U$ for a cofinal set of i. By Urysohn's Lemma there is an $f \in C(Y)$ with $f(\phi(p)) = 1$ and $f = 0$ on U^c. Then $f(\phi(p_i)) \not\to f(\phi(p))$, i.e. $f \circ \phi$ is not continuous at p, a contradiction. Thus ϕ is continuous.

Now we tackle (i). We will need to bring in some Banach algebra machinery. If A is a unital commutative Banach algebra, write $\hat{A}$ for the set of unital homomorphisms from A to $\mathbb{C}$. Recall the definitions of spectrum and spectral radius ().

XX.16.2.14. Proposition. Let A be a unital commutative Banach algebra, and $\phi \in \hat{A}$. If $x \in A$, then $\phi(x) \in \sigma_A(x)$. In particular, $|\phi(x)| \leq r_\sigma(x) \leq \|x\|$. Thus ϕ is bounded and $\|\phi\| = 1$.

Proof: Since $\phi(x - \phi(x)1) = 0$, $x - \phi(x)1$ is not invertible in A. The inequality in the second statement is then obvious, so $\|\phi\| \leq 1$. But $\phi(1) = 1$, so $\|\phi\| = 1$. 👍

XX.16.2.15. Corollary. Let A be a unital commutative C*-algebra, and $\phi \in \hat{A}$. Then ϕ is a *-homomorphism.

Proof: It suffices to show that if $x = x^* \in A$, then $\phi(x) \in \mathbb{R}$. But $\sigma_A(x) \subseteq \mathbb{R}$ by XX.16.1.9.. 👍

There is a converse to XX.16.2.14.:

XX.16.2.16. PROPOSITION. Let A be a unital commutative Banach algebra and $x \in A$. If $\lambda \in \sigma_A(x)$, then there is a $\phi \in \hat{A}$ with $\phi(x) = \lambda$.

PROOF: Since $x - \lambda 1$ is not invertible in A, it generates a proper ideal of A which is contained in a maximal ideal M of A. Then M is closed (XX.12.2.17.), so A/M is a Banach algebra which is a field. Thus $A/M \cong \mathbb{C}$ (XX.12.3.28.), and the quotient map $\phi : A \to A/M$ can be regarded as an element of $\hat{A}$ with $\phi(x - \lambda 1) = 0$. 👍

XX.16.2.17. Let A be a unital commutative Banach algebra. By XX.16.2.14., $\hat{A}$ can be regarded as a subset of the closed unit ball B of A^*. It is easily checked that $\hat{A}$ is closed in B in the weak-* topology (topology of pointwise convergence on A). Since B is a compact Hausdorff space in the weak-* topology (), $\hat{A}$ is a compact Hausdorff space in the topology of pointwise convergence on A. If $x \in A$, define $\check{x} : \hat{A} \to \mathbb{C}$ by $\check{x}(\phi) = \phi(x)$ for $\phi \in \hat{A}$. Then $\check{x} \in C(\hat{A})$.

XX.16.2.18. DEFINITION. The map $\gamma : A \to C(\hat{A})$ defined by $\gamma(x) = \check{x}$ is a unital homomorphism, called the *Gelfand transform* of A.

We have that γ is bounded and $\|\gamma\| = 1$ by XX.16.2.14.. The Gelfand transform is not injective or surjective for general unital commutative Banach algebras. However:

XX.16.2.19. LEMMA. Let A be a unital commutative C*-algebra. Then the Gelfand transform $\gamma : A \to C(\hat{A})$ is an isometric *-isomorphism.

PROOF: It follows from XX.16.2.15. that γ is a *-homomorphism. Let $x \in A$. Then by () there is a $\lambda \in \sigma_A(x^*x)$ with $|\lambda| = \|x^*x\| = \|x\|^2$, and hence by XX.16.2.16. there is a $\phi \in \hat{A}$ with $|\phi(x)|^2 = |\phi(x^*x)| = \|x^*x\| = \|x\|^2$. Since $\|\gamma\| = 1$, γ is isometric, and hence A can be regarded as the *-subalgebra $\gamma(A)$ of $C(\hat{A})$, which is complete and thus closed. The subalgebra $\gamma(A)$ clearly separates points of $\hat{A}$ and contains the constant functions, so $\gamma(A) = C(\hat{A})$ by the Stone-Weierstrass Theorem (XX.10.2.3.). 👍

XX.16.2.20. This completes the proof of (i). Part (iv) is essentially just a recapitulation of (i)–(iii); the only technicality is the easily verified fact that if X, Y, and Z are compact Hausdorff spaces and $\phi : X \to Y$ and $\psi : Y \to Z$ are continuous, then $(\psi \circ \phi)\check{} = \check{\phi} \circ \check{\psi}$.

This completes the proof of Theorem XX.16.2.1.. 💪

Applications

Probably the most important application of Gelfand's Theorem is continuous functional calculus, discussed in XX.16.2.25.. But here are a couple of other applications of some interest.

XX.16.2.21. Let X be a topological space. Let $BC(X)$ be the set of bounded continuous complex-valued functions on X. With pointwise addition and multiplication, pointwise complex conjugation as adjoint, and supremum norm, $BC(X)$ becomes a unital commutative C*-algebra. Thus $\widehat{BC(X)} \cong C(Y)$ for a compact Hausdorff space Y. Each point of X gives, by point evaluation, an element of $\widehat{BC(X)}$ and hence a point of Y, and it is easy to check that the induced map ϕ from X to Y is continuous. If X is completely regular, a simple argument yields that ϕ is a homeomorphism onto its image, so X can be identified with a subspace of Y. In this case, it is straightforward to identify Y with the Stone-Čech compactification βX. In fact, realizing βX as $\widehat{BC(X)}$ or $\mathrm{Prim}(BC(X))$ is perhaps the most elegant and efficient construction of the Stone-Čech compactification.

XX.16.2.22. Here is another interesting and modestly useful application. Give $A = L^\infty(\mathbb{R})$ pointwise addition, multiplication, and complex conjugation, and the usual essential supremum norm. Then A becomes a unital commutative C*-algebra, so $A \cong C(Y)$ for some compact Hausdorff space Y. The space Y is rather horrendous as a topological space: it is extremally disconnected and not first countable. But it is occasionally useful to realize A as an algebra of continuous functions on a bad space instead of as measurable functions on a nicer space. The same construction can be done for $L^\infty(X, \mathcal{A}, \mu)$ for any measure space $(X, \mathcal{A}, \mu)$. If done for $\mathbb{N}$ with counting measure one obtains $Y = \beta\mathbb{N}$ as in XX.16.2.21.. These $L^\infty(X, \mathcal{A}, \mu)$ are commutative von Neumann algebras (XX.5.5.1.) when realized as algebras of multiplication operators on $L^2(X, \mathcal{A}, \mu)$.

Continuous Functional Calculus

The classification of commutative C*-algebras leads to a far-reaching and powerful extension of holomorphic functional calculus (XX.12.4.).

XX.16.2.23. Let A be a unital C*-algebra, and x a normal (e.g. self-adjoint or unitary) element of A, i.e. $x^*x = xx^*$. Let B be the C*-subalgebra of A generated by x and 1; B is the closure of the set of polynomials in x and x^*. Then B is a unital commutative C*-algebra, and thus $B \cong C(X)$ for a compact Hausdorff space $X = \hat{B}$.

XX.16.2.24. The space $\hat{B}$ can be easily described. If $\phi \in \hat{B}$, then $\phi(x) \in \sigma_B(x) = \sigma_A(x)$ (XX.16.2.14., XX.16.1.12.). Conversely, for each $\lambda \in \sigma_A(x)$ there is a unique $\phi_\lambda \in \hat{B}$ with $\phi_\lambda(x) = \lambda$ by XX.16.2.16. [for uniqueness, if $\phi, \psi \in \hat{B}$ with $\phi(x) = \psi(x) = \lambda$, then $\phi(x^*) = \psi(x^*) = \bar{\lambda}$ by XX.16.2.15., so $\phi(p(x, x^*)) = \psi(p(x, x^*)) = p(\lambda, \bar{\lambda})$ for any polynomial p of two variables with complex coefficients; such elements are dense in B, so $\phi = \psi$]. Thus

$$\hat{B} = \{\phi_\lambda : \lambda \in \sigma_A(x)\} .$$

More precisely, the function $\check{x} : \hat{B} \to \mathbb{C}$ defined by $\check{x}(\phi_\lambda) = \lambda$ is a bijection from $\hat{B}$ to $\sigma_A(x)$. Since $\check{x}$ is continuous and $\hat{B}$ and $\sigma_A(x)$ are compact Hausdorff spaces, it is a homeomorphism. We thus have:

XX.16.2.25. Theorem. [Continuous Functional Calculus, Unital Version] Let A be a unital C*-algebra, x a normal element of A, and B the C*-subalgebra of A generated by x and 1. Then there is an isometric *-isomorphism $\theta : C(\sigma_A(x)) \to B$ with $\theta(\iota) = x$, where $\iota(\lambda) = \lambda$ for $\lambda \in \sigma_A(x) \subseteq \mathbb{C}$.

XX.16.2.26. We have that θ is just $(\check{x})^{\check{}}$. If $f \in C(\sigma_A(x))$, we normally write $f(x)$ instead of $\theta(f)$; thus $\iota(x) = x$. This notation is consistent with previous notation: if p is a polynomial with complex coefficients, then $p(x)$ is the element of B defined in XX.12.4.1.. More generally, if f is holomorphic on an open disk containing $\sigma_A(x)$, then $f(x)$ is the element defined by a power series as in XX.12.4.3.. Thus "functional calculus" can be thought of as a way to apply suitable functions to elements of a C*-algebra. Note that holomorphic functional calculus is defined for arbitrary elements of a C*-algebra (or even a general Banach algebra), but continuous functional calculus only works for normal elements of a C*-algebra.

There is a nonunital version of continuous functional calculus, which can be easily proved from the unital version by simply adding a unit (XX.16.1.2.). If X is a compact subset of $\mathbb{C}$, write $C_0(X)$ for the set of continuous functions $f : X \cup \{0\} \to \mathbb{C}$ with $f(0) = 0$.

XX.16.2.27. Theorem. [Continuous Functional Calculus, General Version] Let A be a C*-algebra, x a normal element of A, and B the C*-subalgebra of A generated by x. Then there is an isometric *-isomorphism $\theta : C_0(\sigma_A(x)) \to B$ with $\theta(\iota) = x$, where $\iota(\lambda) = \lambda$ for $\lambda \in \sigma_A(x) \cup \{0\} \subseteq \mathbb{C}$.

Note that if $0 \notin \sigma_A(x)$, then A is automatically unital and 1 is in the C*-subalgebra of A generated by x.

The following properties of functional calculus are easy consequences:

XX.16.2.28. Proposition. Let A be a C*-algebra and $x \in A$ a normal element. Then

(i) For any $f \in C_0(\sigma_A(x))$, $\sigma_A(f(x)) = \{f(\lambda) : \lambda \in \sigma_A(x)\}$.

(ii) If $f \in C_0(\sigma_A(x))$ and $g \in C_0(f(\sigma_A(x))) = C_0(\sigma_A(f(x)))$, so $g \circ f \in C_0(\sigma_A(x))$, then $(g \circ f)(x) = g(f(x))$.

(iii) If $f_n, f \in C_0(\sigma_A(x))$ and $f_n \to f$ uniformly on $\sigma_A(x)$, then $f_n(x) \to f(x)$.

(iv) If B is a C*-algebra and $\phi : A \to B$ a *-homomorphism, then $\phi(f(x)) = f(\phi(x))$ for any $f \in C_0(\sigma_A(x))$.

There is a useful joint continuity generalization of (iii):

XX.16.2.29. Proposition. Let Y be a compact subset of $\mathbb{C}$, and (f_n) a sequence of elements of $C_0(Y)$ converging uniformly on Y to f. Let A be a C*-algebra, (x_n) a sequence of normal elements of A with $x_n \to x \in A$ and $\sigma_A(x_n) \subseteq Y$ (so $\sigma_A(x) \subseteq Y$). Then $f_n(x_n) \to f(x)$.

Proof: Let $\epsilon > 0$, and approximate f uniformly on Y within $\epsilon/4$ by a function of the form $g(\lambda) = p(\lambda, \bar{\lambda})$, where p is a polynomial of two variables (XX.10.3.4.); then $g(x) = p(x, x^*)$, so

$$\|f_n(x_n) - f(x)\| \leq$$

$$\|f_n(x_n) - f(x_n)\| + \|f(x_n) - g(x_n)\| + \|p(x_n, x_n^*) - p(x, x^*)\| + \|g(x) - f(x)\|.$$

If n is large enough that $|f_n - f|$ is uniformly less than $\epsilon/4$ on Y, the first term is less than $\epsilon/4$; the second and fourth terms are $< \epsilon/4$ by choice of g, and the third term goes to 0 as $n \to \infty$ by continuity of addition, multiplication, and involution. Thus $\|f_n(x_n) - f(x)\| < \epsilon$ for all sufficiently large n. 👍

This result is often used when (f_n) is a constant sequence.

Another simple consequence of functional calculus is:

XX.16.2.30. Corollary. Let A be a C*-algebra, and x a normal element of A. Then

x is self-adjoint if and only if $\sigma_A(x) \subseteq \mathbb{R}$.

x is unitary if and only if $\sigma_A(x) \subseteq \{\lambda : |\lambda| = 1\}$.

x is a projection if and only if $\sigma_A(x) \subseteq \{0, 1\}$.

Proof: (i): Let $f(\lambda) = Re(\lambda) = \frac{1}{2}(\lambda + \bar{\lambda})$. Then $f(x) = \frac{1}{2}(x + x^*)$ is the real part of x. We have $f(x) = x$ if and only if $f = \iota$ on $\sigma_A(x)$, i.e. if and only if $\sigma_A(x) \subseteq \mathbb{R}$. The other parts are similar. 👍

Uniqueness of Norm on a C*-Algebra

Another important consequence of Gelfand's Theorem is that the norm on a C*-algebra is unique in a strong sense. Thus C*-algebras are quite rigid.

We first prove one of the fine structure properties of commutative C*-algebras (XX.16.2.3.).

XX.16.2.31. Lemma. Let A and B be commutative C*-algebras, and let $\alpha : A \to B$ be a *-homomorphism. If α is injective, it is isometric.

Proof: Write $A = C^*(Z, *)$ and $B = C^*(W, *)$. We have $\alpha = \check{\phi}$ for a pointed continuous function $\phi : (W, *) \to (Z, *)$. If α is injective, then ϕ is surjective: set $Y = \phi(W)$ and suppose there is a $p \in Z \setminus Y$. By Urysohn's Lemma there is an $f \in C(Z)$ with $f = 0$ on Y and $f(p) = 1$. Then $f \in C^*(Z, *)$ since $* \in Y$, and $\alpha(f) = 0$, contradicting that α is injective.

Thus, if $f \in C^*(Z, *)$, we have

$$\|f\| = \sup_{p \in Z} |f(p)| = \sup_{q \in W} |f(\phi(q))| = \|\alpha(f)\| .$$

👍

XX.16.2.32. Corollary. Let A and B be C*-algebras, $\alpha : A \to B$ an injective *-homomorphism. Then α is isometric.

Proof: If $x \in A$, then $\|x\|^2 = \|x^*x\|$ and $\|\alpha(x)\|^2 = \|\alpha(x^*x)\|$, so replacing A and B by $C^*(x^*x)$ and $C^*(\alpha(x^*x))$ respectively, it suffices to assume A and B are commutative, where the result follows from the lemma. 👍

XX.16.2.33. Corollary. If A is a C*-algebra with respect to a norm $\|\cdot\|$, then there is no other C*-norm $\|\cdot\|'$ on A (with the same algebraic structure and involution), even incomplete.

Proof: Let B be the completion of A with respect to $\|\cdot\|'$. Then B is a C*-algebra, and the natural injective *-homomorphism from $(A, \|\cdot\|)$ to B is isometric. 👍

XX.16.2.34. A *-algebra thus has at most one norm making it a (complete) C*-algebra; if it has one, it has no other (even incomplete) norm satisfying the C*-axiom (XX.16.2.33.). [A C*-algebra can have other C*-seminorms; and a *-algebra (e.g. a polynomial ring) can have many different (incomplete) norms satisfying the C*-axiom.] Note that by XX.16.1.10., the norm in a C*-algebra is completely determined by the algebraic structure: if $x \in A$, then

$$\|x\| = \|x^*x\|^{1/2} = [r_\sigma(x^*x)]^{1/2}$$

and $r_\sigma(x^*x)$ is a purely algebraic quantity.

XX.16.3. Algebraic and Order Theory of C*-Algebras

Positive Elements in a C*-Algebra

XX.16.3.1. DEFINITION. Let A be a C*-algebra. An element $x \in A$ is *positive* if $x = x^*$ and $\sigma_A(x) \subseteq [0, \infty)$. The set of positive elements in A is denoted A_+. If $x \in A_+$, we write $x \geq 0$.

Note that by XX.16.1.12., the property of being positive is independent of the containing C*-algebra, i.e. if B is a C*-subalgebra of A, then $B_+ = B \cap A_+$. A positive element of $C_0(X)$ is just a function taking only nonnegative real values.

The following facts are obvious from functional calculus:

XX.16.3.2. PROPOSITION. Let A be a C*-algebra and $x, y \in A$. Then

(i) If $x \geq 0$ and $-x \geq 0$, then $x = 0$.

(ii) If x is normal, then $x^*x \geq 0$. In particular, if $x = x^*$, then $x^2 \geq 0$.

(iii) If $x \geq 0$, then $\|x\| = \max\{\lambda : \lambda \in \sigma(x)\}$.

(iv) If $x = x^*$ and $\|x\| \leq 2$, then $x \geq 0$ if and only if $\|1 - x\| \leq 1$ (in $\tilde{A}$).

(v) If $x, y \geq 0$ and $xy = yx$, then $x + y$ and xy are positive.

(vi) If $x = x^*$, then there is a unique decomposition $x = x_+ - x_-$, where $x_+, x_- \geq 0$ and $x_+x_- = 0$. We have $x_+, x_- \in C^*(x)$. [$x_+ = f(x)$ and $x_- = g(x)$, where $f(t) = \max(t, 0)$ and $g(t) = -\min(t, 0)$.] Thus every element of A is a linear combination of four positive elements.

(vii) Every positive element of a C*-algebra has a unique positive square root. More generally, if $x \geq 0$ and α is a positive real number, there is a positive element $x^\alpha \in C^*(x)_+$; these elements satisfy $x^\alpha x^\beta = x^{\alpha+\beta}$, $x^1 = x$, and $\alpha \mapsto x^\alpha$ is continuous. If x is invertible x^α is also defined for $\alpha \leq 0$.

(viii) $(x, \lambda) \geq 0$ in $A^\dagger$ if and only if $x = x^*$ and $\lambda \geq \|x_-\|$.

For (vii), set $x^\alpha = g_\alpha(x)$, where $g_\alpha(t) = t^\alpha$. Then $g_{\alpha^{-1}}(x^\alpha) = x$. If $b \in A_+$ with $g_{\alpha^{-1}}(b) = x$, then x commutes with b, hence x^α commutes with b since it is a limit of polynomials in x; thus $b = x^\alpha$ in $C^*(x^\alpha, b) \subseteq A$ (we clearly have uniqueness in a commutative C*-algebra).

Two crucial facts about positive elements are significant generalizations of (ii) and (v):

XX.16.3.3. PROPOSITION. Let A be a C*-algebra. Then

(i) A_+ is a closed cone () in A; in particular, if $x, y \geq 0$ then $x + y \geq 0$.

(ii) If $x \in A$, then $x^*x \geq 0$.

PROOF: To prove (i), since $\mathbb{R}_+A_+ \subseteq A_+$ it suffices to show that $A_+ \cap B$ is a closed convex set, where $B = B_1(A)$ is the closed unit ball in A. But $A_+ \cap B$ is the intersection of the closed convex sets A_{sa}, B, and $\{x : \|1 - x\| \leq 1\}$ (XX.16.3.2.(iv)).

The proof of property (ii) uses the fact that if $x = a + ib \in A$ with $a, b \in A_{sa}$, then $x^*x \geq 0$ if and only if $xx^* \geq 0$ by XX.12.3.20., and

$$x^*x + xx^* = 2a^2 + 2b^2 \geq 0$$

by (i) and XX.16.3.2.(ii). If $x^*x = c = c_+ - c_-$ as in XX.16.3.2.(vi) and $y = xc_-$, then $-y^*y = c_-^3 \geq 0$, so

$$yy^* = (y^*y + yy^*) + (-y^*y) \geq 0$$

and thus $y^*y \geq 0$. So $y^*y = 0$ by XX.16.3.2.(i), $c_- = 0$, $x^*x \geq 0$. 👍

XX.16.3.4. These properties were mysterious in the early days of the subject, and property (ii) was originally assumed as an axiom for C*-algebras, in the equivalent form that $1 + x^*x$ is invertible for all x. (GELFAND, NAIMARK, and SEGAL were all convinced that the additional axiom was superfluous, and said so in their papers.) Property (i) was proved independently in [Fuk52] and [KV53], and (ii) was proved by KAPLANSKY (cf. [Sch52]; this is one of the few significant results in mathematics whose proof first appeared in the *Mathematical Reviews*!) [Kap68] and [DB86] contain an account of the history of the development of the C*-axioms.

XX.16.3.5. COROLLARY. If A is a C*-algebra and $a \in A_+$, then $x^*ax = (a^{1/2}x)^*(a^{1/2}x) \geq 0$ for any $x \in A$.

By XX.16.3.2.(vi), there are "many" positive elements of a C*-algebra. Unital C*-algebras also have "many" unitaries:

XX.16.3.6. PROPOSITION. Every element of a unital C*-algebra A is a linear combination of four unitaries. In fact, if $x = x^* \in A$ and $\|x\| \leq 2$, then x is a sum of two unitaries in $\mathcal{U}(A)_o$ (where $\mathcal{U}(A)_o$ is the connected component of 1 in $\mathcal{U}(A)$).

PROOF: Set $a = x/2$. Then $1 - a^2 \geq 0$ and $a \pm i(1-a^2)^{1/2}$ are unitaries. If $0 \leq t \leq 1$, then $(ta \pm i(1-t^2a^2)^{1/2})$ gives a path of unitaries from $\pm i1$ to $a \pm i(1-a^2)^{1/2}$, so $a \pm i(1-a^2)^{1/2} \in \mathcal{U}(A)_o$. 👍

Invertible elements have a polar decomposition analogous to ():

XX.16.3.7. Proposition. Let A be a unital C*-algebra, and $x \in A$ an invertible element. Then there are a unique unitary $u \in A$ and $a \in A_+$ with $x = ua$. In fact, $a = (x^*x)^{1/2}$ and $u = x(x^*x)^{-1/2}$.

Proof: If $u = x(x^*x)^{-1/2}$ (note that x^*x is invertible), then $u^*u = 1$, and in particular u is left invertible. We have $uu^* = x(x^*x)^{-1}x^*$ is invertible, hence u is also right invertible, i.e. u is invertible and $u^{-1} = u^*$, so u is unitary. Obviously u and $a = (x^*x)^{1/2} \geq 0$ give such a decomposition. For uniqueness, if $x = ua = vb$, then $x^*x = a^2 = b^2$, so $a = b$ (XX.16.3.2.(vii)), and then $u = v = xa^{-1}$. 👍

Noninvertible elements of a C*-algebra do not have an analogous polar decomposition in general; but if x is any element of a C*-algebra A and $0 < \alpha < 1/2$, then there is a $u \in A$ (not unique in general) such that $x = u(x^*x)^\alpha$, so x has an approximate polar decomposition.

XX.16.3.8. Because of XX.16.3.3.(i), it makes sense in a C*-algebra A to write $x \leq y$ if $y - x \geq 0$. This defines a translation-invariant partial order on A (it is usually only used on A_{sa}). Note that $x \leq y$ is well defined independent of the containing C*-algebra. If $x = x^*$ and $\sigma_A(x) \subseteq [\alpha, \beta]$, then $\alpha 1 \leq x \leq \beta 1$ (in $A^\dagger$). If $a \leq b$, then $x^*ax \leq x^*bx$ for any x. If $x \in A$, $a \in A_+$, then $x^*ax \leq x^*(\|a\|1)x = \|a\|x^*x$. In $C_0(X)$, if f and g are self-adjoint (i.e. real-valued), then $f \leq g$ in $C_0(X)$ if and only if $f(x) \leq g(x)$ for all $x \in X$.

XX.16.3.9. Proposition. Let A be a unital C*-algebra, $x, y \in A$. If $0 \leq x \leq y$ and x is invertible, then y is invertible and $0 \leq y^{-1} \leq x^{-1}$.

Proof: If x is invertible, then $\epsilon 1 \leq x$ for some $\epsilon > 0$, so $\epsilon 1 \leq y$ and y is invertible by XX.16.2.1., and $x^{-1}, y^{-1} \geq 0$. The inequality $y^{-1} \leq x^{-1}$ is obvious from XX.16.2.1. if x and y commute, and in particular if $y = 1$. For the general case, if $x \leq y$, then

$$y^{-1/2}xy^{-1/2} \leq y^{-1/2}yy^{-1/2} = 1$$

$$1 \leq (y^{-1/2}xy^{-1/2})^{-1} = y^{1/2}x^{-1}y^{1/2}$$

$$y^{-1} = y^{-1/2}1y^{-1/2} \leq y^{-1/2}(y^{1/2}x^{-1}y^{1/2})y^{-1/2} = x^{-1}.$$

👍

Approximate Units

While a C*-algebra need not be unital, it always contains an approximate unit (sometimes called an approximate identity) which can often be used as a substitute. Some, but not all, other Banach algebras have approximate units in an analogous sense.

XX.16.3.10. Definition. Let A be a C*-algebra. An *approximate unit* for A is a net (h_λ) of positive elements of A of norm ≤ 1, indexed by a directed set Λ, such that $h_\lambda x \to x$ for all $x \in A$. If $h_\lambda \leq h_\mu$ for $\lambda \leq \mu$, the approximate unit (h_λ) is *increasing*.

An approximate unit (h_λ) is *idempotent* if each h_λ is a projection; it is *almost idempotent* if $h_\lambda \ll h_\mu$ for $\lambda < \mu$.

An approximate unit (h_λ) is *sequential* if $\Lambda = \mathbb{N}$; it is *continuous* if $\Lambda = (0, \infty)$ (or a cofinal subinterval) and $\lambda \mapsto h_\lambda$ is continuous.

In many references, approximate units are assumed to be increasing, but we will not do so since there are situations where this assumption is not natural. At the other extreme, the assumption that $\|h_\lambda\| \leq 1$ for all λ, or even that $\|h_\lambda\|$ is bounded, is not universal.

An approximate unit in a unital C*-algebra is just a net of (eventually) invertible positive elements in the unit ball converging to 1.

If A is a C*-algebra and I is a dense (two-sided) ideal in A (the case $I = A$ is the most important one!), let Λ_I be the set of all positive elements of I of norm strictly less than 1. Give Λ_I its ordering as a subset of A_+.

XX.16.3.11. PROPOSITION.

(i) Λ_I is a directed set.

(ii) Λ_I is an increasing approximate unit for A.

PROOF: (i): By XX.16.3.9.(iii), $h \mapsto (1-h)^{-1} - 1$ is an order-isomorphism of Λ_I onto I_+, which is a directed set. (The computation is done in $A^\dagger$, but it is easy to check that the map and its inverse $h \mapsto 1 - (1+h)^{-1}$ map I into I by considering the quotient map $A^\dagger \to A^\dagger/I$.)

(ii): It suffices to show that for any $x \in A$, the decreasing net

$$\{x^*(1-h)x : h \in \Lambda_I\}$$

converges to 0, for then

$$\|(1-h)x\|^2 = \|x^*(1-h)^2x\| \leq \|x^*(1-h)x\| \to 0.$$

Since Λ_I is dense in Λ_A [if $x \in A_+$, choose $y \in I$ close to $x^{1/2}$; then $y^*y \in I_+$ is close to x], we need only find for each x and ϵ an $h \in \Lambda_A$ with $\|x^*(1-h)x\| < \epsilon$. By XX.16.3.2.(vi), we may assume $x \geq 0$ and $\|x\| < 1$. But then $\|x(1-x^{1/n})x\| < \epsilon$ for sufficiently large n. ☝

XX.16.3.12. COROLLARY. If A is a C*-algebra and I is a dense (two-sided) ideal in A, then there is an increasing approximate unit for A contained in I. If A is separable, then the increasing approximate unit may be chosen to be sequential or continuous.

To construct an increasing sequential approximate unit for a separable A, let $\{x_n\}$ be a dense set in A, and inductively find h_n in Λ_I with $h_n \geq h_{n-1}$ and $\|h_n x_k - x_k\| < 1/n$ for $1 \leq k \leq n$. For a continuous approximate unit, set $h_t = (n+1-t)h_n + (t-n)h_{n+1}$ for $n \leq t \leq n+1$.

Closed Ideals and Quotients

In this section, "ideal" means "two-sided ideal" unless otherwise specified.

XX.16.3.13. PROPOSITION. Let A be a C*-algebra, and J a closed ring-theoretic ideal in A. Then J is closed under scalar multiplication.

PROOF: This is obvious if A is unital. For general A, let (h_λ) be an approximate unit for A. If $x \in J$ and $\alpha \in \mathbb{C}$, then $(\alpha h_\lambda)x \in J$ for all λ and $(\alpha h_\lambda)x = \alpha(h_\lambda x) \to \alpha x$. Since J is closed, $\alpha x \in J$. ☝

XX.16.3.14. PROPOSITION. Let J be a closed ideal in a C*-algebra A, and π the quotient map from A to the Banach algebra A/J. Then

(i) J is self-adjoint (so J is a C*-algebra and there is an induced involution on A/J).

(ii) If (h_λ) is an approximate unit for J, then for all $x \in A$,

$$\|\pi(x)\| = \lim_\lambda \|x(1-h_\lambda)\| = \inf \|x(1-h_\lambda)\|.$$

(iii) A/J is a C*-algebra in the quotient norm.

PROOF: (i): If $x \in J$, then $(x^*x)^{1/n} \in J$ for all n.

$$\|x^* - (x^*x)^{1/n}x^*\|^2 = \|x - x(x^*x)^{1/n}\|^2$$

$$= \|(1-(x^*x)^{1/n})x^*x(1-(x^*x)^{1/n})\| \to 0.$$

(ii): $\|\pi(x)\| = \inf\{\|x+y\| : y \in I\} \leq \inf \|x - xh_\lambda\|$. But if $y \in J$, then

$$\|x - xh_\lambda\| \leq \|(x+y)(1-h_\lambda)\| + \|y - yh_\lambda\| \leq \|x+y\| + \|y - yh_\lambda\|,$$

so $\limsup \|x - xh_\lambda\| \leq \|\pi(x)\|$.

(iii): For any $x \in A$,

$$\|\pi(x)\|^2 = \inf \|x(1-h_\lambda)\|^2 = \inf \|(1-h_\lambda)x^*x(1-h_\lambda)\|$$

$$\leq \inf \|x^*x(1-h_\lambda)\| = \|\pi(x^*x)\| = \|\pi(x)^*\pi(x)\|.$$



XX.16.3.15. COROLLARY. Let A and B be C*-algebras, $\phi : A \to B$ a *-homomorphism. Then $\phi(A)$ is closed in B, and therefore a C*-subalgebra which is isometrically isomorphic to A/J, where $J = Ker(\phi)$.

PROOF: J is closed since ϕ is continuous (XX.16.1.11.), and ϕ induces an injective *-homomorphism from the C*-algebra A/J to B, which is an isometry by XX.16.2.32.. 

XX.16.3.16. PROPOSITION. Let A be a C*-algebra.

(i) If I and J are closed ideals in A, then $I \cap J = IJ$, where IJ is the set of finite sums of elements of the form xy with $x \in I$, $y \in J$.

(ii) If I is a closed ideal in A and J is a closed ideal of I, then J is an ideal of A.

(iii) If I, J, K are closed ideals of A, then $(I+K) \cap (J+K) = (I \cap J) + K$.

PROOF: (i): Clearly $IJ \subseteq I \cap J$, and any product of two elements of $I \cap J$ is in IJ. But it is an easy consequence of functional calculus that any self-adjoint element of a C*-algebra is a product of two elements (in fact, any element of a C*-algebra is the product of two elements, so it is even unnecessary to allow finite sums in the definition of IJ).

(ii): I is itself a C*-algebra, hence has an approximate unit (h_λ). If $x \in A$ and $y \in J$, then $y \in I$, so $h_\lambda y \to y$, and thus $xh_\lambda y \to xy$. We have $xh_\lambda \in I$ and thus $(xh_\lambda)y \in J$ for all λ since J is an ideal of I, and so $xy \in J$ since J is closed. A similar argument shows that $yx \in J$.

(iii) follows immediately from (i) and the fact that

$$(I+K)(J+K) \subseteq IJ + K \subseteq (I \cap J) + K \subseteq (I+K) \cap (J+K)$$

(true in any ring). 👍

XX.16.3.17. A C*-algebra can have many nonclosed ideals, and their structure can be enormously complicated. (See, for example, [GJ76] for a description of the ideal structure of $C(X)$.) Several of the properties of closed ideals can fail for nonclosed ideals: they need not be self-adjoint, a self-adjoint ideal need not be hereditary or even positively generated, and an ideal of an ideal of A need not be an ideal of A. See [Bla06] for examples.

XX.16.4. Representations of Banach *-Algebras and C*-Algebras

Before about 1980, representation theory was the core of the theory of C*-algebras. It has since become less central, but it is still an extremely important part of the theory. Many applications of the theory involve representations. And there are even basic structural facts about C*-algebras which can only be reasonably obtained via representation theory.

XX.16.4.1. DEFINITION. A *representation* of a Banach *-algebra A is a *-homomorphism from A to $\mathcal{B}(\mathcal{H})$ for some Hilbert space $\mathcal{H}$. (Technically this should be called a *-representation; we will not consider representations which are not *-representations.) We will mostly only consider representations of C*-algebras.

Two representations π and ρ of A on $\mathcal{X}$ and $\mathcal{Y}$ respectively are *(unitarily) equivalent* if there is a unitary operator $U \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$ with $U\pi(x)U^* = \rho(x)$ for all $x \in A$.

A *subrepresentation* of a representation π on $\mathcal{H}$ is the restriction of π to a closed invariant subspace of $\mathcal{H}$ (a subspace $\mathcal{Y}$ of $\mathcal{H}$ is *invariant* if $\pi(x)\xi \in \mathcal{Y}$ for all $x \in A$, $\xi \in \mathcal{Y}$).

A representation is *irreducible* if it has no nontrivial closed invariant subspaces.

If π_i $(i \in \Lambda)$ is a representation of A on $\mathcal{H}_i$, then the *sum* $\oplus_i \pi_i$ of the π_i is the diagonal sum acting on $\oplus_i \mathcal{H}_i$. If each π_i is equivalent to a fixed representation ρ, then $\oplus_i \pi_i$ is a *multiple* or *amplification* of ρ by *card*(Λ).

XX.16.4.2. A representation is always norm-decreasing (XX.16.1.11.) and hence continuous. Thus the kernel of a representation is a closed ideal. A representation with kernel 0 is called *faithful*. By XX.16.2.32., a faithful representation of a C*-algebra is isometric.

XX.16.4.3. If J is a closed ideal of A, then any representation of A/J gives a representation of A by composition with the quotient map, whose kernel contains J. Conversely, if π is a representation of A whose kernel contains J, then π drops to a representation of A/J.

XX.16.4.4. PROPOSITION. If π is a representation of A on $\mathcal{H}$ and $\mathcal{X}$ is a closed subspace of $\mathcal{H}$, then the following are equivalent:

(i) $\mathcal{X}$ is invariant under $\pi(A)$.

(ii) $\mathcal{X}^\perp$ is invariant under $\pi(A)$.

(iii) $P_\mathcal{X} \in \pi(A)'$.

XX.16.4.5. If $\mathcal{N}$ is the largest subspace of $\mathcal{H}$ on which $\pi(x) = 0$ for all $x \in A$, then $\mathcal{N}$ is closed, and $\mathcal{X} = \mathcal{N}^\perp$ is an invariant subspace, called the *essential subspace* of π. $\mathcal{X}$ is the closed span of $\{\pi(x)\xi : x \in A,\ \xi \in \mathcal{H}\}$. If A is unital, then $P_\mathcal{X} = \pi(1_A)$. The representation π is *nondegenerate* (or *essential*) if $\mathcal{X} = \mathcal{H}$. An irreducible representation is always nondegenerate.

XX.16.4.6. PROPOSITION. Let π be a representation of a C*-algebra A on $\mathcal{H}$, and $\mathcal{X}$ the essential subspace of π. If (h_λ) is an approximate unit for A, then $\pi(h_\lambda) \to P_\mathcal{X}$ strongly.

PROOF: Fix $x \in A$ and $\zeta \in \mathcal{X}$. Since $h_\lambda x \to x$ in norm, $\pi(h_\lambda x) \to \pi(x)$ in norm, so $\pi(h_\lambda x)\zeta \to \pi(x)\zeta$. We have

$$\pi(h_\lambda)[\pi(x)\zeta] = \pi(h_\lambda x)\zeta \to \pi(x)\zeta.$$

Since vectors of the form $\pi(x)\zeta$ ($x \in A$, $\zeta \in \mathcal{X}$) are dense in $\mathcal{X}$, and (h_λ) is uniformly bounded, for any $\zeta \in \mathcal{X}$ we have $\pi(h_\lambda)\zeta \to \zeta$. Since $\pi(h_\lambda)\eta = 0$ for $\eta \in \mathcal{X}^\perp$, for any $\xi \in \mathcal{H}$ write $\xi = \zeta + \eta$ with $\zeta \in \mathcal{X}$, $\eta \in \mathcal{X}^\perp$; then

$$\pi(h_\lambda)\xi = \pi(h_\lambda)\zeta + \pi(h_\lambda)\eta = \pi(h_\lambda)\zeta \to \zeta = P_\mathcal{X}\xi\ .$$



XX.16.4.7. PROPOSITION. If J is a closed ideal in A, let $\mathcal{E}$ be the essential subspace of $\pi|_J$. Then $\mathcal{E}$ is invariant under $\pi(A)$, so $P_\mathcal{E} \in \pi(A)'$.

PROOF: Fix an approximate unit (h_λ) for J. If $\xi \in \mathcal{E}$, then $\pi(h_\lambda)\xi \to \xi$ (XX.16.4.6.). If $x \in A$, then $\pi(xh_\lambda)\xi \in \mathcal{E}$ since $xh_\lambda \in J$, and $\pi(xh_\lambda)\xi = \pi(x)[\pi(h_\lambda)\xi] \to \pi(x)\xi$ since $\pi(x)$ is continuous; thus $\pi(x)\xi \in \mathcal{E}$ since $\mathcal{E}$ is closed. 

XX.16.4.8. If J is a closed ideal in A, let $\mathcal{E}$ be the essential subspace of $\pi|_J$. Then $\mathcal{E}$ is invariant under $\pi(A)$, so π decomposes into a sum of a representation essential on J and a representation which is zero on J (a representation of A/J). Conversely, if ρ is a nondegenerate representation of J on an $\mathcal{H}$, and (h_λ) is an approximate unit for J, then for any $a \in A$ the net $(\rho(ah_\lambda))$ converges strongly in $\mathcal{B}(\mathcal{H})$ to an operator we call $\rho_A(a)$ (cf. **??**), defining a representation of A extending ρ, and we have $\rho_A(A)'' = \rho(J)''$. (This ρ_A is the unique extension of ρ to a representation of A on $\mathcal{H}$.)

XX.16.4.9. Note that for a nondegenerate representation π, the strong (or weak) closure of $\pi(A)$ is $\pi(A)''$ by the Bicommutant Theorem (XX.5.5.2.); for a general π, the strong closure of $\pi(A)$ is $P_{\mathcal{X}}\pi(A)''$, where $\mathcal{X}$ is the essential subspace of π. (Note that $P_{\mathcal{X}}$ is a central projection in $\pi(A)''$.)

We get the following fact from an application of the Bicommutant Theorem:

XX.16.4.10. PROPOSITION. Let π be a representation of a C*-algebra A on a Hilbert space $\mathcal{H}$. Then the following are equivalent:

(i) π is irreducible.

(ii) $\pi(A)' = \mathbb{C}I$.

(iii) $\pi(A)'' = \mathcal{B}(\mathcal{H})$.

(iv) $\pi(A)$ is strongly dense in $\mathcal{B}(\mathcal{H})$.

It follows from XX.16.4.8. that if π is an irreducible representation of A, and J is a closed ideal of A, then $\pi|_J$ is either zero or irreducible.

XX.16.4.11. PROPOSITION. The kernel of an irreducible representation of a C*-algebra is a (closed) prime ideal.

PROOF: The kernel of any representation is closed. Let π be an irreducible representation of a C*-algebra A on $\mathcal{H}$. Replacing A by $A/\ker(\pi)$ we may assume π is faithful. Suppose there are nonzero ideals J and K of A with $J \cap K = \{0\}$. If (h_λ) and (k_μ) are increasing approximate units for J and K respectively, then $(\pi(h_\lambda))$ and $(\pi(k_\mu))$ converge strongly to the projections P and Q onto the closures of $\pi(J)\mathcal{H}$ and $\pi(K)\mathcal{H}$. Then P and Q are nonzero projections in $\mathcal{B}(\mathcal{H})$, and $P \perp Q$. Since $\pi(J)\mathcal{H}$ and $\pi(K)\mathcal{H}$ are invariant under $\pi(A)$, we have $P, Q \in \pi(A)'$, a contradiction.

XX.16.4.12. If π is a representation of a C*-algebra A on a Hilbert space $\mathcal{H}$, and $0 \neq \xi \in \mathcal{H}$, then the set $\{\pi(x)\xi : x \in A\}$ is an invariant subspace of $\mathcal{H}$, and thus its closure is also invariant. If the closure is all of $\mathcal{H}$, ξ is called a *cyclic vector* for π. A representation of A with a cyclic vector is called a *cyclic representation* of A. A cyclic representation is nondegenerate. A simple Zorn's Lemma argument shows that every nondegenerate representation is a direct sum of cyclic representations (in a highly nonunique way in general). If π is an irreducible representation of A on $\mathcal{H}$, then every nonzero vector in $\mathcal{H}$ is a cyclic vector for π.

Positive Linear Functionals and States

The connection between C*-algebras and Hilbert spaces is made via the notion of a state:

XX.16.4.13. DEFINITION. Let A be a C*-algebra. A linear functional ϕ on A is *positive*, written $\phi \geq 0$, if $\phi(x) \geq 0$ whenever $x \geq 0$. A *state* on A is a positive linear functional of norm 1. Denote by $\mathcal{S}(A)$ the set of all states on A, called the *state space* of A.

XX.16.4.14. PROPOSITION. A positive linear functional is bounded.

PROOF: Let ϕ be a positive linear functional on a C*-algebra A. It suffices to show ϕ is bounded on A_+ by XX.16.3.2.(vi); if it is not, for each n choose $x_n \geq 0$ with $\|x_n\| = 1$ and $\phi(x_n) \geq 4^n$, and then $x = \sum_{n=1}^{\infty} 2^{-n} x_n$ satisfies $\phi(x) \geq 2^{-n}\phi(x_n) \geq 2^n$ for all n, a contradiction. 👍

XX.16.4.15. EXAMPLES.

(i) If A is a concrete C*-algebra of operators acting nondegenerately on $\mathcal{H}$ and $\xi \in \mathcal{H}$, and $\phi_\xi(x) = \langle x\xi, \xi\rangle$ for $x \in A$, then ϕ_ξ is a positive linear functional on A of norm $\|\xi\|^2$, so ϕ_ξ is a state if $\|\xi\| = 1$. Such a state is called a *vector state* of A. This example is the origin of the term "state": in the mathematical formulation of quantum mechanics, the states of a physical system are given by probability distributions (unit vectors in an L^2-space), and observables are self-adjoint operators; the value of the observable T on the state ξ is $\langle T\xi, \xi\rangle$ (cf. [BR87]).

(ii) By the Riesz Representation Theorem, there is a one-one correspondence between bounded linear functionals on $C_0(X)$ and finite regular complex Baire measures (complex Radon measures) on X. A bounded linear functional is positive if and only if the corresponding complex measure takes only nonnegative real values (i.e. is an ordinary measure). Thus the states on $C_0(X)$ are precisely given by the regular Baire probability measures on X. Any nonzero homomorphism from $C_0(X)$ to $\mathbb{C}$ is a state by XX.16.2.14., XX.16.2.15., and XX.16.3.2.(ii) and (vii).

A positive linear functional takes real values on self-adjoint elements, and thus $\phi(x^*) = \overline{\phi(x)}$ for any x if $\phi \geq 0$. So if ϕ is a positive linear functional on A, then ϕ defines a pre-inner product on A by $\langle x, y\rangle_\phi = \phi(y^*x)$. Thus we get the CBS inequality:

XX.16.4.16. PROPOSITION. Let A be a C*-algebra, ϕ a positive linear functional on A, x, $y \in A$. Then $|\phi(y^*x)|^2 \leq \phi(x^*x)\phi(y^*y)$. More symmetrically, $|\phi(xy)|^2 \leq \phi(xx^*)\phi(y^*y)$.

XX.16.4.17. PROPOSITION. Let A be a C*-algebra and ϕ a bounded linear functional on A. Then

(i) If $\phi \geq 0$, then $\|\phi\| = \sup\{\phi(x) : x \geq 0,\ \|x\| = 1\} = \lim \phi(h_\lambda)$ for any approximate unit (h_λ) for A (in particular, $\|\phi\| = \phi(1_A)$ if A is unital), and ϕ extends to a positive linear functional on $A^\dagger$ by setting $\phi(1) = \|\phi\|$ (or $\phi(1) = t$ for any $t \geq \|\phi\|$). In particular, any state on A extends uniquely to a state on $A^\dagger$.

(ii) If A is unital and $\phi(1_A) = \|\phi\|$, then $\phi \geq 0$.

PROOF: For the first part of (i), if (h_λ) is an approximate unit for A, we have

$$\|\phi\| \geq \sup\{\phi(x) : x \geq 0,\ \|x\| = 1\} \geq \sup \phi(h_\lambda)\ ;$$

but if $y \in A$, $\|y\| = 1$, $|\phi(y)|^2 > \|\phi\|^2 - \epsilon$, then, using XX.16.4.16.,

$$\|\phi\|^2 - \epsilon < |\phi(y)|^2 = \lim |\phi(h_\lambda^{1/2} y)|^2 \leq \liminf \phi(h_\lambda)\phi(y^* y) \leq \|\phi\| \liminf \phi(h_\lambda).$$

The second statement of (i) follows from the first and XX.16.3.2.(viii).

To prove (ii), if $x \geq 0$, restrict ϕ to $C^*(x, 1) \cong C(X)$; then ϕ corresponds to a complex measure μ on X, and $\mu(X) = |\mu|(X)$, so $\mu \geq 0$. A more elementary alternate argument can be based on XX.16.1.10.: if $\phi(1_A) = \|\phi\| = 1$, and $x \geq 0$, suppose $\phi(x) \notin [0, \infty)$. Then there is a $\lambda \in \mathbb{C}$ and $\rho \geq 0$ with $\sigma_A(x) \subseteq D = \{z \in \mathbb{C} : |z - \lambda| \leq \rho\}$, but $\phi(x) \notin D$. If $y = x - \lambda 1_A$, then y is normal and $\|y\| = r_\sigma(y) \leq \rho$, so $|\phi(x) - \lambda| = |\phi(y)| \leq \|y\| \leq \rho$, a contradiction. 👍

XX.16.4.18. Any sum or nonnegative scalar multiple of positive linear functionals on a C*-algebra A is positive. A weak-* limit of a net of positive linear functionals on A is positive. XX.16.4.17.(i) shows that if ϕ and ψ are positive linear functionals on A, then $\|\phi + \psi\| = \|\phi\| + \|\psi\|$. In particular, a convex combination of states is a state, so if A is unital, $\mathcal{S}(A)$ is a compact convex subset of the dual of A. If A is nonunital, then $\mathcal{S}(A)$, though convex, is not compact or even locally compact, but the set of positive linear functionals on A of norm ≤ 1 is in natural one-one correspondence with $\mathcal{S}(A^\dagger)$, which is a compact convex set.

XX.16.4.19. Definition. A state of a C*-algebra A is *pure* if it is an extreme point of $\mathcal{S}(A)$. Denote the set of pure states of A by $\mathcal{P}(A)$.

XX.16.4.20. If A is unital, then $\mathcal{S}(A)$ is the closed convex hull of the pure states; in general, the set of positive linear functionals of norm ≤ 1 is the closed convex hull of the pure states and the zero functional. The pure states of $C_0(X)$ are precisely the homomorphisms to $\mathbb{C}$, i.e. the states corresponding to point masses (XX.16.4.15.(ii)).

Extension and Existence of States

XX.16.4.21. Proposition. If A is a C*-algebra and B a C*-subalgebra, then every state on B extends to a state on A.

Proof: If A is a C*-algebra and ϕ is a state on A, then by XX.16.4.17.(i) ϕ extends to a state on $A^\dagger$. If A is a unital C*-algebra and B a unital C*-subalgebra, and ϕ is a state on B, by the Hahn-Banach Theorem () ϕ extends to a linear functional ψ on A of norm 1. Since $\psi(1) = \phi(1) = 1$, ψ is a state on A. If B is a (not necessarily unital) C*-subalgebra of a (not necessarily unital) C*-algebra A, and ϕ is a state on B, first extend ϕ to $B^\dagger$, regard $B^\dagger$ as a unital C*-subalgebra of $A^\dagger$ and extend to $A^\dagger$, and then restrict to A to get an extension of ϕ to a state on A. 👍

XX.16.4.22. If ϕ is a pure state on a C*-subalgebra of a (not necessarily unital) C*-algebra A, then the set of extensions of ϕ to A is a weak-* compact convex subset of $\mathcal{S}(A)$, and any extreme point of this set (extreme points exist by the Krein-Milman Theorem ()) is a pure state of A. Thus we get:

XX.16.4.23. PROPOSITION. Let A be a C*-algebra, $x \in A_{sa}$. Then there is a pure state ϕ on A with $|\phi(x)| = \|x\|$.

PROOF: Note that there is a pure state (homomorphism) ϕ on $C^*(x) \cong C_0(\sigma_A(x))$ with $|\phi(x)| = r_\sigma(x) = \|x\|$, and extend ϕ to a pure state on A. ☝

There is also a noncommutative version of the Jordan Decomposition Theorem for signed measures:

XX.16.4.24. THEOREM. Let A be a C*-algebra, and ϕ a bounded self-adjoint linear functional on A. Then there are unique positive linear functionals ϕ_+ and ϕ_- on A with $\phi = \phi_+ - \phi_-$ and $\|\phi\| = \|\phi_+\| + \|\phi_-\|$.

Since every linear functional ϕ can be canonically written as $\phi_{re} + i\phi_{im}$ with ϕ_{re}, ϕ_{im} self-adjoint [define ϕ^* by $\phi^*(x) = \overline{\phi(x^*)}$, and let $\phi_{re} = (\phi + \phi^*)/2$, $\phi_{im} = (\phi - \phi^*)/2i$], every bounded linear functional can be written canonically as a linear combination of four states. A simple consequence of this and the Hahn-Banach Theorem is:

XX.16.4.25. COROLLARY. Let A be a C*-algebra, $x \in A$.

(i) If $\phi(x) = 0$ for all $\phi \in \mathcal{S}(A)$, then $x = 0$.

(ii) $x = x^*$ if and only if $\phi(x) \in \mathbb{R}$ for all $\phi \in \mathcal{S}(A)$.

(iii) $x \geq 0$ if and only if $\phi(x) \geq 0$ for all $\phi \in \mathcal{S}(A)$.

In fact, $\mathcal{S}(A)$ can be replaced by $\mathcal{P}(A)$ in this result.

The GNS Construction

The GNS construction, discovered independently by GELFAND and NAIMARK [GN43] and I. SEGAL [Seg47] (the essential idea appeared earlier in [Gel41]), while quite simple, is an "ingenious construction" (E. HILLE [Hil43]) and one of the most fundamental ideas of the theory of operator algebras. It provides a method for manufacturing representations of C*-algebras.

XX.16.4.26. Let A be a C*-algebra, ϕ a positive linear functional on A. Put a pre-inner product on A by $\langle x, y\rangle_\phi = \phi(y^*x)$. If

$$N_\phi = \{x \in A : \phi(x^*x) = 0\},$$

then N_ϕ is a closed left ideal of A, and $\langle \cdot\,, \cdot \rangle_\phi$ drops to an inner product on A/N_ϕ. If $a \in A$, let $\pi_\phi(a)$ be the left multiplication operator by a on A/N_ϕ, i.e. $\pi_\phi(a)(x + N_\phi) = ax + N_\phi$. Since

$$x^*a^*ax \leq \|a\|^2 x^*x$$

(XX.16.3.8.), $\pi_\phi(a)$ is a bounded operator and $\|\pi_\phi(a)\| \leq \|a\|$; so $\pi_\phi(a)$ extends to a bounded operator, also denoted $\pi_\phi(a)$, on the completion (Hilbert space) $\mathcal{H}_\phi = L^2(A, \phi)$ of A/N_ϕ. The representation π_ϕ is called the *GNS representation* of A associated with ϕ.

XX.16.4.27. There is also a (unique) distinguished vector $\xi_\phi \in \mathcal{H}_\phi$ such that $\phi(a) = \langle \pi_\phi(a)\xi_\phi, \xi_\phi\rangle_\phi$ (and $\|\xi_\phi\|^2 = \|\phi\|$). If A is unital, then ξ_ϕ is just the image of 1_A in $\mathcal{H}_\phi$. In general, let $\tilde{\phi}$ be the unique extension of ϕ to $A^\dagger$ with $\|\tilde{\phi}\| = \|\phi\|$; then from $\tilde{\phi}(1) = \lim \phi(h_\lambda) = \lim \phi(h_\lambda^2)$, where (h_λ) is an approximate unit for A, it follows that $\tilde{\phi}((1-h_\lambda)^2) \to 0$, so A/N_ϕ is dense in $A^\dagger/N_{\tilde{\phi}}$ and thus $\mathcal{H}_\phi$ can be identified with $\mathcal{H}_{\tilde{\phi}}$, and ξ_ϕ may be taken to be $\xi_{\tilde{\phi}}$. This ξ_ϕ is a cyclic vector for π_ϕ, i.e. $\pi_\phi(A)\xi_\phi$ is dense in $\mathcal{H}_\phi$.

XX.16.4.28. This construction is the inverse of XX.16.4.15.(i): if π is a representation of A on a Hilbert space $\mathcal{H}$ and ξ is a cyclic vector for π, and $\phi(a) = \langle \pi(a)\xi, \xi\rangle$, then $(\mathcal{H}_\phi, \pi_\phi, \xi_\phi)$ is unitarily equivalent to $(\mathcal{H}, \pi, \xi)$. Thus there is a precise one-one correspondence between positive linear functionals on A and (cyclic) representations of A with a specified cyclic vector.

XX.16.4.29. Example. Let $A = n$, with Tr the (ordinary) trace on A. Then $\mathcal{H}_{\mathrm{Tr}}$ can be identified with $\mathbb{M}_n$ with the Hilbert-Schmidt norm, with $\xi_{\mathrm{Tr}} = I$. The action π_{Tr} is the ordinary one by left multiplication. $\mathbb{M}_n$ acting on itself in this way is called the *standard form* representation of $\mathbb{M}_n$. Its commutant is also isomorphic (or, more naturally, anti-isomorphic) to $\mathbb{M}_n$, acting by right multiplication.

XX.16.4.30. If ϕ is a positive linear functional on A, then ϕ has a canonical extension to a positive linear functional on $\mathcal{B}(\mathcal{H}_\phi)$ of the same norm, also denoted ϕ, by $\phi(T) = \langle T\xi_\phi, \xi_\phi\rangle_\phi$.

The next simple result is a version of the Radon-Nikodým Theorem (the classical Radon-Nikodým theorem is essentially the special case where $A = L^\infty(X, \mu)$ for a finite measure space (X, μ), and $\phi(f) = \int f\, d\mu$).

XX.16.4.31. Proposition. Let ϕ and ψ be positive linear functionals on A, with $\psi \leq \phi$ (i.e. $\psi(x) \leq \phi(x)$ for all $x \in A_+$). Then there is a unique operator $T \in \pi_\phi(A)' \subseteq \mathcal{B}(\mathcal{H}_\phi)$, with $0 \leq T \leq I$, such that $\psi(x) = \phi(T\pi_\phi(x)) = \langle T\pi_\phi(x)\xi_\phi, \xi_\phi\rangle_\phi$ for all $x \in A$.

Proof: For $x, y \in A$, define $(\pi_\phi(x)\xi_\phi, \pi_\phi(y)\xi_\phi) = \psi(y^*x)$. Then $(\cdot\,,\cdot)$ extends to a bounded sesquilinear form (pre-inner product) on $\mathcal{H}_\phi$, and thus there is a $T \in (\mathcal{H}_\phi)$, $0 \leq T \leq I$, such that $(\eta, \zeta) = \langle T\eta, \zeta\rangle_\phi$ for all $\eta, \zeta \in \mathcal{H}_\phi$ (**??**). If $x, y, z \in A$, then

$$\langle T\pi_\phi(x)[\pi_\phi(z)\xi_\phi], \pi_\phi(y)\xi_\phi\rangle_\phi = \psi(y^*(xz)) = \psi((x^*y)^*z)$$

$$= \langle T\pi_\phi(z)\xi_\phi, \pi_\phi(x^*)[\pi_\phi(y)\xi_\phi]\rangle_\phi = \langle \pi_\phi(x)T[\pi_\phi(z)\xi_\phi], \pi_\phi(y)\xi_\phi\rangle_\phi$$

so, fixing x and letting y, z range over A, we conclude that $T\pi_\phi(x) = \pi_\phi(x)T$. 👍

XX.16.4.32. In fact, if $\psi \leq \phi$ and $\omega = \phi - \psi$, so that $\phi = \psi + \omega$, then π_ϕ is unitarily equivalent to the subrepresentation of $\pi_\psi \oplus \pi_\omega$ with cyclic vector $\xi_\psi \oplus \xi_\omega$, and T is projection onto the first coordinate, compressed to this subspace.

Because of the abundance of states on a C*-algebra described in (XX.16.4.23.), every C*-algebra has many representations. We can say even more:

XX.16.4.33. PROPOSITION. If ϕ is a state of A, then π_ϕ is irreducible if and only if ϕ is a pure state.

PROOF: It follows immediately from XX.16.4.31. and XX.16.4.10. that if π_ϕ is irreducible, then any positive linear functional $\psi \leq \phi$ is a multiple of ϕ, and hence ϕ is pure. Conversely, if ϕ is pure, suppose there is a projection P in $\pi_\phi(A)'$, $P \neq 0, I$. Then $P\xi_\phi \neq 0$, for otherwise

$$0 = \pi_\phi(x)P\xi_\phi = P[\pi_\phi(x)\xi_\phi]$$

for all $x \in A$, so $P = 0$. Similarly, $(I-P)\xi_\phi \neq 0$. For $x \in A$, write

$$\phi_1(x) = \langle \pi_\phi(x)P\xi_\phi, P\xi_\phi\rangle_\phi = \phi(P\pi_\phi(x))$$

$$\phi_2(x) = \langle \pi_\phi(x)(I-P)\xi_\phi, (I-P)\xi_\phi\rangle_\phi = \phi((I-P)\pi_\phi(x)).$$

Then ϕ_1, ϕ_2 are positive linear functionals on A, $\phi = \phi_1 + \phi_2$, and $\phi_1 = \lambda\phi$, $\phi_2 = (1-\lambda)\phi$ for some λ, $0 < \lambda < 1$, because ϕ is pure. For any $\epsilon > 0$, there is an $x \in A$ with $\|\pi_\phi(x)\xi_\phi - P\xi_\phi\|_\phi < \epsilon$ and $\|\pi_\phi(x)\xi_\phi\|_\phi^2 = \phi(x^*x) = \|P\xi_\phi\|_\phi^2$. Then $\|(I-P)\pi_\phi(x)\xi_\phi\|_\phi = \|(I-P)(\pi_\phi(x)\xi_\phi - P\xi_\phi)\|_\phi < \epsilon$, and

$$(1-\lambda)\|P\xi_\phi\|_\phi^2 = (1-\lambda)\phi(x^*x) = \phi_2(x^*x) = \|(I-P)\pi_\phi(x)\xi_\phi\|_\phi^2 < \epsilon^2$$

contradicting $\lambda \neq 1$ and therefore the existence of P. Thus $\pi_\phi(A)' = \mathbb{C}I$. 👍

XX.16.4.34. COROLLARY. If A is a C*-algebra and $x \in A$, then there is an irreducible representation π of A with $\|\pi(x)\| = \|x\|$.

PROOF: Let $\pi = \pi_\phi$, where ϕ is a pure state of A with $\phi((xx^*)^2) = \|(xx^*)^2\| = \|x\|^4$ (XX.16.4.23.). Then

$$\|x\|^2 = \langle xx^*, xx^*\rangle_\phi^{1/2} = \|xx^*\|_\phi = \|\pi(x)x^*\|_\phi \leq \|\pi(x)\|\|x^*\|_\phi \leq \|\pi(x)\|\|x\| \ .$$

👍

By considering sums of irreducible representations, we obtain

XX.16.4.35. COROLLARY. [GELFAND-NAIMARK THEOREM] If A is a C*-algebra, then A has a faithful representation, i.e. A is isometrically isomorphic to a concrete C*-algebra of operators on a Hilbert space $\mathcal{H}$. If A is separable, then $\mathcal{H}$ may be chosen to be separable.

XX.16.4.36. PROPOSITION. Let A be a C*-algebra, B a C*-subalgebra of A, and ρ a representation of B on a Hilbert space $\mathcal{H}_0$. Then ρ can be extended to a representation π of A on a (possibly) larger Hilbert space $\mathcal{H}$, i.e. there is a Hilbert space $\mathcal{H}$ containing $\mathcal{H}_0$ and a representation π of A on $\mathcal{H}$ such that $\mathcal{H}_0$ is invariant under $\pi|_B$ and $\pi|_B$ agrees with ρ on $\mathcal{H}_0$. If ρ is irreducible, then π may be chosen irreducible.

PROOF: Since every representation is a direct sum of cyclic representations, it suffices to show the result for ρ cyclic, i.e. ρ is the GNS representation corresponding to a state ψ of B. Extend ψ to a state ϕ of A (with ϕ pure if ψ is pure; see XX.16.4.22.), and let π be its GNS representation. 👍

XX.16.4.37. Thus every C*-algebra can be faithfully represented on a Hilbert space, but the representation is highly nonunique in general. The major part of representation theory is concerned with how to describe the possibilities. There is a relatively small class of C*-algebras () for which a concrete description of the representations is even possible.

Using the GNS construction, we can obtain an extension of XX.16.4.17.(i) and a sharpening of XX.16.4.21. in the case that the subalgebra is an ideal. This result is also closely related to XX.16.4.8..

XX.16.4.38. PROPOSITION. Let J be a closed ideal in a C*-algebra A, and ψ a positive linear functional on J. Then there is a unique extension of ψ to a positive linear functional ϕ on A with $\|\phi\| = \|\psi\|$. If (h_λ) is an increasing approximate unit for J, then $\phi(x) = \lim \psi(h_\lambda x) = \lim \psi(x h_\lambda)$ for all $x \in A$; if $x \in A_+$, then

$$\phi(x) = \lim \psi(x^{1/2} h_\lambda x^{1/2}) = \sup \psi(x^{1/2} h_\lambda x^{1/2}).$$

Any other positive linear functional ϕ' on A extending ψ satisfies $\phi' \geq \phi$.

PROOF: By XX.16.4.21. there is a ϕ extending ψ with $\|\phi\| = \|\psi\|$, but we will not use this result. Instead we define ϕ directly as follows.

Let π_ψ be the GNS representation of J on $\mathcal{H}_\psi$ with cyclic vector ξ_ψ. Extend π_ψ to a representation π of A on $\mathcal{H}_\psi$ as in XX.16.4.8.. Then ξ_ψ defines a positive linear functional ϕ on A by $\phi(x) = \langle \pi(x)\xi_\psi, \xi_\psi \rangle$. We have $\|\phi\| = \|\xi_\psi\|^2 = \|\psi\|$, and $(\mathcal{H}_\phi, \pi_\phi, \xi_\phi)$ can be naturally identified with $(\mathcal{H}_\psi, \pi, \xi_\psi)$. Since $\pi(h_\lambda) \to I$ strongly, we have, for any $x \in A$,

$$\psi(h_\lambda x) = \langle \pi(h_\lambda)\pi(x)\xi_\psi, \xi_\psi \rangle \to \langle \pi(x)\xi_\psi, \xi_\psi \rangle = \phi(x)$$

and similarly $\psi(x h_\lambda) \to \phi(x)$ and, if $x \geq 0$, $\psi(x^{1/2} h_\lambda x^{1/2}) \to \phi(x)$.

If ϕ' is any extension of ψ to a positive linear functional on A, and $x \in A_+$, then $\phi'(x) \geq \phi'(x^{1/2} h_\lambda x^{1/2}) = \psi(x^{1/2} h_\lambda x^{1/2})$ for all λ, so $\phi'(x) \geq \phi(x)$. Thus $\omega = \phi' - \phi$ is a positive linear functional, and we have $\|\phi'\| = \|\phi\| + \|\omega\| = \|\psi\| + \|\omega\|$, so if $\|\phi'\| = \|\psi\|$, then $\omega = 0$, $\phi' = \phi$. 👍

Primitive Ideal Space and Spectrum

It is common in mathematics (e.g. in algebraic geometry) to associate to a ring A a topological space of ideals of A, generally either the set of prime ideals (denoted $\mathrm{Spec}(A)$) or the set of maximal ideals (denoted $\mathrm{Maxspec}(A)$), with the hull-kernel topology (XIII.7.10.20.). For C*-algebras, the most useful variant is to consider the primitive ideal space:

XX.16.4.39. DEFINITION. Let A be a C*-algebra. A *primitive ideal* of A is an ideal which is the kernel of an irreducible representation of A. Denote by $\mathrm{Prim}(A)$ the set of primitive ideals of A. The topology on $\mathrm{Prim}(A)$ is the hull-kernel topology, i.e. $\{J_i\}^- = \{J : J \supseteq \cap J_i\}$.

XX.16.4.40. A primitive ideal is closed by XX.16.4.2. and prime by XX.16.4.11.. By XX.16.4.34., the intersection of the primitive ideals of a C*-algebra A is $\{0\}$. By considering irreducible representations of A/J, it follows that every closed ideal J of A is an intersection of primitive ideals. In particular, every maximal ideal is primitive. There is thus a natural one-one correspondence between the closed sets in $\mathrm{Prim}(A)$ and the closed ideals of A. If A is commutative, then every closed prime ideal is maximal and hence primitive, and $\mathrm{Prim}(A)$ agrees with the set $\mathrm{Maxspec}(A)$ and may be identified with $\hat{A}$ as in XX.16.2.1..

XX.16.4.41. If J is a closed ideal in A, then we can identify $\mathrm{Prim}(A/J)$ with

$$\mathrm{Prim}_J(A) = \{K \in \mathrm{Prim}(A) : J \subseteq K\},$$

the corresponding closed set in $\mathrm{Prim}(A)$. If $K \in \mathrm{Prim}(A)$, $J \not\subseteq K$, and π is an irreducible representation of A with kernel K, then $\pi|J$ is an irreducible representation of J (XX.16.4.8.) with kernel $K \cap J$. So there is a map ρ_J from

$$\mathrm{Prim}^J(A) = \{K \in \mathrm{Prim}(A) : J \not\subseteq K\}$$

to $\mathrm{Prim}(J)$, defined by $\rho_J(K) = K \cap J$, which is surjective by XX.16.4.8.. Also, ρ_J is injective: if $K_1, K_2 \in \mathrm{Prim}(A)$ and $K_1 \cap J = K_2 \cap J \neq J$, then by XX.16.3.16.(iii) we have

$$(K_1 + K_2) \cap (J + K_2) = (K_1 \cap J) + K_2 = (K_2 \cap J) + K_2 = K_2$$

and, since K_2 is prime and $J + K_2 \neq K_2$, we have $K_1 + K_2 = K_2$, $K_1 \subseteq K_2$. Symmetrically, $K_2 \subseteq K_1$, so $K_1 = K_2$. Thus $\mathrm{Prim}(J)$ may be identified with $\mathrm{Prim}^J(A) = \mathrm{Prim}(A) \setminus \mathrm{Prim}_J(A)$, an open set in $\mathrm{Prim}(A)$.

XX.16.4.42. If J is a closed ideal in a C*-algebra A and $x \in A$, write $\|x\|_J$ for the norm of x mod J. If $\{J_i\}$ is any collection of closed ideals in A and $J = \cap J_i$, then $\|\cdot\|_J = \sup_i \|\cdot\|_{J_i}$ (by XX.16.2.33. it suffices to note that this formula defines a C*-norm on A/J); and if (J_i) is an increasing net of closed ideals with $J = [\cup J_i]^-$, then $\|\cdot\|_J = \inf_i \|\cdot\|_{J_i}$ (this follows immediately from the definition of the quotient norm). A simple consequence is:

XX.16.4.43. Proposition. Let A be a C*-algebra.

(i) If $x \in A$, define $\check{x} : \mathrm{Prim}(A) \to \mathbb{R}_+$ by $\check{x}(J) = \|x\|_J$. Then $\check{x}$ is lower semicontinuous.

(ii) If $\{x_i\}$ is a dense set in the unit ball of A, and $U_i = \{J \in \mathrm{Prim}(A) : \check{x}_i(J) > 1/2\}$, then $\{U_i\}$ forms a base for the topology of $\mathrm{Prim}(A)$.

(iii) If $x \in A$ and $\lambda > 0$, then $\{J \in \mathrm{Prim}(A) : \check{x}(J) \geq \lambda\}$ is compact (but not necessarily closed) in $\mathrm{Prim}(A)$.

Proof: Part (i) is obvious. For (ii), let $J \in \mathrm{Prim}(A)$. A neighborhood of J is a set of the form $V = \{I \in \mathrm{Prim}(A) : K \not\subseteq I\}$ for a closed ideal K not contained in J. Choose $y \in K \setminus J$, and $0 < \epsilon \leq \frac{1}{2}\|y\|_J^2$, and let $x = f_\epsilon(y^*y)$. Then $x \in K$, and $\|x\| = \|x\|_J = 1$. Choose x_i with $\|x_i - x\| < 1/2$; then $J \in U_i$ and $U_i \subseteq V$. To show (iii), let $\{K_i : i \in \Omega\}$ be a set of closed ideals in A, K the closed ideal generated by $\cup K_i$ (so $K = [\sum K_i]^-$), and

$$U_i = \{J \in \mathrm{Prim}(A) : K_i \not\subseteq J\}.$$

Then $\{U_i\}$ is an open cover of $F = \{J \in \mathrm{Prim}(A) : \check{x}(J) \geq \lambda\}$ if and only if $\|x\|_K < \lambda$. In this case, there are $i_1, \ldots, i_n$ such that if $K_0 = K_{i_1} + \cdots + K_{i_n}$, then $\|x\|_{K_0} < \lambda$, so $\{U_{i_1}, \ldots, U_{i_n}\}$ is a cover of F. ☝

XX.16.4.44. Corollary. $\mathrm{Prim}(A)$ is a locally compact space T_0-space which is compact if A is unital. If A is separable, then $\mathrm{Prim}(A)$ is second countable.

Actually, there are several inequivalent definitions of local compactness used for non-Hausdorff spaces (XIII.11.8.20.), so the statement that $\mathrm{Prim}(A)$ is "locally compact" is ambiguous in the non-Hausdorff case. To be precise, if A is a C*-algebra, then XX.16.4.43. shows that every point of $\mathrm{Prim}(A)$ has a neighborhood base of compact sets, i.e. $\mathrm{Prim}(A)$ satisfies XIII.11.8.20.(ii). But a point need not have any closed compact neighborhood, i.e. $\mathrm{Prim}(A)$ does not necessarily satisfy XIII.11.8.20.(iii).

XX.16.4.45. The function $\check{x}$ of XX.16.4.43. is not continuous in general; in fact, if $\mathrm{Prim}(A)$ is not Hausdorff, then the $\check{x}$ cannot all be continuous since $\{\check{x} : x \in A\}$ separates the points of $\mathrm{Prim}(A)$. But if $\mathrm{Prim}(A)$ is Hausdorff, XX.16.4.43.(iii) shows that $\check{x}$ is continuous for all $x \in A$. If A is commutative, or more generally if x is in the center of A, then $\check{x}$ is continuous.

XX.16.4.46. Examples.

(i) Let $A = \mathbb{K}^\dagger$. Then $\mathrm{Prim}(A) = \{0, \mathbb{K}\}$; $\{\mathbb{K}\}$ is a closed point, and $\{0\}$ is a dense open point. Thus $\mathrm{Prim}(A)$ homeomorphic to the Sierpiński space and is not T_1.

(ii) Let A be the C*-algebra of all sequences of elements of $\mathbb{M}_2$ which converge to a diagonal matrix. Then $\mathrm{Prim}(A)$ consists of a sequence of points simultaneously converging to two closed points at infinity. Thus $\mathrm{Prim}(A)$ is T_1 but not Hausdorff.

(iii) Let A be the C*-algebra of all sequences in $\mathbb{K}^\dagger$ converging to a scalar multiple of 1. Then $\mathrm{Prim}(A)$ is the one-point compactification of a disjoint union of two-point spaces as in (i).

If A is commutative, then Gelfand's Theorem (XX.16.2.1.) identifies elements of A with continuous functions on $\hat{A} \cong \mathrm{Prim}(A)$.

XX.16.4.47. For a general C*-algebra A we define $\hat{A}$ to be the set of unitary equivalence classes of irreducible representations of A. There is a natural map from $\hat{A}$ onto $\mathrm{Prim}(A)$ sending π to $ker(\pi)$. This map is not injective in general, i.e. there can be inequivalent representations with the same kernel. There is also an obvious map from $\mathcal{P}(A)$ onto $\hat{A}$, sending a pure state to its GNS representation; the composite map from $\mathcal{P}(A)$ to $\mathrm{Prim}(A)$ is continuous and open when $\mathcal{P}(A)$ is given the relative weak-* topology. The quotient topology (XIII.8.1.1.) on $\hat{A}$ from $\mathcal{P}(A)$, which coincides with the weak topology (XIII.5.1.8.) for the map of $\hat{A}$ onto $\mathrm{Prim}(A)$, is the unique topology on $\hat{A}$ making the maps $\mathcal{P}(A) \to \hat{A}$ and $\hat{A} \to \mathrm{Prim}(A)$ continuous and open (XIII.8.5.1.), but $\hat{A}$ is not T_0 unless the map to $\mathrm{Prim}(A)$ is injective (and hence a homeomorphism).

By XX.16.4.8., if J is a closed ideal in A, we may identify $\hat{J}$ and $\mathrm{Prim}(J)$ with open subsets of $\hat{A}$ and $\mathrm{Prim}(A)$ respectively; the complements are naturally identified with $\widehat{A/J}$ and $\mathrm{Prim}(A/J)$. We may do the same with $\mathcal{P}(J)$ since if $\phi \in \mathcal{P}(A)$, then either $\phi|_J = 0$ or $\phi|J$ is a pure state of J.

XX.16.4.48. Another way of describing the topology on $\hat{A}$ is to fix a Hilbert space $\mathcal{H}$ and let $\mathrm{Irr}_{\mathcal{H}}(A)$ be the set of irreducible representations of A on $\mathcal{H}$. $\mathrm{Irr}_{\mathcal{H}}(A)$ has a natural topology, the topology of elementwise convergence, and the map from $\mathrm{Irr}_{\mathcal{H}}(A)$ to $\hat{A}$ is continuous and open, i.e. $\hat{A}$ has the quotient topology from $\mathrm{Irr}_{\mathcal{H}}(A)$ (more care must be taken in describing the topology if A has irreducible representations of different dimensions. The difficulty can be avoided by stabilization (XX.16.4.55.) if A is separable, since there are natural homeomorphisms $[A \otimes \mathbb{K}]\hat{} \cong \hat{A}$ and $\mathrm{Prim}(A \otimes \mathbb{K}) \cong \mathrm{Prim}(A)$.) This procedure also yields a Borel structure (the *Mackey Borel structure*) on $\hat{A}$, the quotient of the natural Borel structure on $\mathrm{Irr}_{\mathcal{H}}(A)$ induced by its topology, which is finer than the Borel structure on $\hat{A}$ induced by its topology, and which is more useful than the topology if $\hat{A}$ is not T_0 (e.g. the Mackey Borel structure separates points of $\hat{A}$).

Matrix Algebras and Stable Algebras

As an example of a basic fact which can be proved easily by representation theory but for which there seems to be no decent space-free proof, we observe that a matrix algebra over a C*-algebra is also a C*-algebra.

XX.16.4.49. If $\mathcal{H}$ is a Hilbert space, then $\mathcal{B}(\mathcal{H}^n)$ is naturally isomorphic to the matrix algebra $M_n(\mathcal{B}(\mathcal{H}))$. (More generally, $\mathcal{B}(\mathcal{H}^n, \mathcal{H}^m)$ can be identified with the $m \times n$ matrices over $\mathcal{B}(\mathcal{H})$.) So if A is a concrete C*-algebra of operators on $\mathcal{H}$, then the matrix algebra $M_n(A)$ acts naturally as a concrete C*-algebra of operators on $\mathcal{H}^n$ (it is easily seen to be complete). The adjoint of a matrix (a_{ij}) is the matrix (b_{ij}), where $b_{ij} = a_{ji}^*$.

XX.16.4.50. Thus, if A is a C*-algebra, we need only to take a faithful representation π of A on a Hilbert space $\mathcal{H}$; then π defines entrywise a faithful representation $\pi^{(n)}$ of $M_n(A)$ (with involution defined as above) on $\mathcal{H}^n$, and thus $M_n(A)$ is a C*-algebra with the induced operator norm.

XX.16.4.51. By XX.16.2.34., this operator norm is the unique C*-norm on $M_n(A)$ with this involution. When we use matrix algebras over C*-algebras, we will always implicitly use this involution and norm. Note that there is no explicit formula for the norm of a matrix in terms of the entries in general (e.g. if $A = \mathbb{M}_3$ and $n = 2$, cf. XX.12.6.6.). There is, however, a simple estimate: if $a = (a_{ij})$ is a matrix, then

$$\max_{i,j} \|a_{ij}\| \leq \|a\| \leq \sum_{i,j} \|a_{ij}\|$$

since it is easily seen that if b is a matrix with exactly one nonzero entry x, then $\|b\| = \|x\|$.

XX.16.4.52. $M_n(A)$ is isomorphic to $A \otimes \mathbb{M}_n = A \otimes_{\mathbb{C}} \mathbb{M}_n$ (), and it is convenient to use tensor product notation in matrix algebras. In $\mathbb{M}_n$, let e_{ij} be the matrix with 1 in the (i,j)'th entry and zeroes elsewhere, and write $a \otimes e_{ij}$ for the element of $M_n(A)$ with a in the (i,j)'th entry and zeroes elsewhere. The e_{ij} are called the *standard matrix units* in $\mathbb{M}_n$. (There is a similar set of standard matrix units in $\mathbb{K}$.)

XX.16.4.53. As in XX.16.4.49., $\mathcal{B}(\mathcal{H}^\infty)$ can be identified with an algebra of infinite matrices over $\mathcal{B}(\mathcal{H})$ (although it is impossible to give an explicit description of which matrices give bounded operators). If A is a concrete C*-algebra of operators on $\mathcal{H}$, let $M_\infty(A)$ denote the infinite matrices over A with only finitely many nonzero entries. Then $M_\infty(A)$ acts naturally as a *-algebra of bounded operators on $\mathcal{H}^\infty$, and its closure in $\mathcal{B}(\mathcal{H}^\infty)$ is denoted $A \otimes \mathbb{K}$.

XX.16.4.54. If A is an (abstract) C*-algebra, we can form $M_\infty(A)$ in the same manner, and by choosing a faithful representation π of A, $M_\infty(A)$ may be identified with $M_\infty(\pi(A))$ and thus given a norm. Using the uniqueness of norm on $M_n(A)$ for each n, it is easily seen that the norm on $M_\infty(A)$ does not depend on the choice of π.

XX.16.4.55. Definition. The completion of $M_\infty(A)$ is called the *stable algebra* of A, denoted $A \otimes \mathbb{K}$.

This is also a special case of a tensor product of C*-algebras, hence the notation.

XX.16.4.56. Definition. A C*-algebra A is *stable* if $A \cong A \otimes \mathbb{K}$. Two C*-algebras A and B are *stably isomorphic* if $A \otimes \mathbb{K} \cong B \otimes \mathbb{K}$.

Note that a stable C*-algebra is always nonunital.

XX.16.4.57. PROPOSITION. $\mathcal{K}(\mathcal{H}) \otimes \mathbb{M}_n = \mathcal{K}(\mathcal{H}^n)$ and $\mathcal{K}(\mathcal{H}) \otimes \mathbb{K} = \mathcal{K}(\mathcal{H}^\infty)$. In particular,

$$\mathbb{K} \cong M_n(\mathbb{K}) \cong \mathbb{M}_n \otimes \mathbb{K} \cong \mathbb{K} \otimes \mathbb{M}_n$$

for all n, and $\mathbb{K}$ is isomorphic to $\mathbb{K} \otimes \mathbb{K}$.

XX.16.4.58. Similarly, it is easily verified that $(A \otimes \mathbb{K}) \otimes \mathbb{K} \cong A \otimes \mathbb{K}$ for any A, so the stable algebra $A \otimes \mathbb{K}$ of a C*-algebra A is a stable C*-algebra, and A is stably isomorphic to $A \otimes \mathbb{K}$. Two stable C*-algebras are stably isomorphic if and only if they are isomorphic. In the same way, A and $M_n(A)$ are stably isomorphic for any A and n. Stably isomorphic C*-algebras are "the same up to 'size';" a stable C*-algebra has uniformly infinite "size." (This must be taken with a grain of salt in the non-σ-unital case, but can be made precise in a nice way for σ-unital C*-algebras). If X is a locally compact Hausdorff space, then $C_0(X) \otimes \mathbb{K} \cong C_0(X, \mathbb{K})$ (XX.16.1.3.).

XX.16.5. Universal and Enveloping C*-Algebras

XX.16.6. Exercises

XX.16.6.1. Use XX.16.2.14. and XX.16.2.16. to give an alternate proof of XX.16.1.9., as follows. Let B be a commutative C*-algebra, which may be assumed unital by XX.16.1.2., and $x = x^* \in B$.

(a) Suppose $\lambda = \alpha + i\beta \in \sigma_B(x)$ with $\alpha, \beta \in \mathbb{R}$. By XX.16.2.16. there is a $\phi \in \hat{B}$ with $\phi(x) = \lambda$.

(b) For $t \in \mathbb{R}$, set $x_t = x + it1$. Then $\|x_t\|^2 = \|x_t^* x_t\| = \|x^2 + t^2 1\| \leq \|x^2\| + t^2 = \|x\|^2 + t^2$.

(c) On the other hand, $\phi(x_t) = \alpha + i(\beta + t)$, so $|\phi(x_t)|^2 = \alpha^2 + (\beta + t)^2$.

(d) Since $|\phi(x_t)| \leq \|x_t\|$ by XX.16.2.14., we have

$$\alpha^2 + (\beta + t)^2 \leq \|x\|^2 + t^2$$

for all $t \in \mathbb{R}$. Show that this implies that $\beta = 0$.

(e) If A is a general not necessarily commutative C*-algebra (which may be assumed unital) and $x = x^* \in A$, the above argument can be applied to the C*-subalgebra B of A generated by x and 1. Why does this give an alternate proof of the full result XX.16.1.9. (what is the relation between $\sigma_B(x)$ and $\sigma_A(x)$)?

XX.16.6.2. (a) Let I be a closed bounded interval in $\mathbb{R}$, and A the set of bounded functions from I to $\mathbb{C}$ whose real and imaginary parts are in $\mathfrak{BC}_1(I)$ (XVI.2.5.2.). Give A pointwise operations and supremum norm. Show that A is a (unital) commutative C*-algebra. [Use XVI.2.10.7..]

(b) Let B be the set of regulated functions on I (V.8.6.7.). Show that B is a (unital) C*-subalgebra of A containing $C(I)$.

(c) Show that $\hat{B}$ is homeomorphic to $I \times \{0, 1\}$ with the order topology from the lexicographic ordering (). The quotient map to $I \cong \widehat{C(I)}$ is projection onto the first coordinate. Note that $\hat{B}$ is not metrizable, so B is not separable (this has an easy direct proof too).

(d) Try to describe $\hat{A}$. (Note that $\hat{B}$ is a quotient of $\hat{A}$.)

XX.16.6.3. Let A be the set of bounded uniformly continuous functions from $\mathbb{R}$ to $\mathbb{C}$. Show that A is a proper unital C*-subalgebra of $BC(\mathbb{R})$. Thus $\hat{A}$ is a quotient of $\widehat{BC(\mathbb{R})} \cong \beta\mathbb{R}$. Describe the quotient map.

Chapter XXI

Nonlinear Functional Analysis

XXI.1. Calculus on Banach Spaces

Much of vector calculus can be generalized to functions between Banach spaces and, more generally, topological vector spaces. The general principle is that "differentiable" is interpreted as "approximately linear" in an appropriate sense. The main technicalities in the infinite-dimensional case are due to two facts: (1) linear transformations between Banach spaces are not necessarily continuous (bounded), and (2) Banach spaces have various weak topologies distinct from the norm topology.

Banach spaces can be either real or complex. There will thus be two parallel theories; although the basic definitions and results are essentially identical, we must be careful to distinguish them. Thus we fix a field $\mathbb{F}$, either $\mathbb{R}$ or $\mathbb{C}$.

XXI.1.1. Differentiable Functions Between Banach Spaces

We can use the definition of differentiability from vector calculus almost verbatim in the Banach space case:

XXI.1.1.1. DEFINITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces (over $\mathbb{F}$), U an open subset of $\mathcal{X}$, $f : U \to \mathcal{Y}$ a function, and $a \in U$. Then

(i) f is *approximately linear* at a if there is a linear ($\mathbb{F}$-linear) transformation $L : \mathcal{X} \to \mathcal{Y}$ such that

$$\lim_{h \to 0} \frac{f(a+h) - f(a) - L(h)}{\|h\|} = 0 \ .$$

(i) f is *differentiable* at a if there is a bounded linear ($\mathbb{F}$-linear) transformation $L : \mathcal{X} \to \mathcal{Y}$ such that

$$\lim_{h \to 0} \frac{f(a+h) - f(a) - L(h)}{\|h\|} = 0 \ .$$

XXI.1.1.2. Of course, the condition in the definition is equivalent to

$$\lim_{h\to 0} \frac{\|f(a+h)-f(a)-L(h)\|}{\|h\|} = 0 \ .$$

The definition is dependent on the choice of $\mathbb{F}$, and when it is necessary to distinguish between $\mathbb{R}$ and $\mathbb{C}$ we may use the terms "approximately $\mathbb{F}$-linear" and "$\mathbb{F}$-differentiable." Complex Banach spaces have an underlying structure as real Banach spaces, and a $\mathbb{C}$-differentiable function is $\mathbb{R}$-differentiable with respect to the underlying real structures; but since an $\mathbb{R}$-linear function is not $\mathbb{C}$-linear in general, an $\mathbb{R}$-differentiable function is not necessarily $\mathbb{C}$-differentiable.

XXI.1.1.3. In the finite-dimensional case, there is no difference between differentiability and approximate linearity, and a differentiable function is automatically continuous. But the notions are distinct in the infinite-dimensional case: an unbounded linear transformation is obviously approximately linear, but neither differentiable nor continuous. But we have:

XXI.1.1.4. Proposition. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, U an open subset of $\mathcal{X}$, $f : U \to \mathcal{Y}$ a function, and $a \in U$. Then f is differentiable at a if and only if f is approximately linear and continuous at a.

In particular, a linear or affine function is differentiable (by our definition) if and only if it is continuous. Proof: Half the proof is the same as in (): if f is differentiable at a, and L is a bounded operator as in the definition, and $\epsilon > 0$, there is a δ, $0 < \delta < 1$, such that $\|L(h)\| < \frac{\epsilon}{2}$ and

$$\|f(a+h)-f(a)-L(h)\| < \frac{\|f(a+h)-f(a)-L(h)\|}{\|h\|} < \frac{\epsilon}{2}$$

if $\|h\| < \delta$. We then have, for $\|h\| < \delta$,

$$\|f(a+h)-f(a)\| = \|f(a+h)-f(a)-L(h)+L(h)\| \le \|f(a+h)-f(a)-L(h)\| + \|L(h)\| < \frac{\epsilon}{2}+\frac{\epsilon}{2} = \epsilon \ .$$

For the other direction, suppose f is continuous at a and there is a linear transformation L as in the definition of approximate linearity. Choose δ, $0 < \delta < 1$, so that

$$\|f(a+h)-f(a)\| < \frac{\epsilon}{2}$$

$$\|f(a+h)-f(a)-L(h)\| < \frac{\epsilon\|h\|}{2} < \frac{\epsilon}{2}$$

for $\|h\| < \delta$. Then, if $\|h\| < \delta$,

$$\|L(h)\| = \|[f(a+h)-f(a)]-[f(a+h)-f(a)-L(h)]\| \le \|f(a+h)-f(a)\|+\|f(a+h)-f(a)-L(h)\| < \frac{\epsilon}{2}+\frac{\epsilon}{2} = \epsilon$$

so L is continuous and f is differentiable at a. 👍

This proof actually shows slightly more than the stated result, but not when combined with the following observation:

XXI.1.1.5. PROPOSITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, U an open subset of $\mathcal{X}$, $f : U \to \mathcal{Y}$ a function, and $a \in U$. If f is approximately linear at a, then the L in the definition is unique.

PROOF: Suppose L_1 and L_2 are linear transformations from $\mathcal{X}$ to $\mathcal{Y}$ with

$$\lim_{h\to 0} \frac{\|f(a+h) - f(a) - L_1(h)\|}{\|h\|} = 0$$

$$\lim_{h\to 0} \frac{\|f(a+h) - f(a) - L_2(h)\|}{\|h\|} = 0$$

and let $\epsilon > 0$. There is a δ, $0 < \delta < 1$, such that

$$\frac{\|f(a+h) - f(a) - L_1(h)\|}{\|h\|} < \frac{\epsilon}{2}$$

$$\frac{\|f(a+h) - f(a) - L_2(h)\|}{\|h\|} < \frac{\epsilon}{2}$$

for $\|h\| < \delta$. Then, if $\|h\| < \delta$,

$$\frac{\|L_2(h) - L_1(h)\|}{\|h\|} = \frac{\|[f(a+h) - f(a) - L_1(h)] - [f(a+h) - f(a) - L_2(h)]\|}{\|h\|}$$

$$\leq \frac{\|f(a+h) - f(a) - L_1(h)\|}{\|h\|} + \frac{\|f(a+h) - f(a) - L_2(h)\|}{\|h\|} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \ .$$

Let $x \in \mathcal{X}$, $\|x\| = 1$. Then, if $0 < \alpha < \delta$, $\|\alpha x\| < \delta$, so

$$\frac{\|L_2(\alpha x) - L_1(\alpha x)\|}{\|\alpha x\|} = \|L_2(x) - L_1(x)\| < \epsilon$$

and since $\epsilon > 0$ is arbitrary, $L_1(x) = L_2(x)$. Since x is an arbitrary vector of norm 1, $L_1 = L_2$. 👍

XXI.1.1.6. DEFINITION. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, U an open subset of $\mathcal{X}$, $f : U \to \mathcal{Y}$ a function, and $a \in U$. If f is differentiable at a, the *derivative* of f at a is the unique operator $Df(a)$ satisfying

$$\lim_{h\to 0} \frac{f(a+h) - f(a) - Df(a)h}{\|h\|} = 0 \ .$$

We will only use the notation $Df(a)$ if f is differentiable at a, i.e. if $Df(a)$ is a bounded operator. Thus $Df(a) \in \mathcal{B}(\mathcal{X}, \mathcal{Y})$.

XXI.1.1.7. Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, U an open subset of $\mathcal{X}$, and $f : U \to \mathcal{Y}$ a function. We say f is differentiable on U if it is differentiable at every point of U. If so, Df is a function from $\mathcal{X}$ to $\mathcal{B}(\mathcal{X}, \mathcal{Y})$. We say f is *continuously differentiable*, or $\mathcal{C}^1$, on U if Df is a continuous function from U to $\mathcal{B}(\mathcal{X}, \mathcal{Y})$ with its norm topology.

We can also define higher-order differentiability since $\mathcal{B}(\mathcal{X}, \mathcal{Y})$ is itself a Banach space: f is second-order differentiable at $a \in U$ if Df is differentiable at a, and f is $\mathcal{C}^2$ on U if Df is $\mathcal{C}^1$ on U. Higher-order differentiability and $\mathcal{C}^r$ $(2 < r \leq \infty)$ can be defined analogously.

XXI.1.1.8. EXAMPLE. [Die60, 8.3.2] Let A be a unital Banach algebra, and write A^{-1} for the set of invertible elements in A. Then A^{-1} is an open set in A (XX.12.2.11.). Define $f : A^{-1} \to A^{-1} \subseteq A$ by $f(x) = x^{-1}$. Then f is continuous on A^{-1} (XX.12.2.14.). In fact, f is differentiable on A^{-1} and, if $a \in A^{-1}$, $[Df(a)]x = a^{-1}xa^{-1}$. For note that if $u \in A$, $\|u\| < 1$, then $1 + u \in A^{-1}$ and the infinite series

$$\sum_{k=0}^{\infty}(-1)^k u^k$$

converges to $(1 + u)^{-1}$ (XX.12.2.9.); we then have that

$$\sum_{k=2}^{\infty}(-1)^k u^k = u^2 \sum_{k=0}^{\infty}(-1)^k u^k = \left[\sum_{k=0}^{\infty}(-1)^k u^k\right] - 1 + u$$

converges to $(1 + u)^{-1} - 1 + u = u^2(1 + u)^{-1}$. Thus we obtain (XX.12.2.9.)

$$\|(1 + u)^{-1} - 1 + u\| = \|u^2(1 - u)^{-1}\| \leq \|u\|^2 \frac{1}{1 - \|u\|} = \frac{\|u\|^2}{1 - \|u\|} \ .$$

If $h \in A$, $\|h\| < \|a^{-1}\|^{-1}$, then $\|a^{-1}h\| < 1$, so $1 + a^{-1}h \in A^{-1}$ and hence $a + h = a(1 + a^{-1}h)$ is invertible, with inverse $(1 + a^{-1}h)^{-1}a^{-1}$, and

$$(a + h)^{-1} - a^{-1} = [(1 + a^{-1}h)^{-1} - 1]a^{-1} \ .$$

Thus, taking $u = a^{-1}h$ above, we have

$$\|(a + h)^{-1} - a^{-1} + a^{-1}ha^{-1}\| = \|([(1 + a^{-1}h)^{-1} - 1] + a^{-1}h)a^{-1}\| \leq \|a^{-1}\|\|(1 + a^{-1}h)^{-1} - 1 + a^{-1}h\|$$

$$\leq \|a^{-1}\| \frac{\|a^{-1}h\|^2}{1 - \|a^{-1}h\|} = \frac{\|a^{-1}\|^3\|h\|^2}{1 - \|a^{-1}h\|} \leq 2\|a^{-1}\|^3\|h\|^2$$

if $\|h\| \leq \frac{1}{2\|a^{-1}\|}$. Thus, for $\|h\| \leq \frac{1}{2\|a^{-1}\|}$,

$$\frac{\|(a + h)^{-1} - a^{-1} + a^{-1}ha^{-1}\|}{\|h\|} \leq 2\|a^{-1}\|^3\|h\|$$

which goes to 0 as $h \to 0$.

XXI.1.2. The Implicit and Inverse Function Theorems

The Implicit and Inverse Function Theorems, and the General Implicit/Inverse Function Theorem VIII.5.2.1., have nearly verbatim versions for functions between Banach spaces.

XXI.1.2.1. If $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces, there are many ways of defining a norm on $\mathcal{X} \times \mathcal{Y}$ giving the product topology (). Any of these can be used in this section, but for specificity let us standardize on the ∞-norm: $\|(x, y)\| = \max(\|x\|, \|y\|)$.

XXI.1.2.2. We also need to define analogs of $D_x f$ and $D_y f$. If L is a linear transformation from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{Z}$, write $L_{\mathcal{X}}$ for the linear transformation from $\mathcal{X}$ to $\mathcal{Z}$ given by $L_{\mathcal{X}}(x) = L(x, 0)$, and $L_{\mathcal{Y}}$ the linear transformation from $\mathcal{Y}$ to $\mathcal{Z}$ given by $L_{\mathcal{Y}}(y) = L(0, y)$. If L is bounded, then $L_{\mathcal{X}}$ and $L_{\mathcal{Y}}$ are bounded (and conversely). If f is a function from an open subset U of $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{Z}$, and $(a, b) \in U$, and f is differentiable at (a, b), write $D_{\mathcal{X}} f(a, b) = [Df(a, b)]_{\mathcal{X}}$ and $D_{\mathcal{Y}} f(a, b) = [Df(a, b)]_{\mathcal{Y}}$. Then $D_{\mathcal{X}} f(a, b) \in \mathcal{B}(\mathcal{X}, \mathcal{Z})$ and $D_{\mathcal{Y}} f(a, b) \in \mathcal{B}(\mathcal{Y}, \mathcal{Z})$.

The statements of the theorems are as follows:

XXI.1.2.3. Theorem. [Implicit Function Theorem, Banach Smooth Version] Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, U an open subset of $\mathcal{X} \times \mathcal{Y}$, $F : U \to \mathcal{Y}$ a function, and $(a, b) \in U$ with $F(a, b) = 0$. Suppose F is $\mathcal{C}^r$ on U for some r, $1 \leq r \leq \infty$. If $D_{\mathcal{Y}} F(a, b)$ is invertible, then there are:

(i) An open set V in $\mathcal{X} \times \mathcal{Y}$ containing (a, b), with $V \subseteq U$, such that $D_{\mathcal{Y}} F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open set W in $\mathcal{X}$ containing a;

(iii) A unique function $f : W \to \mathcal{Y}$ such that $f(a) = b$ and, for all $x \in W$, $(x, f(x)) \in V$ and

$$F(x, f(x)) = 0 \ .$$

Additionally, f is $\mathcal{C}^r$ on W, and for any $x \in W$, we have

$$Df(x) = -D_{\mathcal{Y}} F(x, f(x))^{-1} D_{\mathcal{X}} F(x, f(x))$$

(in particular, $D_{\mathcal{Y}} F(x, f(x))$ is invertible for all $x \in W$).

XXI.1.2.4. Theorem. [Inverse Function Theorem, Banach Space Version] Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces, U an open subset of $\mathcal{X}$, $F : U \to \mathcal{Y}$ a $\mathcal{C}^r$ function for some r, $1 \leq r \leq \infty$, and $a \in U$. If $Df(a)$ is invertible, then there is a neighborhood V of a, $V \subseteq U$, such that

(i) f is one-to-one on V.

(ii) $f(V)$ is open in $\mathcal{Y}$.

(iii) $Df(x)$ is invertible for all $x \in V$.

(iv) f^{-1} is $\mathcal{C}^r$ on $f(V)$ and, for all $y \in f(V)$,

$$D(f^{-1})(y) = [Df(f^{-1}(y))]^{-1} \ .$$

XXI.1.2.5. Theorem. [General Implicit/Inverse Function Theorem] Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be Banach spaces, U an open subset of $\mathcal{X} \times \mathcal{Y}$, $F : U \to \mathcal{Z}$ a $\mathcal{C}^r$ function, $a \in \mathcal{X}$, $b \in \mathcal{Y}$, $c \in \mathcal{Z}$ with $(a, b) \in U$ and $F(a, b) = c$. If $D_{\mathcal{Y}} F(a, b)$ is invertible, then there are:

(i) An open neighborhood V of (a, b) in $\mathcal{X} \times \mathcal{Y}$, contained in U, such that $D_{\mathcal{Y}} F(x, y)$ is invertible for all $(x, y) \in V$;

(ii) An open neighborhood W of (a, c) in $\mathcal{X} \times \mathcal{Z}$;

(iii) A unique function $G : W \to \mathcal{Y}$ satisfying $G(a, c) = b$ and, for all $(x, z) \in W$, $(x, G(x, z)) \in V$ and

$$F(x, G(x, z)) = z \ .$$

Additionally, G is $\mathcal{C}^r$ on W, and for all $(x, z) \in W$,

$$[DG(x, z)](h, k) = -[D_{\mathcal{Y}} F(x, G(x, z))^{-1} D_{\mathcal{X}} F(x, G(x, z))]h + [D_{\mathcal{Y}} F(x, G(x, z))^{-1}]k$$

for $h \in \mathcal{X}, k \in \mathcal{Y}$.

XXI.1.2.6. Note that in these theorems it is allowed for $\mathcal{X}$, $\mathcal{Y}$, and/or $\mathcal{Z}$ to be the zero Banach space; this is mostly uninteresting except if the $\mathcal{X}$ in XXI.1.2.5. is zero, in which case the result is precisely the Inverse Function Theorem XXI.1.2.4..

XXI.1.2.7. The special case of the Inverse Function Theorem XXI.1.2.4. where f is linear is the Bounded Inverse Theorem. (But the Bounded Inverse Theorem is used in the proof or, depending on one's point of view, is implicit in the assumption that $Df(a)$ is invertible.)

Proofs of the Theorems

We gave two proofs of the theorems in the finite-dimensional case. The induction proof () will not work in the Banach space case, but the fixed-point proof () does with little modification.

XXI.1.2.8. It will suffice to prove the Inverse Function Theorem XXI.1.2.4., since then XXI.1.2.5. will follow as in () (cf. ()), and as in (), the Implicit Function Theorem XXI.1.2.3. is a corollary of XXI.1.2.5. by taking $\mathcal{Z} = \mathcal{Y}$ and $c = 0$, and setting $f(x) = G(x, 0)$.

The Rank and Straightening Theorems

XXI.1.2.9. There are serious difficulties in generalizing the Rank and Straightening Theorems to the infinite-dimensional case, due to at least four problems:

(i) A bounded operator between Banach spaces need not have closed range.

(ii) A closed subspace of a Banach space, even the null space or closure of the range of a bounded operator, need not have a closed complementary subspace.

(iii) A Banach space does not have a canonical basis, and need not have any (Schauder) basis at all.

(iv) It is unclear in the infinite-dimensional case how to define "constant rank."

Even if the Banach spaces are Hilbert spaces, so (ii) does not occur and (iii) can be managed, (i) and (iv) make a full analog of the Rank Theorem impossible.

Nonetheless, there are infinite-dimensional versions of the theorems, with some additional hypotheses:

XXI.1.3. Integration in Banach Spaces

XXI.2. Calculus of Variations

Running roughly parallel to ordinary calculus on the real line or Euclidean space is another version of calculus called *calculus of variations*, which can be thought of as calculus on function spaces. A nutshell description is that the calculus of variations is the study of what functions maximize or minimize an associated quantity (actually, as pointed out in [GF63], this should be called "calculus of variations in the narrow sense," since there is much more to calculus than solving max/min problems, but calculus on function spaces is not well developed beyond this aspect). Two elementary but critical examples are:

(i) Of all curves between two fixed endpoints, the (unique) one with the minimum arc length is a straight line segment ("a line is the shortest distance between two points").

(ii) Of all plane figures bounded by a curve of fixed arc length, the (unique) one of maximum area is a circle (disk).

A somewhat more sophisticated example is the *brachistochrone problem*: of all curves through two fixed points, the one along which a particle will travel under the force of gravity in the least time is a cycloid ().

The historical development of the calculus of variations is also in parallel with the development of ordinary calculus, and many of the big names are the same: Newton, the Bernoullis, Leibniz, Euler (considered the "father" of the calculus of variations), Lagrange, Legendre, Jacobi, Weierstrass, and more recently Hilbert, Lebesgue, Noether, Hadamard, Tonelli, and Marston Morse. It has gone through periods of intense study and relative neglect, with a recent resurgence of interest, in part due to its close connections with control theory.

The place of the calculus of variations in mathematics is somewhere between ordinary calculus, differential geometry, and the theory of differential equations. The calculus of variations is usually thought of as part of "applied mathematics" but, as with differential equations, it cannot be cleanly categorized as such (indeed, the distinction between "pure" and "applied" mathematics is a very hazy one in general!) Although much of the calculus of variations historically predates the development of functional analysis, a modern treatment falls most appropriately under "nonlinear functional analysis."

The calculus of variations is one of the parts of mathematics most intimately connected with physics (although the applications now go far beyond physics to many areas of pure and applied mathematics). The connection is not only historical, with many of the motivations and applications of the theory coming from important problems in physics, but goes much deeper: the "variational principles" of the calculus of variations can be viewed not only as mathematical principles, but also as deep and fundamental physical laws ("principles of least action"). See [HT96] for a delightful and mostly nontechnical treatment.

Our treatment of the calculus of variations will be far from complete; we will merely try to hit the high points and place the theory in context. There are a number of texts covering the theory in detail, such as [Bol61], [For60], [Car65]–[Car67],[**?**], [Bli46], [Par62], [vB04], [Ewi85], [GH96a]–[GH96b], [JLJ98], ..., and several Russian texts available in English translation such as [Akh88], [Esc62], [GF63]. For "calculus of variations in the large," see [Mor96] and [**?**]. See [Gol80] for a history of the subject through the nineteenth century.

XXI.2.1. The Isoperimetric Problem

XXI.2.1.1. As pointed out by Dirichlet, there is a subtle gap in the proof of the isoperimetric problem we have outlined. We have shown that no simple closed curve of arc length L other than a circle can enclose

a region of maximal area; thus we have only shown that a circle is the only simple closed curve which *can* enclose a maximal area, not that it *does*. In other words, we have shown that *if* there is *any* curve enclosing a maximal area, it must be a circle, but we have not yet shown there is a curve realizing the maximum. Filling this gap turns out to be quite tricky – we must use some type of compactness argument.

XXI.2.1.2. For a simple example of the difficulty, consider the following argument (attributed to PERRON). This argument may seem silly, but it is actually far less silly than it appears. Let N be the largest natural number. If n is any natural number other than 1, then $n^2 > n$, so N cannot be n. Therefore $N = 1$.

Chapter XXII

Measure and Topology; Generalized Functions and Distributions

This chapter covers two seemingly distinct subjects: Borel measures on locally compact spaces, and the theory of generalized functions and distributions on $\mathbb{R}$ and $\mathbb{R}^n$. The common theme is that both can be naturally and profitably regarded in functional analysis terms, as continuous linear functionals on spaces of ordinary functions.

XXII.1. Measures on Locally Compact Spaces

XXII.1.1. Regular Measures

XXII.1.2. The Riesz Representation Theorem

XXII.2. Distributions

XXII.2.1. Generalized Functions on $\mathbb{R}$

XXII.2.1.1. The notion of a real-valued (or complex-valued) function on a space seems to entail specification of its values at points of the space. But a functional analysis point of view is to regard functions as themselves points of another space, with properties not closely tied to their values at particular points (e.g. as objects which can be integrated).

XXII.2.1.2. This shift of point of view already appears when L^p spaces are considered. The elements of L^p are technically not functions, but equivalence classes of functions which agree a.e.; thus it does not make sense to talk about the value of an "L^p-function" at any particular point – only the general behavior at "most" points is determined and relevant.

XXII.2.1.3. Another example is the "Dirac δ-function": this is a fictitious function with certain desirable properties which no actual function can have. We have an interpretation of this "function" as a measure (point mass), but we want to somehow regard it as a "generalized function."

XXII.2.1.4. There are (at least) two worthwhile approaches toward generalizing the notion of a function which are well developed in the literature, via functionals on spaces of "test functions" and as limits of ordinary functions. We will begin with the first approach, which was the first historically and which is the dominant one, and return to the second in (). It is useful to have both points of view, since some aspects of the theory are simpler from each.

XXII.2.1.5. We will work on a general open set Ω in $\mathbb{R}$ or, later, $\mathbb{R}^n$, which we may take to be connected without essential loss of generality (the analysis on a general open set can be almost completely reduced to the connected components). Thus, in the one-dimensional case, we may take Ω to be an open interval. The most important case is $\Omega = \mathbb{R}$ (or, later, $\mathbb{R}^n$).

We begin with two natural sets of actual functions on an open interval $\Omega \subseteq \mathbb{R}$:

XXII.2.1.6. DEFINITION. An *ordinary function* on Ω is a function from Ω to $\mathbb{R}$ (or $\mathbb{C}$) which is Lebesgue measurable, and Lebesgue integrable over each closed bounded subinterval of Ω. Denote the set of ordinary functions by $\mathcal{L}^1_{\text{loc}}(\Omega)$ ("locally integrable functions"). Write $L^1_{\text{loc}}(\Omega)$ for the set of equivalence classes of locally integrable functions, with functions agreeing a.e. identified.

A *test function* is a $\mathcal{C}^\infty$ function from Ω to $\mathbb{R}$ (or $\mathbb{C}$) with compact support, i.e. which is zero outside a compact (closed bounded) subinterval of Ω. Denote the set of test functions by $\mathcal{D}(\Omega) = \mathcal{C}^\infty_c(\Omega)$ (or just $\mathcal{D}$ if Ω is understood); the notation $\mathcal{K}(\Omega)$ is also common in the literature.

We also write $\mathcal{E}(\Omega)$ for $\mathcal{C}^\infty(\Omega)$, the set of $\mathcal{C}^\infty$ functions on Ω with no restriction on support and not necessarily bounded. We have $\mathcal{D}(\Omega) \subseteq \mathcal{E}(\Omega)$.

As usual, by slight abuse we will think of the elements of $L^1_{\text{loc}}(\Omega)$ as "functions" although, as noted above, it does not make sense to consider their values at individual points. The sets $\mathcal{L}^1_{\text{loc}}(\Omega)$, $L^1_{\text{loc}}(\Omega)$, and $\mathcal{D}(\Omega)$ are vector spaces in the obvious way. The set $\mathcal{C}(\Omega)$ of continuous functions on Ω is a subspace of $\mathcal{L}^1_{\text{loc}}(\Omega)$, and the quotient map from $\mathcal{L}^1_{\text{loc}}(\Omega)$ to $L^1_{\text{loc}}(\Omega)$ is injective on this subspace, i.e. there is a natural embedding of $\mathcal{C}(\Omega)$ into $L^1_{\text{loc}}(\Omega)$. Since $\mathcal{D}(\Omega) \subseteq \mathcal{E}(\Omega) \subseteq \mathcal{C}(\Omega)$, they can be regarded as subspaces of $L^1_{\text{loc}}(\Omega)$.

XXII.2.1.7. Each ordinary function f defines a linear functional on $\mathcal{D}(\Omega)$ by

$$\langle f, \phi \rangle = \int_\Omega f \phi \, d\lambda$$

for $\phi \in \mathcal{D}(\Omega)$. (Note that for each ϕ, the integral is really over a compact subinterval.) These linear functionals are actually defined on a much larger space of functions than $\mathcal{D}$, e.g. the space of bounded measurable functions on Ω with compact support, but we only apply them to test functions for reasons which will become apparent. The notation suggests an inner product, and in fact extends the standard inner product on $\mathcal{D}(\Omega)$ regarded as a subspace of $\mathcal{L}^1_{\text{loc}}(\Omega)$. The inner product view is a useful one, cf. ().

If $f = g$ a.e., then f and g define the same linear functional. Conversely:

XXII.2.1.8. PROPOSITION. Let f and g be ordinary functions. If $\langle f, \phi\rangle = \langle g, \phi\rangle$ for all test functions ϕ, then $f = g$ a.e.

PROOF: By linearity, it suffices to show that if f is an ordinary function and $\langle f, \phi\rangle = 0$ for every test function ϕ, then $f = 0$ a.e. Fix an interval $[a, b] \subseteq \Omega$, and for $x \in [a, b]$ define

$$F(x) = \int_a^x f\, d\mu \ .$$

Then F is absolutely continuous on $[a, b]$, and $F' = f$ a.e. on $[a, b]$. Suppose F is not constant on $[a, b]$, and fix $x_1, x_2 \in (a, b)$ with $F(x_1) < F(x_2)$, and fix a real number c with $F(x_1) < c < F(x_2)$. By continuity, there are disjoint closed subintervals $[a_1, b_1]$ and $[a_2, b_2]$ of $[a, b]$ containing x_1 and x_2 respectively in their interiors, such that $F(x) < c$ for all $x \in [a_1, b_1]$ and $F(x) > c$ for all $x \in [a_2, b_2]$. Then there is a $\mathcal{C}^\infty$ function ψ supported on $[a_1, b_1] \cup [a_2, b_2]$, with ψ positive on $[a_1, b_1]$ and negative on $[a_2, b_2]$, and such that

$$\int_{a_2}^{b_2} \psi(t)\, dt = -\int_{a_1}^{b_1} \psi(t)\, dt \ .$$

Set

$$\phi(x) = \int_a^x \psi(t)\, dt$$

for $x \in \mathbb{R}$. Then ϕ is $\mathcal{C}^\infty$ and vanishes outside $[a, b]$, hence is a test function, and $\phi' = \psi$. Using Integration by Parts, we have, since $\phi(a) = \phi(b) = 0$ and $(F - c)' = f$ a.e.,

$$\int_{\mathbb{R}} f\phi\, d\lambda = \int_a^b f\phi\, d\lambda = (F(b) - c)\phi(b) - (F(a) - c)\phi(a) - \int_a^b (F(t) - c)\psi(t)\, dt$$

$$= -\int_a^b (F(t) - c)\psi(t)\, dt = -\int_{a_1}^{b_1} (F(t) - c)\psi(t)\, dt - \int_{a_2}^{b_2} (F(t) - c)\psi(t)\, dt > 0$$

since both integrals are strictly negative because the integrands are strictly negative, contradicting that $\int_{\mathbb{R}} f\phi\, d\lambda = 0$.

Thus F is constant on $[a, b]$ (actually identically 0 since $F(a) = 0$), and hence $f = F' = 0$ a.e. on $[a, b]$. Since $[a, b]$ is arbitrary, $f = 0$ a.e. on Ω.

Convergence in $\mathcal{D}(\Omega)$

XXII.2.1.9. We define convergence of sequences in $\mathcal{D}$ as follows. If (ϕ_n) is a sequence in $\mathcal{D}$, $\phi \in \mathcal{D}$, then $\phi_n \to \phi$ in $\mathcal{D}$ if and only if there is a fixed closed bounded interval $[a, b] \subseteq \Omega$ such that all ϕ_n vanish outside $[a, b]$ and $\phi_n^{(k)} \to \phi^{(k)}$ uniformly on $[a, b]$ for each $k \geq 0$.

Note that this notion of convergence is strictly stronger than uniform convergence of (ϕ_n) and all its derivatives, since the ϕ_n must all be supported on the same fixed compact subinterval.

XXII.2.1.10. This notion of convergence of sequences in $\mathcal{D}$ is precisely convergence of sequences for a certain locally convex vector space topology on $\mathcal{D}$ (XXII.2.1.14.). Although this topology is not first countable, continuity of linear functionals is completely determined by convergence of sequences in $\mathcal{D}$ (XXII.2.1.15.).

XXII.2.1.11. PROPOSITION. Let $f \in \mathcal{L}^1_{\text{loc}}(\Omega)$. Then the linear functional defined on $\mathcal{D}(\Omega)$ by f is sequentially continuous in the sense that if $\phi_n \to \phi$ in $\mathcal{D}(\Omega)$, then $\langle f, \phi_n\rangle \to \langle f, \phi\rangle$.

PROOF: If $\phi_n \to \phi$, fix an $[a,b]$ outside which all ϕ_n vanish. Since $\phi_n \to \phi$ uniformly on $[a,b]$, they are uniformly bounded, i.e. there is an M such that $|\phi_n(x)| \le M$ for all n and all $x \in [a,b]$. We have $f\phi_n \to f\phi$ pointwise on $[a,b]$, and $|f\phi_n| \le M|f|$ on $[a,b]$, so

$$\int_{\mathbb{R}} f\phi_n \, d\lambda = \int_a^b f\phi_n \, d\lambda \to \int_a^b f\phi \, d\lambda = \int_{\mathbb{R}} f\phi \, d\lambda$$

by the Dominated Convergence Theorem. 👍

It follows (XXII.2.1.15.) that the linear functional is actually continuous for the topology on $\mathcal{D}$. Note that we did not use the full strength of convergence in $\mathcal{D}$ in the proof: the linear functional is even continuous for the weaker topology on $\mathcal{D}$ as a subspace of $\mathcal{C}_c(\Omega)$ with its topology from ().

Topology and Convergence in $\mathcal{D}(\Omega)$

XXII.2.1.12. Let $I = [a,b]$ be an interval. Define norms $\|\cdot\|_n$ $(n \ge 0)$ on $\mathcal{C}^\infty([a,b])$ by

$$\|\phi\|_{n,I} = \sum_{k=0}^{n} \left[\max_{a \le x \le b} |\phi^{(k)}(x)| \right] .$$

Then $\mathcal{C}^\infty([a,b])$ is a Fréchet space with its topology from this family of norms (), and

$$\mathcal{C}^\infty_0([a,b]) = \{\phi \in \mathcal{C}^\infty([a,b]) : \phi^{(k)}(a) = \phi^{(k)}(b) = 0 \text{ for all } k \ge 0\}$$

is a closed subspace, hence also a Fréchet space. If $[a,b] \subseteq \Omega$, then $\mathcal{C}^\infty_0([a,b])$ can be naturally considered a subspace of $\mathcal{C}^\infty_c(\Omega)$, and

$$\mathcal{C}^\infty_c(\Omega) = \bigcup_{[a,b]\subseteq\Omega} \mathcal{C}^\infty_0([a,b]) = \bigcup_{n=1}^{\infty} \mathcal{C}^\infty_0([a_n,b_n])$$

where $([a_n,b_n])$ is an increasing sequence of intervals with union Ω. We give $\mathcal{D} = \mathcal{C}^\infty_c(\Omega)$ the inductive limit topology. Note that if $[a,b] \subseteq [c,d]$, the embedding of $\mathcal{C}^\infty_0([a,b])$ into $\mathcal{C}^\infty_0([c,d])$ is a homeomorphism onto its image, which is a closed set; thus the restriction of the inductive limit topology to $\mathcal{C}^\infty_0([a,b])$ is its Fréchet topology. The topology on $\mathcal{D}(\Omega)$ is locally convex, but not metrizable (not first countable).

XXII.2.1.13. Alternatively, we can give a local base at 0 for the topology on $\mathcal{D}$ as follows. Let $\boldsymbol{c} = (c_1, c_2, \dots)$ be a sequence of strictly positive real numbers. Then for each $n \in \mathbb{N}$ set

$$U_{n,\boldsymbol{c}} = \{\phi \in \mathcal{D}(\Omega) : |\phi^{(k)}(x)| < c_m \text{ for } a_m \le x \le b_m \ \forall m, 0 \le k \le n\} .$$

It is a straightforward exercise to show that these sets are convex and open and (for all n and $\boldsymbol{c}$) form a local base at 0, and translates a local base at any ϕ.

XXII.2.1.14. PROPOSITION. If (ϕ_n) is a sequence in $\mathcal{D}(\Omega)$, $\phi \in \mathcal{D}(\Omega)$, then $\phi_n \to \phi$ in the inductive limit topology if and only if $\phi_n \to \phi$ in the sense of XXII.2.1.9..

PROOF: It is obvious that if $\phi_n \to \phi$ in the sense of XXII.2.1.9., then $\phi_n \to \phi$ in the inductive limit topology. Conversely, suppose $\phi_n \to 0$ in the inductive limit topology. Then clearly $\phi_n^{(k)} \to 0$ uniformly on Ω for all k (for any k and $\epsilon > 0$ consider $U_{k,\boldsymbol{c}}$, where $\boldsymbol{c} = (\epsilon, \epsilon, \dots)$). To see that the ϕ_n must all vanish outside some fixed $[a_m, b_m]$, suppose not. Then, passing to a subsequence, we may assume that for each n there is an $x_n \in \Omega$ with $x_n \notin [a_n, b_n]$ and $\phi_n(x_n) \neq 0$. Choose a decreasing sequence (c_n) of strictly positive real numbers with $c_n < |\phi_n(x_n)|$. If $\boldsymbol{c} = (c_1, c_2, \dots)$, then $\phi_n \notin U_{1,\boldsymbol{c}}$ for any n, a contradiction. 👍

XXII.2.1.15. COROLLARY. A linear functional f on $\mathcal{D}(\Omega)$ is continuous if and only if $\langle f, \phi_n \rangle \to \langle f, \phi \rangle$ whenever $\phi_n \to \phi$.

PROOF: A linear functional on $\mathcal{D}(\Omega)$ is continuous if and only if its restriction to $\mathcal{C}_0^\infty([a,b])$ is continuous for every $[a,b] \subseteq \Omega$. But $\mathcal{C}_0^\infty([a,b])$ is metrizable, so continuity of functionals on it is determined by sequential continuity. 👍

XXII.2.1.16. PROPOSITION. A linear functional f on $\mathcal{D}(\Omega)$ is continuous if and only if, for every closed bounded interval $I \subseteq \Omega$, there is an n and $c > 0$ such that

$$|f(\phi)| \leq c\|\phi\|_{n,I}$$

for all ϕ supported in I.

PROOF: If the condition is satisfied and $\phi_k \to \phi$ in $\mathcal{D}(\Omega)$, all the ϕ_k are supported in some I. Let n and c correspond to I. The restriction of f to the functions supported in I is bounded with respect to the norm $\|\cdot\|_{n,I}$, hence continuous (), and hence is also continuous with respect to the stronger $\mathcal{D}$-topology.

Conversely, if the condition is not satisfied for some I, there is for each n a ϕ_n supported in I such that $f(\phi_n) = 1$ and $\|\phi_n\|_{n,I} < \frac{1}{n}$. Then $\phi_n \to 0$ in $\mathcal{D}(\Omega)$ but $f(\phi_n) \not\to 0$, so f is not continuous. 👍

XXII.2.1.17. Similarly, if (I_n) is an increasing sequence of closed bounded intervals with union Ω, we give $\mathcal{E}(\Omega)$ the topology defined by the norms $\|\cdot\|_{n,I_n}$. Thus $\phi_n \to \phi$ in $\mathcal{E}(\Omega)$ if and only if $\phi_n^{(k)} \to \phi^{(k)}$ u.c. on Ω for every k. Then $\mathcal{E}(\Omega)$ is a Fréchet space. Note that the topology on $\mathcal{D}(\Omega)$ is strictly stronger than the relative topology from $\mathcal{E}(\Omega)$. ✋

Generalized Functions

We thus make the following definition:

XXII.2.1.18. DEFINITION. A *generalized function* or *distribution* on Ω is a (sequentially) continuous linear functional on $\mathcal{D}(\Omega)$. A generalized function which is defined by an ordinary function as above is called an *ordinary function*. Denote the set of generalized functions on Ω by $\mathcal{D}'(\Omega)$ (or just $\mathcal{D}'$ if Ω is understood); $\mathcal{K}'(\Omega)$ is also used.

There is slight abuse of terminology here: an ordinary function in this sense corresponds to an a.e. equivalence class of ordinary functions in the sense of XXII.2.1.6., via the correspondence $f \mapsto \langle f, \cdot \rangle$. This correspondence is injective by XXII.2.1.8., i.e. $L^1_{\text{loc}}(\Omega)$ can be thought of as a subspace of the vector space $\mathcal{D}'(\Omega)$.

The notation $\mathcal{D}'$ is standard in distribution theory for the dual space (cf. ()), and has nothing to do with derivatives as defined in XXII.2.1.25.; the notation $\mathcal{D}^*$ would be less confusing, but we stick to convention.

There are many other generalized functions:

XXII.2.1.19. EXAMPLES. (i) Define δ by

$$\langle \delta, \phi \rangle = \phi(0) \ .$$

If $\phi_n \to \phi$ in $\mathcal{D}(\mathbb{R})$, then certainly $\phi_n \to \phi$ pointwise, so $\langle \delta, \phi_n \rangle \to \langle \delta, \phi \rangle$, and δ is continuous. This is the "Dirac δ-function."

(ii) More generally, let μ be a Radon measure on $\mathbb{R}$, i.e. a signed or complex Borel measure which is finite on bounded sets. Define a linear functional on $\mathcal{D}(\mathbb{R})$ by

$$\langle \mu, \phi \rangle = \int_{\mathbb{R}} \phi \, d\mu \ .$$

If ϕ is supported in $[a, b]$, $\mu|_{[a,b]}$ is finite, so the integral is defined. If $\phi_n \to \phi$ in $\mathcal{D}$, the ϕ_n are all supported in some $[a, b]$ and converge uniformly to ϕ, so, since $\mu|_{[a,b]}$ is finite,

$$\int_{\mathbb{R}} \phi_n \, d\mu = \int_a^b \phi_n \, d\mu \to \int_a^b \phi \, d\mu = \int_{\mathbb{R}} \phi \, d\mu \ .$$

Thus the linear functional is continuous. By the Radon-Nikodým Theorem, the ordinary functions can be identified with the Radon measures which are absolutely continuous with respect to Lebesgue measure: if $\mu \ll \lambda$, then μ is identified with its density function $\left[\frac{d\mu}{d\lambda}\right]$.

(iii) Functions which are not ordinary functions can define generalized functions. For example, $f(x) = \frac{1}{x}$ is not an ordinary function since it is not locally integrable around 0. But we can use Cauchy principal value to define a linear functional:

$$\langle f, \phi \rangle = (PV) \int_{\mathbb{R}} \frac{\phi(x)}{x} \, d\lambda$$

$$= \lim_{\epsilon \to 0} \left[\int_{-\infty}^{-\epsilon} \frac{\phi(x)}{x} \, dx + \int_{\epsilon}^{\infty} \frac{\phi(x)}{x} \, dx \right]$$

$$= \lim_{\epsilon \to 0} \left[\int_{\epsilon}^{\infty} \frac{\phi(x)}{x} \, dx - \int_{\epsilon}^{\infty} \frac{\phi(-x)}{x} \, dx \right]$$

$$= \lim_{\epsilon \to 0} \int_{\epsilon}^{\infty} \frac{\phi(x) - \phi(-x)}{x} \, dx \ .$$

Since ϕ is differentiable at 0,

$$\lim_{x \to 0} \frac{\phi(x) - \phi(-x)}{x}$$

exists (and equals $2\phi'(0)$), i.e. $\psi(x) = \frac{\phi(x) - \phi(-x)}{x}$ extends to a continuous function on $\mathbb{R}$ (with compact support); so

$$\langle f, \phi \rangle = \int_{0}^{\infty} \frac{\phi(x) - \phi(-x)}{x} \, dx \ .$$

This linear functional is obviously continuous, and this generalized function is commonly denoted $\frac{1}{x}$.

(iv) We can do higher-order versions of (i) and (ii): for any k, we can define

$$\langle \delta^{(k)}, \phi \rangle = (-1)^k \phi^{(k)}(0)$$

and, more generally, if μ is a Radon measure,

$$\langle \mu^{(k)}, \phi \rangle = (-1)^k \int_{\mathbb{R}} \phi^{(k)} \, d\mu$$

(the reason for the conventional factor of $(-1)^k$ will appear in XXII.2.1.25.).

(v) Define

$$\langle f, \phi \rangle = \sum_{k=0}^{\infty} \phi^{(k)}(k)$$

(this is a finite sum for any ϕ).

XXII.2.1.20. There is a natural topology on $\mathcal{D}'(\Omega)$, the topology of pointwise convergence ("weak-* topology"). A local base for this topology at 0 consists of sets of the form

$$U(\phi_1, \ldots, \phi_m, \epsilon) = \{ f \in \mathcal{D}'(\Omega) : |\langle f, \phi_k \rangle| < \epsilon \text{ for } 1 \leq k \leq m \}$$

where $\phi_1, \ldots, \phi_m \in \mathcal{D}(\Omega)$ and $\epsilon > 0$. A net (f_i) in $\mathcal{D}'(\Omega)$ converges to $f \in \mathcal{D}'(\Omega)$ if and only if $\langle f_i, \phi \rangle \to \langle f, \phi \rangle$ for all $\phi \in \mathcal{D}(\Omega)$. This topology is discussed in more detail in ().

XXII.2.1.21. In references and applications, this topology is rarely if ever referred to explicitly; practically all use is in terms of convergence of sequences of generalized functions, i.e. a sequence (f_n) in $\mathcal{D}'(\Omega)$ converges to $f \in \mathcal{D}'(\Omega)$ if and only if $\langle f_n, \phi \rangle \to \langle f, \phi \rangle$ for all $\phi \in \mathcal{D}(\Omega)$; in most references, the topology is simply described via convergence of sequences. (Actually, although this topology is not first countable, it is completely determined, at least on bounded sets, by convergence of sequences just as the topology on $\mathcal{D}(\Omega)$, so the sequential approach is not as sloppy as it might appear.)

Multiplication of Generalized Functions by Smooth Functions

XXII.2.1.22. If f is an ordinary function and α is a $\mathcal{C}^\infty$ function on Ω, not necessarily of compact support or even bounded, we can define αf by usual function multiplication, and it is an ordinary function. If ϕ is a test function, then $\alpha\phi$ is also a test function, and

$$\langle \alpha f, \phi \rangle = \int_\Omega (\alpha f)\phi \, d\lambda = \int_\Omega f(\alpha\phi) \, d\lambda = \langle f, \alpha\phi \rangle \ .$$

Thus, if f is a generalized function, we can define αf as a generalized function by

$$\langle \alpha f, \phi \rangle = \langle f, \alpha\phi \rangle$$

for $\phi \in \mathcal{D}$. If $\phi_n \to \phi$ in $\mathcal{D}$, then all the ϕ_n are supported in some interval $[a, b]$, hence $\alpha\phi_n$ is also supported in $[a, b]$, and by repeated applications of the product rule and standard limit theorems we have $(\alpha\phi_n)^{(k)} \to (\alpha\phi)^{(k)}$ uniformly on $[a, b]$ for each k, i.e. $\alpha\phi_n \to \alpha\phi$ in $\mathcal{D}$, so

$$\langle \alpha f, \phi_n \rangle = \langle f, \alpha\phi_n \rangle \to \langle f, \alpha\phi \rangle = \langle \alpha f, \phi \rangle$$

and αf is a continuous functional. This product agrees with the pointwise product if f is an ordinary function, and αf is again an ordinary function.

XXII.2.1.23. Although the product of an ordinary function with a locally bounded measurable function is well defined and is another ordinary function, there is no reasonable definition of the product of a generalized function with an ordinary function which is not infinitely differentiable (), let alone the product of two generalized functions.

Differentiation of Generalized Functions

Ordinary functions do not have (ordinary) derivatives in general. Even a function which is differentiable everywhere need not have a derivative which is an ordinary function. However, remarkably, generalized functions have (generalized) derivatives of all orders if the "derivative" is properly defined.

XXII.2.1.24. To get started, suppose f is an ordinary function on Ω which is absolutely continuous on each closed bounded subinterval; then f' (defined a.e.) is an ordinary function (XIX.5.2.21.). If ϕ is a test function supported in $[a, b]$, then using Integration by Parts (XIX.5.5.2.) we have:

$$\langle f', \phi \rangle = \int_\Omega f'\phi \, d\lambda = \int_a^b f'\phi \, d\lambda$$

$$= f(b)\phi(b) - f(a)\phi(a) - \int_a^b f\phi' \, d\lambda = -\int_a^b f\phi' \, d\lambda = -\int_\Omega f\phi' \, d\lambda = -\langle f, \phi' \rangle$$

since $\phi(a) = \phi(b) = 0$.

We thus make the following definition:

XXII.2.1.25. Definition. Let f be a generalized function on Ω. Define a new generalized function Df on Ω by

$$\langle Df, \phi \rangle = -\langle f, \phi' \rangle$$

for $\phi \in \mathcal{D}(\Omega)$.

XXII.2.1.26. Since ϕ' is also a test function for every $\phi \in \mathcal{D}$, $\langle Df, \phi\rangle$ is defined. And since whenever $\phi_n \to \phi$ in $\mathcal{D}$, $\phi_n' \to \phi'$ in $\mathcal{D}$, the linear functional Df is (sequentially) continuous.

XXII.2.1.27. If f is differentiable (everywhere) on Ω in the ordinary sense, and f' is an ordinary function, then $Df = f'$ (regarded as a generalized function). More generally, if f is locally absolutely continuous on Ω, i.e. absolutely continuous on closed bounded subintervals of Ω, then Df in the distributional sense is an ordinary function which agrees (a.e.) with the ordinary derivative of f. But if f is not locally absolutely continuous, even if f is differentiable everywhere, the generalized function Df in this sense is not an ordinary function and does not agree with the ordinary derivative of f.

XXII.2.1.28. **Terminology:** There is an inconsistency in the literature about names for this type of derivative. The terms *distributional derivative* and *weak derivative* are used synonymously by most authors, but not always in the same sense: some authors use these terms only in the case where the derivative of an ordinary function in the distributional sense is again an ordinary function. Since it is inefficient to have two terms for the same thing, we will use *distributional derivative* in the general setting (i.e. every generalized function has a distributional derivative which is again a generalized function), and use the term *weak derivative* only in the case where the distributional derivative (and, therefore, the original function too) is an ordinary function. Thus an ordinary function always has a distributional derivative, but may or may not have a weak derivative (if it has a weak derivative, we say it is *weakly differentiable*).

Note also that weak derivatives, when they exist at all, are only determined a.e.; thus it is an abuse of language to talk about *the* weak derivative of a function. If g is a weak derivative for f, then any function equal to g a.e. is also a weak derivative for f (and in fact for any function equal to f a.e.)

XXII.2.1.29. The notation Df for the distributional derivative is in conflict with the notation of () which should cause no confusion. The term "weak derivative," as we use it, also has an unfortunate slight incongruity: we can have a function, e.g. (), which is differentiable but not weakly differentiable.

XXII.2.1.30. If f is locally absolutely continuous on Ω, then it is weakly differentiable on Ω, and its ordinary derivative (defined a.e.) is a weak derivative. Conversely, if f is an ordinary function which is weakly differentiable, then f is equal a.e. to a locally absolutely continuous function g whose ordinary derivative is a weak derivative for f (XXII.2.1.50.). Thus the notions of weak differentiability and weak derivative are (up to a.e. ambiguity) not really a new concept in the one-dimensional case. The situation is much more delicate in higher dimensions: a weakly differentiable function need not even be equal a.e. to a continuous function on any open set (). It is the higher-dimensional case where weak derivatives are especially important.

XXII.2.1.31. We can, of course, repeat the process to define derivatives of all orders: if f is a generalized function, we define $D^n f$ by

$$\langle D^n f, \phi\rangle = (-1)^n \langle f, \phi^{(n)}\rangle$$

for $\phi \in \mathcal{D}$. Thus every generalized function is infinitely differentiable in the distributional sense.

XXII.2.1.32. Examples. (i) Let f be the step function (Heaviside function) $\chi_{[0,+\infty)}$. If ϕ is a test function supported in $[a, b]$, with $a < 0 < b$, we have

$$\phi(0) = \phi(0) - \phi(a) = \int_a^0 \phi'\, d\lambda$$

by the Fundamental Theorem of Calculus and, since $\int_a^0 \phi'\,d\lambda + \int_0^b \phi'\,d\lambda = \int_a^b \phi'\,d\lambda = \phi(b) - \phi(a) = 0$, we have

$$\phi(0) = -\int_0^b \phi'\,d\lambda = -\int_a^b f\phi'\,d\lambda = -\langle f, \phi'\rangle = \langle Df, \phi\rangle$$

so Df is the Dirac δ-function.

(ii) More generally, if f is a measurable function of locally bounded variation, then f can be regarded as the cdf for a Radon measure μ (), and Df is the generalized function defined by μ (XXII.2.1.19.(ii)).

(iii) Let $f(x) = \log|x|$ $(x \neq 0)$. Then f is an ordinary function. If ϕ is a test function, supported in $[-n, n]$, and we set $\psi(x) = \phi(x) - \phi(-x)$ for $x \in \mathbb{R}$, we have

$$\langle Df, \phi\rangle = -\langle f, \phi'\rangle = -\int_{-n}^{n} f\phi'\,d\lambda$$

$$= -\int_{-n}^{0} f\phi'\,d\lambda - \int_0^n f\phi'\,d\lambda = -\lim_{\epsilon\to 0}\left[\int_{-n}^{-\epsilon} f\phi'\,d\lambda + \int_\epsilon^n f\phi'\,d\lambda\right]$$

$$= -\lim_{\epsilon\to 0}\left[\int_\epsilon^n \phi'(-x)\log x\,dx + \int_\epsilon^n \phi'(x)\log x\,dx\right]$$

$$= -\lim_{\epsilon\to 0}\int_\epsilon^n [\phi'(x) + \phi'(-x)]\log x\,dx = -\lim_{\epsilon\to 0}\int_\epsilon^n \psi'(x)\log x\,dx\ .$$

We can evaluate this last integral by parts to obtain

$$\int_\epsilon^n \psi'(x)\log x\,dx = \psi(n)\log n - \psi(\epsilon)\log\epsilon - \int_\epsilon^n \frac{\psi(x)}{x}\,dx\ .$$

We have $\psi(n) = 0$, and

$$\lim_{\epsilon\to 0}\psi(\epsilon)\log\epsilon = \lim_{\epsilon\to 0}\left[\frac{\phi(\epsilon) - \phi(-\epsilon)}{\epsilon}(\epsilon\log\epsilon)\right] = \left[\lim_{\epsilon\to 0}\frac{\phi(\epsilon) - \phi(-\epsilon)}{\epsilon}\right]\left[\lim_{\epsilon\to 0}\epsilon\log\epsilon\right] = 2\phi'(0)\cdot 0 = 0\ .$$

Thus we have

$$\langle Df, \phi\rangle = -\lim_{\epsilon\to 0}\int_\epsilon^n \psi'(x)\log x\,dx$$

$$= -\lim_{\epsilon\to 0}\left[-\psi(\epsilon)\log\epsilon - \int_\epsilon^n \frac{\psi(x)}{x}\,dx\right] = \lim_{\epsilon\to 0}\int_\epsilon^n \frac{\phi(x) - \phi(-x)}{x}\,dx = \left\langle \frac{1}{x}, \phi\right\rangle$$

so Df is the generalized function $\frac{1}{x}$ of XXII.2.1.19.(iii).

The Product Rule holds for the limited multiplication we have defined:

XXII.2.1.33. PROPOSITION. Let f be a generalized function, and α a $\mathcal{C}^\infty$ function. Then

$$D(\alpha f) = \alpha' f + \alpha(Df) \ .$$

PROOF: If ϕ is a test function, we have

$$\begin{aligned} \langle D(\alpha f), \phi\rangle &= -\langle \alpha f, \phi'\rangle = -\langle f, \alpha\phi'\rangle \\ &= -\langle f, (\alpha\phi)' - \alpha'\phi\rangle = -\langle f, (\alpha\phi)'\rangle + \langle f, \alpha'\phi\rangle \\ &= \langle Df, \alpha\phi\rangle + \langle f, \alpha'\phi\rangle = \langle \alpha(Df), \phi\rangle + \langle \alpha' f, \phi\rangle \ . \end{aligned}$$



A Unified Approach

Multiplication by smooth functions and differentiation are examples of a general class of extensions to distributions of operations on test functions.

XXII.2.1.34. PROPOSITION. Let Ω_1 and Ω_2 be regions, and $T : \mathcal{D}(\Omega_1) \to \mathcal{D}(\Omega_2)$ a (sequentially) continuous linear function. If there is a (sequentially) continuous (automatically linear) function $T^* : \mathcal{D}(\Omega_2) \to \mathcal{D}(\Omega_1)$ such that

$$\langle T\phi, \psi\rangle = \langle \phi, T^*\psi\rangle$$

for all $\phi \in \mathcal{D}(\Omega_1)$, $\psi \in \mathcal{D}(\Omega_2)$, then there is a unique (sequentially) continuous extension $\bar{T} : \mathcal{D}'(\Omega_1) \to \mathcal{D}'(\Omega_2)$ of T. The extension $\bar{T}$ satisfies

$$\langle \bar{T} f, \psi\rangle = \langle f, T^*\psi\rangle$$

for all $f \in \mathcal{D}'(\Omega_1)$, $\psi \in \mathcal{D}(\Omega_2)$. $\bar{T}$ is also linear.

PROOF: For $f \in \mathcal{D}'(\Omega_1)$, the formula defines a linear functional $\bar{T}f$ on $\mathcal{D}(\Omega_2)$, so we need only check that this linear functional is (sequentially) continuous, i.e. if $f \in \mathcal{D}'(\Omega_1)$ and $\psi_n \to \psi$ in $\mathcal{D}(\Omega_2)$, then $\langle \bar{T}f, \psi_n\rangle \to \langle \bar{T}f, \psi\rangle$. But

$$\langle \bar{T}f, \psi_n\rangle = \langle f, T^*\psi_n\rangle \to \langle f, T^*\psi\rangle = \langle \bar{T}f, \psi\rangle$$

by continuity of T^* and f.

Uniqueness of $\bar{T}$ follow from the density of $\mathcal{D}(\Omega_1)$ in $\mathcal{D}'(\Omega_1)$ () and continuity of T^*, and linearity is straightforward.



This result is often used in the case $\Omega_1 = \Omega_2$.

XXII.2.1.35. EXAMPLES. (i) Let Ω be a region and $\alpha \in \mathcal{C}^\infty(\Omega)$. Set $\Omega_1 = \Omega_2 = \Omega$ and $T\phi = \alpha\phi$. Then the hypotheses are satisfied with $T^* = T$, and $\bar{T}$ is multiplication by α in the sense of XXII.2.1.22..

(ii) Let $\Omega_1 = \Omega_2 = \Omega$, and $T\phi = \phi'$. Then the hypotheses are satisfied with $T^* = -T$, and $\bar{T}$ is distributional differentiation D.

(iii) Let $c \in \mathbb{R}$, and let Ω_1 be a region. Set $\Omega_2 = c + \Omega_1$ and $T_c\phi(x) = \phi(x - c)$. Then the hypotheses are satisfied with $T_c^* = T_{-c}$, and $\bar{T}_c$ is "translation by c."

(iv) Let Ω_1 be a region, and $\Omega_2 = -\Omega = \{-x : x \in \Omega\}$. Define $R : \Omega_1 \to \Omega_2$ by $(R\phi)(x) = \phi(-x)$. Then the hypotheses are satisfied with $R^* = R$, and $\bar{R}$ is "reflection."

We can alternately phrase distributional differentiation in usual terms via translation. We state this only for $\Omega = \mathbb{R}$, but the same can be done with care for general Ω.

XXII.2.1.36. PROPOSITION. let $f \in \mathcal{D}'(\mathbb{R})$. Then

$$Df = \lim_{\epsilon \to 0} \frac{1}{\epsilon} (T_\epsilon f - f) \ .$$

(where $\mathcal{D}'(\mathbb{R})$ has the topology of pointwise convergence, cf. XXII.2.1.20.).

Convolution

We now restrict to the case $\Omega = \mathbb{R}$.

XXII.2.1.37. Recall (XVIII.5.1.4.) that if ϕ and ψ are test functions, then the convolution $\phi * \psi$ is defined by

$$(\phi * \psi)(x) = \int_{-\infty}^{\infty} \phi(y)\psi(x - y)\, dy$$

and that $\phi * \psi$ is again a test function (XVIII.5.1.7., XVIII.5.1.8.). We have $\phi * \psi = \psi * \phi$ ().

XXII.2.1.38. PROPOSITION. Let (ϕ_n), ϕ, ψ be in $\mathcal{D}(\mathbb{R})$, with $\phi_n \to \phi$ in $\mathcal{D}(\mathbb{R})$. Then $\phi_n * \psi \to \phi * \psi$ in $\mathcal{D}(\mathbb{R})$.

PROOF: This follows immediately from XVIII.5.1.16., XVIII.5.1.7., and XVIII.5.1.8.. 👍

XXII.2.1.39. Let θ be a fixed element of $\mathcal{D}(\mathbb{R})$. Define $S_\theta : \mathcal{D}(\mathbb{R}) \to \mathcal{D}(\mathbb{R})$ by $S_\theta\phi = \theta * \phi$. Then S_θ is linear, and continuous by XXII.2.1.38.. If $\phi, \psi \in \mathcal{D}(\mathbb{R})$, we have

$$\langle S_\theta\phi, \psi \rangle = \int_{-\infty}^{\infty} (\phi * \theta)(x)\psi(x)\, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(y)\theta(x - y)\psi(x)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(y)[R\theta](y - x)\psi(x)\, dx\, dy = \int_{-\infty}^{\infty} \phi(y)([R\theta] * \phi)(y)\, dy = \langle \phi, S_{R\theta}\psi \rangle$$

(the integrals are actually only over a bounded rectangle). So the hypotheses of XXII.2.1.34. are satisfied with $S_\theta^* = S_{R\theta}$, and thus S_θ extends to $\bar{S}_\theta : \mathcal{D}'(\mathbb{R}) \to \mathcal{D}'(\mathbb{R})$. If $f \in \mathcal{D}'(\mathbb{R})$, we define $f * \theta = \bar{S}_\theta f$, i.e. for $\psi \in \mathcal{D}(\mathbb{R})$ we have

$$\langle f * \theta, \psi \rangle = \langle f, (R\theta) * \psi \rangle \ .$$

Thus the convolution of a generalized function with a test function is defined, extending usual convolution. We can also write $\theta * f = f * \theta$.

XXII.2.1.40. Continuity of $\bar{S}_\theta$ means that, if $f_n \to f$ in $\mathcal{D}'(\mathbb{R})$, then $f_n * \theta \to f * \theta$ in $\mathcal{D}'(\mathbb{R})$.

XXII.2.1.41. Proposition. Let $f \in \mathcal{D}'(\mathbb{R})$ and $\theta \in \mathcal{D}(\mathbb{R})$. Then $D(f * \theta) = (Df) * \theta$.

Proof: If $\psi \in \mathcal{D}(\mathbb{R})$, then

$$\langle D(f * \theta), \psi \rangle = -\langle f * \theta, \psi' \rangle = -\langle f, (R\theta) * (\psi') \rangle = -\langle f, [(R\theta) * \psi]' \rangle = \langle Df, (R\theta) * \psi \rangle = \langle (Df) * \theta, \psi \rangle$$

by (). 👍

XXII.2.1.42. Example. Let $\phi, \psi \in \mathcal{D}(\mathbb{R})$. Then

$$\langle \delta * \phi, \psi \rangle = \langle \delta, (R\phi) * \psi \rangle = [(R\phi) * \psi](0) = \int_{-\infty}^{\infty} \phi(-y)\psi(0-y)\, dy = \langle \phi, \psi \rangle$$

so $\delta * \phi = \phi$, i.e. convolution with δ is the identity on $\mathcal{D}(\mathbb{R})$.

We will show () that whenever $f \in \mathcal{D}'(\mathbb{R})$ and $\theta \in \mathcal{D}(\mathbb{R})$, we have that $f * \theta$ is actually an ordinary function which is $\mathcal{C}^\infty$, i.e. $f * \theta \in \mathcal{E}(\mathbb{R})$. It is in $\mathcal{D}(\mathbb{R})$ if and only if f has compact support (XXII.2.1.56.).

Integration of Generalized Functions on $\mathbb{R}$

We will show that every generalized function has an antiderivative in $\mathcal{D}'$, and two antiderivatives must differ by a constant.

XXII.2.1.43. Definition. Let

$$\mathcal{D}_0(\Omega) = \left\{ \phi \in \mathcal{D}(\Omega) : \int_\Omega \phi(x)\, dx = 0 \right\}$$

(denoted $\mathcal{D}_0$ when Ω is understood). $\mathcal{D}_0$ is a closed subspace of $\mathcal{D}$.

XXII.2.1.44. A function ϕ is in $\mathcal{D}_0$ if and only if $\phi = \psi'$ for some $\psi \in \mathcal{D}$: if $\psi \in \mathcal{D}$, supported in $[a,b]$, then

$$\int_\Omega \psi'(x)\,dx = \int_a^b \psi'(x)\,dx = \psi(b) - \psi(a) = 0$$

and, conversely, if $\phi \in \mathcal{D}_0$, supported on $[a,b]$, for $x \in \Omega$ set

$$\psi(x) = \int_a^x \phi(t)\,dt\ .$$

Then $\psi' = \phi$, and ψ is supported on $[a,b]$. In fact, each $\phi \in \mathcal{D}_0$ is the derivative of a *unique* $\psi \in \mathcal{D}$ (since such ψ must have compact support). Thus there is a linear map $\rho : \mathcal{D}_0 \to \mathcal{D}$ sending ϕ to ψ, which is continuous (i.e. $\psi \mapsto \psi'$ is a linear homeomorphism from $\mathcal{D}$ onto $\mathcal{D}_0$); $\rho(\psi') = \psi$ for any $\psi \in \mathcal{D}$.

XXII.2.1.45. There is a projection π from $\mathcal{D}$ onto $\mathcal{D}_0$, although not a canonical one. Fix a test function ϕ_1 with $\int_\Omega \phi_1(x)\,dx = 1$, and for $\phi \in \mathcal{D}$ set

$$\pi(\phi) = \phi - \left[\int_\Omega \phi(x)\,dx\right]\phi_1\ .$$

Then π is an idempotent linear map from $\mathcal{D}$ onto $\mathcal{D}_0$. If $\phi_n \to \phi$ in $\mathcal{D}$, then

$$\int_\Omega \phi_n(x)\,dx = \langle 1, \phi_n\rangle \to \langle 1, \phi\rangle = \int_\Omega \phi(x)\,dx$$

and it follows that π is continuous.

XXII.2.1.46. If $f \in \mathcal{D}'$, define $I(f)$ by

$$\langle I(f), \phi\rangle = -\langle f, \rho(\pi(\phi))\rangle\ .$$

Since π and ρ are continuous and linear, $I(f) \in \mathcal{D}'$. If $\phi \in \mathcal{D}$, we have

$$\langle DI(f), \phi\rangle = -\langle I(f), \phi'\rangle = \langle f, \rho(\pi(\phi'))\rangle = \langle f, \phi\rangle$$

since $\pi(\phi') = \phi'$ and $\rho(\phi') = \phi$. Thus $DI(f) = f$, i.e. $I(f)$ is an antiderivative for f.

XXII.2.1.47. Note that the generalized function $I(f)$ depends on the choice of π and hence ϕ_1, and thus is not canonical. However, by XXII.2.1.49., the ambiguity is only up to adding a constant. The next simple result is key, not just here but in the applications of distribution theory to differential equations generally:

XXII.2.1.48. Proposition. Let $f \in \mathcal{D}'$. If $Df = 0$, then f is a constant function (regarded as a generalized function).

Proof: If $\phi \in \mathcal{D}_0$, then $\phi = \psi'$ for a $\psi \in \mathcal{D}$, so

$$\langle f, \phi\rangle = \langle f, \psi'\rangle = -\langle Df, \psi\rangle = 0\ .$$

Set $C = \langle f, \phi_1\rangle$, where ϕ_1 is as in XXII.2.1.45.. Then, for general $\phi \in \mathcal{D}$, we have $\phi = \alpha\phi_1 + \pi(\phi)$, where $\alpha = \int_\Omega \phi(x)\,dx$, so

$$\langle f, \phi\rangle = \langle f, \alpha\phi_1\rangle + \langle f, \pi(\phi)\rangle = \alpha C = \langle C, \phi\rangle\ .$$



XXII.2.1.49. Corollary. Let $f, g \in \mathcal{D}'$. If $Df = Dg$, then f and g differ by a constant.

XXII.2.1.50. If f is an ordinary function, and F is an antiderivative for f in the ordinary sense (i.e. F is locally absolutely continuous and $F' = f$ a.e., cf. ()), then F is an antiderivative in the distribution sense. Thus if g is a generalized function with a weak derivative Dg which is an ordinary function, and G is an ordinary antiderivative for Dg, then $G = g$ plus a constant a.e., i.e. g is equal a.e. to a locally absolutely continuous function (in particular, g is an ordinary function) whose ordinary derivative is equal to Dg a.e.

Support of a Generalized Function

XXII.2.1.51. Recall that the (closed) support of a real- or complex-valued continuous function f is the closure of its cozero set, i.e.

$$supp(f) = \overline{\{x : f(x) \neq 0\}} \ .$$

The support of a discontinuous ordinary function is somewhat tricky to define directly (). But the support of a generalized function is easy to define:

XXII.2.1.52. Definition. Let f be a generalized function, and $E \subseteq \mathbb{R}$ a closed set. Then f is *supported on* E (or *carried on* E) if $\langle f, \phi \rangle = 0$ for every test function ϕ whose (closed) support is disjoint from E. The smallest closed set on which f is supported is called the *support* of f.

For this to make sense, it must be shown that there is a smallest such set. By the next result, the support does indeed exist and is the intersection of all closed sets on which f is supported (there is at least one such set, $\mathbb{R}$).

XXII.2.1.53. Proposition. Let f be a generalized function, and $\{E_i : i \in I\}$ a collection of closed sets in $\mathbb{R}$. If f is supported on each E_i, then f is supported on $E = \cap_{i \in I} E_i$.

Proof: Let ϕ be a test function whose support K is disjoint from E. The open sets $U_i = \mathbb{R} \setminus E_i$ cover K. Since K is compact, a finite number $U_{i_1}, \ldots, U_{i_n}$ cover K. The sets $\{U_{i_1}, \ldots, U_{i_n}, \mathbb{R} \setminus K\}$ are a finite open cover of $\mathbb{R}$. Let $\{h_1, \ldots, h_{n+1}\}$ be a smooth partition of unity subordinate to a shrinking of this open cover (), and set $\phi_j = h_j \phi$. Then the ϕ_j are test functions, and $\phi = \phi_1 + \cdots + \phi_{n+1}$. We have $\phi_{n+1} = 0$ since the support of h_{n+1} is disjoint from K. For $1 \leq j \leq n$ we have $\langle f, \phi_j \rangle = 0$ since the support of ϕ_j is disjoint from E_{i_j}. Thus $\langle f, \phi \rangle = 0$ and f is supported on E. 👍

XXII.2.1.54. Thus the support of a generalized function is defined. Note that the support of a nonzero generalized function is nonempty (the support of 0 is $\emptyset$). If f is a continuous ordinary function, the support of f as a generalized function is just the (closed) support of f as a continuous function. More generally, if f is an ordinary function, then the support of f as a generalized function is the complement of the largest open set U for which $f = 0$ a.e. on U (in fact, the support of a Radon measure μ, regarded as a generalized function, is the complement of the largest open μ-null set; cf. XIII.1.6.6. for existence of such an open set). Thus the support of a function need not be compact. If ϕ is a test function, the (closed) support of ϕ' is contained in the support of ϕ, so if f is a generalized function, the support of f' is contained in the support of f.

XXII.2.1.55. EXAMPLES. (i) The support of δ, or $\delta^{(k)}$ for any k, is $\{0\}$.

(ii) It is not true that $\langle f, \phi\rangle = 0$ in general if ϕ just vanishes on the support of f. For example, if ϕ is a test function with $\phi(0) = 0$ but $\phi'(0) \neq 0$, then $\langle \delta', \phi\rangle \neq 0$. However, if $\phi^{(n)}$ vanishes on the support of f for all n, it is true that $\langle f, \phi\rangle = 0$, since such a ϕ is a limit in $\mathcal{D}$ of a sequence of test functions vanishing in a neighborhood of $supp(f)$.

Distributions with Compact Support

The notion of support of a generalized function is arguably more interesting for generalized functions on $\mathbb{R}^n$, $n > 1$, since in that case a generalized function can be supported on a curve or surface. But even on $\mathbb{R}$, generalized functions with compact support are especially important.

XXII.2.1.56. If f is a generalized function on Ω with compact support S, and $\phi \in \mathcal{D}(\Omega)$, then $\langle f, \phi\rangle$ depends only on the values of ϕ in a neighborhood of S, not on all of Ω: if Λ is an open neighborhood of S contained in Ω, and ϕ and ψ are test functions with $\phi = \psi$ on Λ, then $\psi = \phi + (\psi - \phi)$, and the support of $\psi - \phi$ is disjoint from S, so $\langle f, \psi - \phi\rangle = 0$ and

$$\langle f, \psi\rangle = \langle f, \phi\rangle + \langle f, \psi - \phi\rangle = \langle f, \phi\rangle \ .$$

XXII.2.1.57. The distribution f does not even depend in any essential way on Ω, and can be extended in a natural way to a much larger class of "test functions." Let Λ be an open neighborhood of S whose closure is compact and contained in Ω, and let h be a $\mathcal{C}^\infty$ function supported in Λ and identically 1 on a smaller neighborhood of S. If ψ is any $\mathcal{C}^\infty$ function on $\mathbb{R}$, then $\phi = h\psi \in \mathcal{D}(\Omega)$ and $(1-h)\psi$ is a $\mathcal{C}^\infty$ function on $\mathbb{R}$ whose support is disjoint from S, so it makes sense to define $\langle f, (1-h)\psi\rangle = 0$ and, since $\psi = \phi + (1-h)\psi$,

$$\langle f, \psi\rangle = \langle f, \phi\rangle + \langle f, (1-h)\psi\rangle = \langle f, \phi\rangle \ .$$

If (ψ_n) is a sequence in $\mathcal{C}^\infty(\mathbb{R})$, and there is a $\mathcal{C}^\infty$ function ψ on $\mathbb{R}$ such that $\psi_n^{(k)} \to \psi^{(k)}$ uniformly on Λ for each k, irrespective of what values the ψ_n and ψ take outside Λ, set $\phi_n = h\psi_n$ and $\phi = h\psi$; then $\phi_n, \phi \in \mathcal{D}(\Omega)$ and $\phi_n \to \phi$, so

$$\langle f, \psi_n\rangle = \langle f, \phi_n\rangle \to \langle f, \phi\rangle = \langle f, \psi\rangle \ .$$

Thus the extended linear functional f has a much broader continuity property: it is continuous for the relative topology on $\mathcal{D}(\Omega)$ from the Fréchet topology of $\mathcal{E}(\Omega)$, and extends uniquely to a continuous linear functional on $\mathcal{E}(\Omega)$. Conversely, the restriction to $\mathcal{D}(\Omega)$ of a continuous linear functional on $\mathcal{E}(\Omega)$ is in $\mathcal{D}'(\Omega)$. In fact, we have:

XXII.2.1.58. PROPOSITION. Let Ω be a region. The restrictions to $\mathcal{D}(\Omega)$ of continuous linear functionals on $\mathcal{E}(\Omega)$ are precisely the generalized functions on Ω of compact support.

XXII.2.1.59. Thus we can identify the space $\mathcal{E}'(\Omega)$ of continuous linear functionals on $\mathcal{E}(\Omega)$ with the generalized functions on Ω of compact support.

Topology on $\mathcal{D}'(\Omega)$

XXII.2.1.60. We now discuss the topology on $\mathcal{D}'(\Omega)$ (XXII.2.1.20.) in more detail.

A simple but important property of this topology is that differentiation is continuous:

XXII.2.1.61. PROPOSITION. Let (f_i) be a net in $\mathcal{D}'(\Omega)$. If $f_i \to f$ in $\mathcal{D}'(\Omega)$, then $Df_i \to Df$.

PROOF:

$$\langle Df_i, \phi\rangle = -\langle f_i, \phi'\rangle \to -\langle f, \phi'\rangle = \langle Df, \phi\rangle$$

for any $\phi \in \mathcal{D}(\Omega)$. 👍

By iteration, if $f_i \to f$, then $D^k f_i \to D^k f$ for any k.

XXII.2.1.62. If (f_n) is a sequence (even net) of ordinary functions on Ω, and $f_n \to f$ u.c., then $f_n \to f$ in $\mathcal{D}'(\Omega)$. But the restriction of the topology of $\mathcal{D}'(\Omega)$, even to $\mathcal{C}^\infty(\Omega)$, is far weaker than the topology of u.c. convergence; see XXII.2.1.63..

XXII.2.1.63. EXAMPLES. (i) Let (f_n) be a sequence of locally absolutely continuous functions on Ω, converging u.c. to f on Ω. Then the sequence (f_n') converges to Df in $\mathcal{D}'(\Omega)$. If f is also absolutely continuous, then $f_n' \to f'$ in $\mathcal{D}'(\Omega)$. Note that (f_n') does not converge to f' in any usual functional sense (cf. I.6.1.18.).

(ii) Here is an interesting example of (i). Let $f_n(x) = \frac{1}{n}\sin nx$. Then $f_n \to 0$ uniformly on $\mathbb{R}$. Thus the sequence $f_n'(x) = \cos nx$ converges to 0 in $\mathcal{D}'(\mathbb{R})$, i.e.

$$\lim_{n\to\infty}\left(\int_{\mathbb{R}} \phi(x)\cos nx\, dx\right) = 0$$

for any $\phi \in \mathcal{D}(\mathbb{R})$ (this can be verified directly by integration by parts). This is a version of the Riemann-Lebesgue Lemma (). Note that (f_n') does not converge to 0 in any ordinary functional sense, even pointwise.

Perhaps the most important property of $\mathcal{D}'(\Omega)$ with this topology is that it is (sequentially) complete:

XXII.2.1.64. THEOREM. Let (f_n) be a sequence in $\mathcal{D}'(\Omega)$, and suppose the sequence $(\langle f_n, \phi\rangle)$ converges (i.e. is a Cauchy sequence) for every $\phi \in \mathcal{D}(\Omega)$. Then there is a $f \in \mathcal{D}'(\Omega)$ with $f_n \to f$.

XXII.2.1.65. The definition of f is obvious: if $\phi \in \mathcal{D}(\Omega)$, define

$$\langle f, \phi\rangle = \lim_{n\to\infty}\langle f_n, \phi\rangle\ .$$

It is easily verified that this f is a linear functional on $\mathcal{D}(\Omega)$. We need to show that f is continuous, i.e. $f \in \mathcal{D}'(\Omega)$; then $f_n \to f$ by definition. For this we need a Lemma which is perhaps of independent interest:

XXII.2.1.66. LEMMA. Let (f_n) be a sequence in $\mathcal{D}(\Omega)$ such that $(\langle f_n, \phi\rangle)$ converges for every $\phi \in \mathcal{D}(\Omega)$. Notationally write

$$\langle f, \phi\rangle = \lim_{n\to\infty} \langle f_n, \phi\rangle \ .$$

If (ϕ_n) is a sequence in $\mathcal{D}(\Omega)$ with $\phi_n \to \phi$ in $\mathcal{D}(\Omega)$, then the sequence $(\langle f_n, \phi_n\rangle)$ converges to $\langle f, \phi\rangle$.

XXII.2.1.67. Using the Lemma, we prove that the limit functional is continuous, completing the proof of XXII.2.1.64..

PROOF: It suffices to show that if $\phi_n \to 0$ in $\mathcal{D}(\Omega)$, then $\langle f, \phi_n\rangle \to 0$. If not, passing to a subsequence, there is an $\epsilon > 0$ such that $|\langle f, \phi_n\rangle| > 2\epsilon$ for all n. Then for each n there is an m_n such that $|\langle f_{m_n}, \phi_n\rangle| > \epsilon$. But the subsequence $\langle f_{m_n}, \phi\rangle$ converges to $\langle f, \phi\rangle$ for any ϕ, so by the Lemma $\langle f_{m_n}, \phi_n\rangle \to \langle f, 0\rangle = 0$, a contradiction. 👍

XXII.2.1.68. It remains to prove Lemma XXII.2.1.66..

Infinite Series

XXII.2.1.69. Convergence of infinite series in $\mathcal{D}(\Omega)$ and in $\mathcal{D}'(\Omega)$ has the evident meaning: convergence of the sequence of partial sums. Thus in $\mathcal{D}(\Omega)$ we have

$$\phi = \sum_{k=1}^{\infty} \phi_k$$

if and only if all the ϕ_k are supported in the same compact subinterval of Ω and $\sum_{k=1}^{\infty} \phi_k^{(m)}$ converges uniformly to $\phi^{(m)}$ for all m.

In $\mathcal{D}'(\Omega)$ we have

$$f = \sum_{k=1}^{\infty} f_k$$

if and only if

$$\sum_{k=1}^{\infty} \langle f_k, \phi\rangle$$

converges to $\langle f, \phi\rangle$ for all $\phi \in \mathcal{D}(\Omega)$. By XXII.2.1.64., if

$$\sum_{k=1}^{\infty} \langle f_k, \phi\rangle$$

converges for all $\phi \in \mathcal{D}(\Omega)$, then the series $\sum_{k=1}^{\infty} f_k$ converges in $\mathcal{D}'(\Omega)$.

XXII.2.1.70. Because of the linearity and continuity of differentiation, term-by-term differentiation is valid in both $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$: if $\phi = \sum_{k=1}^{\infty} \phi_k$ in $\mathcal{D}(\Omega)$, then $\sum_{k=1}^{\infty} \phi_k^{(m)}$ converges to $\phi^{(m)}$ in $\mathcal{D}(\Omega)$ for any m, and if $f = \sum_{k=1}^{\infty} f_k$ in $\mathcal{D}'(\Omega)$, then $\sum_{k=1}^{\infty} D^m f_k$ converges to $D^m f$ in $\mathcal{D}'(\Omega)$ for any m.

We can decompose an arbitrary distribution as a convergent series of distributions of compact support whose supports "go off to infinity":

XXII.2.1.71. PROPOSITION. Let f be a distribution on $\Omega \subseteq \mathbb{R}$. Then there is a convergent series

$$f = \sum_{k=1}^{\infty} f_k$$

where the f_k have compact support S_k and any compact subset of Ω meets only finitely many S_k.

PROOF: Cover Ω by a sequence $\{I_n : n \in \mathbb{Z}\}$ of open intervals such that $\bar{I}_n$ is compact and contained in Ω, and so that only consecutive $\bar{I}_n$ overlap (), and let $\{h_n : n \in \mathbb{Z}\}$ be a smooth partition of unity subordinate to the I_n (). Set $f_n = h_n f$. Then $h_n f$ is a distribution on Ω whose support is contained in $\bar{I}_n$. If C is a compact subset of Ω, then only finitely many I_n meet C. If $\phi \in \mathcal{D}(\Omega)$, then

$$\langle f, \phi \rangle = \sum_{n \in \mathbb{Z}} \langle f, h_n \phi \rangle = \sum_{n \in \mathbb{Z}} \langle h_n f, \phi \rangle$$

(the sums converge since only finitely many terms are nonzero), so

$$f = \sum_{n \in \mathbb{Z}} h_n f \ .$$



XXII.2.2. Order of Generalized Functions and a Sequential Approach

XXII.2.2.1. DEFINITION. Let Ω be an open interval in $\mathbb{R}$, and $f \in \mathcal{D}'(\Omega)$. Then f has *order* $\leq m$ if there are ordinary functions $f_0, f_1, \ldots, f_m$ such that

$$\langle f, \phi \rangle = \sum_{k=0}^{m} \langle f_k, \phi^{(k)} \rangle$$

for all $\phi \in \mathcal{D}(\Omega)$. The distribution f has *order* m if it has order $\leq m$ but does not have order $\leq m-1$. Write $s(f) = m$ if f has order m.

XXII.2.2.2. In the one-dimensional case, this definition can be simplified. Note that if f is an ordinary function and F is any antiderivative of f, then $\langle f, \phi\rangle = -\langle F, \phi'\rangle$. So if $f \in \mathcal{D}'(\Omega)$ has order $\leq m$ via ordinary functions $f_0, \ldots, f_m$, and for each k we let F_k be a function with $F_k^{(m-k)} = f_k$, then for any ϕ we have

$$\langle f, \phi\rangle = \sum_{k=0}^{m} \langle f_k, \phi^{(k)}\rangle = \sum_{k=0}^{m} (-1)^{m-k} \langle F_k, \phi^{(m)}\rangle = \left\langle \sum_{k=0}^{m} (-1)^{m-k} F_k, \phi^{(m)} \right\rangle$$

i.e. there is an ordinary function F such that $\langle f, \phi\rangle = \langle F, \phi^{(m)}\rangle$, and it follows that f is the m'th distributional derivative of an ordinary function. (Things are more complicated in the multidimensional case.)

XXII.2.2.3. Examples. (i) A distribution has order 0 if and only if it is an ordinary function.

(ii) If f is an ordinary function, then $D^m f$ has order $\leq m$ for any m, and will have order $\leq m-1$ if and only if f is absolutely continuous; if f' is absolutely continuous it will have order $\leq m-2$, etc.

(iii) As a special case of (ii) δ has order ≤ 1 by (), but does not have order 0 by (i), so $s(\delta) = 1$.

(iv) More generally, if μ is a Radon measure on $\mathbb{R}$, regarded as a generalized function as in (), then $s(\mu) \leq 1$; $s(\mu) = 0$ if and only if $\mu \ll \lambda$ by the Radon-Nikodým Theorem.

(v) The distribution of () on $\mathbb{R}$ does not have finite order. There is a similar example on any Ω, taking a strictly increasing sequence (x_k) approaching the right end of Ω and setting

$$\langle f, \phi\rangle = \sum_{k=1}^{\infty} \phi^{(k)}(x_k)$$

for $\phi \in \mathcal{D}(\Omega)$ (this is a finite sum for any such ϕ).

XXII.2.2.4. By XXII.2.2.3.(v), a distribution on Ω need not have finite order. But it is a remarkable and significant fact that a distribution on Ω always has locally finite order (XXII.2.2.6.).

XXII.2.2.5. If Ω and Λ are open intervals in $\mathbb{R}$, with $\Lambda \subseteq \Omega$, there is a restriction map from $\mathcal{D}'(\Omega)$ to $\mathcal{D}'(\Lambda)$ (since $\mathcal{D}(\Lambda)$ is a subspace of $\mathcal{D}(\Omega)$), which is neither injective nor surjective if $\Lambda \neq \Omega$. We will be especially interested in the case where Λ is *compactly contained* in Ω, i.e. the closure of Λ is bounded and contained in Ω (we write $\Lambda \subseteq_c \Omega$; notations such as $\Lambda \subset\subset \Omega$ are also sometimes used).

We will prove the following theorem:

XXII.2.2.6. Theorem. Let $\Lambda \subseteq_c \Omega$ and $f \in \mathcal{D}'(\Omega)$. Then the restriction of f to Λ has finite order.

XXII.2.2.7. The idea of the proof is to show that the restriction of f to Λ is not only in $\mathcal{D}'(\lambda)$, i.e. continuous for the usual topology on $\mathcal{D}(\Lambda)$, but actually continuous for a weaker topology determined by only finitely many derivatives. This will be established using two lemmas.

XXII.2.2.8. LEMMA. Let $\Lambda = (a,b)$ be compactly contained in Ω, and $f \in \mathcal{D}'(\Omega)$. Then there is an $m \in \mathbb{N}$ and a constant C such that

$$|\langle f, \phi\rangle| \leq C \sum_{k=0}^{m} \|\phi^{(k)}\|_\infty$$

for all $\phi \in \mathcal{D}(\Lambda)$.

PROOF: Suppose not. Then for each m there is a $\phi_m \in \mathcal{D}(\Lambda)$ such that

$$|\langle f, \phi_m\rangle| > m \sum_{k=0}^{m} \|\phi_m^{(k)}\|_\infty \ .$$

This inequality is unchanged if ϕ_m is multiplied by a nonzero constant, so we may assume that

$$\sum_{k=0}^{m} \|\phi_m^{(k)}\|_\infty = \frac{1}{m}$$

by renormalizing ϕ_m; thus $|\langle f, \phi_m\rangle| > 1$ for all m. Then $\phi_m \to 0$ in $\mathcal{D}'(\Omega)$ since the ϕ_m are all supported in $\bar{\Lambda}$ and $\lim_{m\to\infty} \|\phi_m^{(k)}\|_\infty = 0$ for all k. Thus $\langle f, \phi_m\rangle \to 0$ since f is continuous, a contradiction. 👍

XXII.2.2.9. LEMMA. Let $\Lambda = (a,b)$ be compactly contained in Ω, and $f \in \mathcal{D}'(\Omega)$. Then there is an $m \in \mathbb{N}$ and a constant C_2 such that

$$|\langle f, \phi\rangle| \leq C_2 \sum_{k=1}^{m+1} \|\phi^{(k)}\|_2$$

for all $\phi \in \mathcal{D}(\Lambda)$, where

$$\|\phi\|_2 = \left[\int_a^b \phi^2 \, d\lambda\right]^{1/2} \ .$$

PROOF: Let m and C be as in XXII.2.2.8., and let $\phi \in \mathcal{D}(\Lambda)$. For $0 \leq k \leq m$, we have, for all $x \in [a,b]$,

$$\phi^k(x) = \int_a^x \phi^{(k+1)} \, d\lambda \ .$$

Applying the CBS inequality with the function $\phi^{(k+1)}$ and the constant function 1, we have, for $x \in [a,b]$,

$$|\phi^{(k)}(x)|^2 = \left|\int_a^x 1 \cdot \phi^{(k+1)} \, d\lambda\right|^2 \leq \left[\int_a^x 1^2 \, d\lambda\right] \left[\int_a^x (\phi^{(k+1)})^2 \, d\lambda\right]$$

$$\leq \left[\int_a^b 1^2 \, d\lambda\right] \left[\int_a^b (\phi^{(k+1)})^2 \, d\lambda\right] = (b-a)\|\phi^{(k+1)}\|_2^2$$

and thus

$$\|\phi^{(k)}\|_\infty \leq \sqrt{b-a}\|\phi^{(k+1)}\|_2$$

$$|\langle f, \phi\rangle| \leq C \sum_{k=0}^{m} \|\phi^{(k)}\|_\infty \leq C\sqrt{b-a} \sum_{k=1}^{m+1} \|\phi^{(k)}\|_2 \ .$$



We now prove Theorem XXII.2.2.6..

PROOF: Let $\Lambda = (a, b) \subseteq_c \Omega$, and let m and C_2 be as in Lemma XXII.2.2.9.. Put an inner product on $\mathcal{D}(\Lambda)$ by

$$\langle \phi, \psi\rangle = \sum_{k=1}^{m+1} \left[\int_a^b \phi^{(k)} \psi^{(k)} \, d\lambda \right]$$

and let $\|\cdot\|$ be the corresponding norm. Then f is a bounded linear functional on $\mathcal{D}(\Lambda), \|\cdot\|)$ by Lemma XXII.2.2.9., so extends to a bounded linear functional on the completion $\mathcal{H}$. We may identify $\mathcal{H}$ with a subspace of

$$\bigoplus_{k=1}^{m+1} L^2([a, b])$$

with its natural Hilbert space norm. By the Riesz Representation Theorem (), there is a vector $\xi \in \mathcal{H}$ such that

$$\langle f, \eta\rangle = \langle \xi, \eta\rangle$$

for all $\eta \in \mathcal{H}$ (where $\langle\cdot,\cdot\rangle$ on the left is the action of f on $\mathcal{H}$ and the $\langle\cdot,\cdot\rangle$ on the right is the inner product in $\mathcal{H}$). We have

$$\eta = (f_1, \ldots, f_{n+1})$$

for some $f_1, \ldots, f_{m+1} \in \mathcal{L}^2([a, b]) \subseteq \mathcal{L}^1([a, b])$, i.e.

$$\langle f, \phi\rangle = \sum_{k=1}^{m+1} \langle f_k, \phi^{(k)}\rangle \ .$$



Putting together XXII.2.2.6. and XXII.2.2.2., we can rephrase the result:

XXII.2.2.10. THEOREM. Let f be a distribution on $\Omega \subseteq \mathbb{R}$, and Λ an open interval compactly contained in Ω. Then there is an $m \in \mathbb{N} \cup \{0\}$ and an absolutely continuous function F on $\bar{\Lambda}$ such that the restriction of f to Λ is $D^m F$ in $\mathcal{D}'(\Lambda)$.

XXII.2.2.11. Note that under the hypotheses of XXII.2.2.10.:

(i) If the conclusion holds for some m, the F is not quite unique, but it is unique up to adding a polynomial of degree $< m$ (regarded as an ordinary function), by repeated applications of XXII.2.1.49..

(ii) If the conclusion holds for some m, it also holds for all larger m by just taking antiderivatives. The smallest such m is one more than the order of f as defined in XXII.2.2.1..

XXII.2.2.12. If f has compact support S in Ω, and Λ is an open neighborhood of S compactly contained in Ω, let $h \in \mathcal{D}(\Lambda)$ with h identically 1 on a smaller neighborhood of S. Then the restriction of f to Λ is hf, and the extension of hf to $\mathcal{C}^\infty(\mathbb{R})$ (XXII.2.1.57.) is the same as the extension of f. The restriction of f to Λ has finite order, and it follows that f itself has finite order, i.e. is the m'th distributional derivative of an absolutely continuous function, for some m.

Using XXII.2.1.71., we obtain:

XXII.2.2.13. Corollary. Let Ω be an open interval in $\mathbb{R}$, and $f \in \mathcal{D}'(\Omega)$. Then there is a sequence (F_k) of absolutely continuous functions on Ω (some of the F_k may be 0) and a sequence (m_k) in $\mathbb{N} \cup \{0\}$ such that

(i) Each F_k has compact support $S_k \subseteq \Omega$.

(ii) Any compact subset of Ω meets only finitely many S_k.

(iii) The infinite series

$$\sum_{k=0}^{\infty} D^{m_k} F_k$$

converges to f in $\mathcal{D}'(\Omega)$.

Conversely, any such infinite series with the F_k continuous (or just locally integrable) and satisfying (i) and (ii) converges to an element of $\mathcal{D}'(\Omega)$.

We can extend the finite order to convergent sequences:

XXII.2.2.14. Theorem. Let Ω be an open interval in $\mathbb{R}$, and (f_n), f with $f_n \to f$ in $\mathcal{D}'(\Omega)$. If $\Lambda \subseteq_c \Omega$, then there is a sequence (F_n) of absolutely continuous functions and an absolutely continuous F, and an m such that:

(i) All the F_n and F are supported on a compact set $C \subseteq \Omega$ containing Λ.

(ii) $F_n \to F$ uniformly on C.

(iii) The restrictions of $D^m F_n$ and $D^m F$ to Λ agree with the restrictions of f_n and f to Λ.

In particular, the restrictions of the f_n and f to Λ have uniformly bounded order.

Thus every convergent sequence in $\mathcal{D}'(\Omega)$ is locally the sequence of derivatives of some fixed order of a uniformly convergent sequence of (absolutely) continuous functions. (This is only true for *sequences*, not nets! The distinction is related to the fact that a weak-* convergent sequence in the dual of a Banach space must be uniformly bounded (XX.8.4.7.), but a weak-* convergent net need not be. In fact, even a convergent net of real numbers is not necessarily bounded.)

This result is equivalent to:

XXII.2.2.15. THEOREM. Let Ω be an open interval in $\mathbb{R}$, and $(f_n) \in \mathcal{D}'(\Omega)$ with $f_n \to 0$ in $\mathcal{D}'(\Omega)$. If $\Lambda \subseteq_c \Omega$, then there is a sequence (F_n) of absolutely continuous functions, and an m such that:

(i) All the F_n are supported on a compact set $C \subseteq \Omega$ containing Λ.

(ii) $F_n \to 0$ uniformly on C.

(iii) The restrictions of $D^m F_n$ to Λ agree with the restrictions of f_n to Λ.

In particular, the restrictions of the f_n to Λ have uniformly bounded order.

XXII.2.2.16. Theorem XXII.2.2.15. is not an immediate special case of XXII.2.2.14., but follows easily: under the hypotheses of XXII.2.2.15. we get (F_n) and F satisfying the conclusion of XXII.2.2.14., with $D^m F = 0$, so we can replace F_n by $F_n - F$ to get XXII.2.2.15.. Conversely, under the hypotheses of XXII.2.2.14., applying XXII.2.2.15. to the sequence $f_n - f$ we get an m_1 and a sequence $G_n \to 0$ with $D^{m_1} G_n = f_n - f$. Also, by XXII.2.2.10. there is an m_2 and F so that $f = D^{m_2} F$ on Λ. We may increase m_1 or m_2 by taking suitable antiderivatives, so we may assume $m_1 = m_2 = m$. Then we may take $F_n = G_n + F$.

Thus it suffices to prove XXII.2.2.15.. The proof follows the scheme of the proof of XXII.2.2.6., using modified versions of the two lemmas.

XXII.2.2.17. LEMMA. Let $\Lambda = (a, b)$ be compactly contained in Ω, and (f_n) a sequence in $\mathcal{D}'(\Omega)$ converging to 0. Then there is an $m \in \mathbb{N}$ and a constant C such that

$$|\langle f_n, \phi \rangle| \leq C \sum_{k=0}^{m} \|\phi^{(k)}\|_\infty$$

for all $\phi \in \mathcal{D}(\Lambda)$ and all n.

PROOF: Suppose not. Then for each m there is a $\phi_m \in \mathcal{D}(\Lambda)$ such that

$$|\langle f_n, \phi_m \rangle| > m \sum_{k=0}^{m} \|\phi_m^{(k)}\|_\infty$$

for infinitely many n. This inequality is unchanged if ϕ_m is multiplied by a nonzero constant, so we may assume that

$$\sum_{k=0}^{m} \|\phi_m^{(k)}\|_\infty = \frac{1}{m}$$

by renormalizing ϕ_m; thus, for each m, $|\langle f_n, \phi_m \rangle| > 1$ for infinitely many n. Inductively choose a strictly increasing sequence (n_m) so that $|\langle f_{n_m}, \phi_m \rangle| > 1$ for all m. Then $\phi_m \to 0$ in $\mathcal{D}'(\Omega)$ since the ϕ_m are all supported in $\bar{\Lambda}$ and $\lim_{m\to\infty} \|\phi_m^{(k)}\|_\infty = 0$ for all k. Thus $\langle f_{n_m}, \phi_m \rangle \to 0$ by XXII.2.1.66., a contradiction. ☝

XXII.2.2.18. LEMMA. Let $\Lambda = (a,b)$ be compactly contained in Ω, and (f_n) a sequence in $\mathcal{D}'(\Omega)$ converging to 0. Then there is an $m \in \mathbb{N}$ and a constant C_2 such that

$$|\langle f_n, \phi\rangle| \leq C_2 \sum_{k=1}^{m+1} \|\phi^{(k)}\|_2$$

for all $\phi \in \mathcal{D}(\Lambda)$ and all n, where

$$\|\phi\|_2 = \left[\int_a^b \phi^2 \, d\lambda\right]^{1/2}.$$

PROOF: The proof is essentially identical to the proof of XXII.2.2.9.. 👍

XXII.2.2.19. Thus, if we let $\mathcal{H}$ be the completion of $\mathcal{D}(\Lambda)$ with respect to the inner product

$$\langle \phi, \psi\rangle = \sum_{k=1}^{m+1} \left[\int_a^b \phi^{(k)}\psi^{(k)} \, d\lambda\right]$$

the sequence (f_n), regarded as a sequence in $\mathcal{H}$, is uniformly bounded and converges to 0 weakly, i.e. $\langle f_n, \eta\rangle \to 0$ for any $\eta \in \mathcal{H}$ (it suffices to check this for $\eta = \phi \in \mathcal{D}(\Lambda)$, where it is true by assumption).

Convolutions Revisited

A significant consequence of the local finite order of distributions is:

XXII.2.2.20. THEOREM. Let $f \in \mathcal{D}'(\mathbb{R})$ and $\phi \in \mathcal{D}(\mathbb{R})$. Then $f * \phi$ is an ordinary function which is $\mathcal{C}^\infty$.

PROOF: It suffices to show that the restriction of $f * \phi$ to any interval $[c,d]$ is an ordinary function which is $\mathcal{C}^\infty$. As in XVIII.5.1.11., the restriction of $f * \phi$ to $[c,d]$ is the same as the restriction of $g * \phi$ to $[c,d]$, where g is the restriction of f to a suitable finite interval. So we may assume f has compact support, and thus finite order, i.e. there is a continuous function F and $k \geq 0$ such that $f = D^k F$. Then $f * \phi = D^k(F * \phi)$. But $F * \phi$ is $\mathcal{C}^\infty$ (XVIII.5.1.8.), so $D^k(F * \phi) = (F * \phi)^{(k)}$ is also $\mathcal{C}^\infty$. 👍

There is a direct proof of this theorem not using (explicitly) the finite order result [].

XXII.2.2.21. DEFINITION. An *ai-sequence* is a sequence (ϕ_n) of nonnegative functions in $\mathcal{D}(\mathbb{R})$ such that

(i) ϕ_n is supported on $[-\delta_n, \delta_n]$ for some sequence (δ_n) with $0 < \delta_n \leq 1$ for all n and $\lim_{n\to\infty} \delta_n = 0$.

(ii) $\displaystyle\int_{-\infty}^{\infty} \phi_n(x)\, dx = \int_{-\delta_n}^{\delta_n} \phi_n(x)\, dx = 1.$

The requirement that the ϕ_n be nonnegative is not essential, but simplifies some arguments.

XXII.2.2.22. PROPOSITION. Let (ϕ_n) be an ai-sequence, and $\psi \in \mathcal{D}(\mathbb{R})$. Then $\phi_n * \psi \to \psi$ in $\mathcal{D}(\mathbb{R})$.

PROOF: Fix ψ. If ψ is supported on $[a,b]$, then $\phi_n * \psi^{(k)}$ is supported on $[a-1, b+1]$ for all n and k. Fix k. If $\epsilon > 0$, then since $\psi^{(k)}$ is uniformly continuous there is a $\delta > 0$ [not the Dirac delta function!] such that $|\psi^{(k)}(x) - \psi^{(k)}(y)| < \epsilon$ whenever $|x-y| < \delta$. Fix N such that $\delta_n < \delta$ for all $n \geq N$. If $n \geq N$, then, for any x, we have

$$\psi^{(k)}(x) - \epsilon = \int_{-\delta_n}^{\delta_n} \phi_n(y)[\psi^{(k)}(x) - \epsilon]\,dy \leq (\phi_n * \psi)^{(k)}(x) = [\phi_n * (\psi^{(k)})](x) = \int_{-\delta_n}^{\delta_n} \phi_n(y)\psi^{(k)}(x-y)\,dy$$

$$\leq \int_{-\delta_n}^{\delta_n} \phi_n(y)[\psi^{(k)}(x) + \epsilon]\,dy = \psi^{(k)}(x) + \epsilon\ .$$

So $(\phi_n * \psi)^{(k)} \to \psi^{(k)}$ uniformly. 👍

XXII.2.2.23. $\mathcal{D}(\mathbb{R})$ is an associative, commutative topological algebra under convolution. It does not have a unit (the Dirac delta function is the "theoretical unit," cf. XXII.2.1.42.). What XXII.2.2.22. says is that an ai-sequence is an *approximate identity* or *approximate unit* (hence the terminology).

XXII.2.2.24. PROPOSITION. Let (ϕ_n) be an ai-sequence, and $f \in \mathcal{D}'(\mathbb{R})$. Then $f * \phi_n \to f$ in $\mathcal{D}'(\mathbb{R})$.

PROOF: Let $\psi \in \mathcal{D}(\mathbb{R})$. Then

$$\langle f * \phi_n, \psi \rangle = \langle f, (R\phi_n) * \psi \rangle \to \langle f, \psi \rangle$$

since $(R\phi_n)$ is also an ai-sequence. 👍

XXII.2.2.25. THEOREM. $\mathcal{D}(\mathbb{R})$ is dense in $\mathcal{D}'(\mathbb{R})$ (hence $\mathcal{E}(\mathbb{R})$, $C(\mathbb{R})$, and $L^1_{\text{loc}}(\mathbb{R})$ are also dense).

PROOF: If $f \in \mathcal{D}'(\mathbb{R})$, let f_n be the restriction of f to $[-n, n]$. It is clear that $f_n \to f$ in $\mathcal{D}'(\mathbb{R})$. So distributions of compact support are dense in $\mathcal{D}'(\mathbb{R})$. Now let f be a distribution of compact support. If (ϕ_n) is an ai-sequence, then $f * \phi_n$ is a $\mathcal{C}^\infty$ function by XXII.2.2.20., and has compact support (), hence is in $\mathcal{D}(\mathbb{R})$, and $f * \phi_n \to f$ in $\mathcal{D}'(\mathbb{R})$ by XXII.2.2.24.. 👍

Using a somewhat more delicate argument, it can be similarly shown that $\mathcal{D}(\Omega)$ is dense in $\mathcal{D}'(\Omega)$ for any Ω.

We can extend convolution to distributions of compact support. As in (), we have:

XXII.2.2.26. PROPOSITION. Let $f \in \mathcal{D}'(\mathbb{R})$ have support in $[a,b]$. If $\phi \in \mathcal{D}(\mathbb{R})$ has support in $[c,d]$, then $f * \phi$ has support in $[a+c, b+d]$. In particular, $f * \phi \in \mathcal{D}(\mathbb{R})$.

XXII.2.2.27. Now let $f, g \in \mathcal{D}'(\mathbb{R})$, with g of compact support. We can define $f * g$ by

$$\langle f * g, \psi \rangle = \langle f, (\bar{R}g) * \psi \rangle$$

(note that $\bar{R}g$ also has compact support). This extends the previous definition of $f * g$ if $g \in \mathcal{D}(\mathbb{R})$.

XXII.2.2.28. We can also define $g * f$ to be $f * g$. This definition makes additional sense because of the next fact, which generalizes ().

XXII.2.2.29. PROPOSITION. Let $g \in \mathcal{D}'(\mathbb{R})$ with compact support. Then $S_g : \mathcal{D}(\mathbb{R}) \to \mathcal{D}(\mathbb{R})$ defined by $S_g(\phi) = g * \phi$ extends to $\tilde{S}_g : \mathcal{E}(\mathbb{R}) \to \mathcal{E}(\mathbb{R})$, and $(\tilde{S}_g \phi) * \psi = \tilde{S}_g(\phi * \psi)$ for all $\phi \in \mathcal{E}(\mathbb{R})$, $\psi \in \mathcal{D}(\mathbb{R})$. [Recall that if $\phi \in \mathcal{E}(\mathbb{R})$, $\psi \in \mathcal{D}(\mathbb{R})$, then $\phi * \psi$ is defined and in $\mathcal{E}(\mathbb{R})$.]

XXII.2.2.30. Thus if $f, g \in \mathcal{D}'(\mathbb{R})$ and g has compact support, and $\phi \in \mathcal{D}(\mathbb{R})$, then $f * \phi \in \mathcal{E}(\mathbb{R})$, so $\tilde{S}_g(f * \phi)$ makes sense. Thus we get a map $\bar{S}_g : \mathcal{D}'(\mathbb{R}) \to \mathcal{D}'(\mathbb{R})$ defined by $[\bar{S}_g f]\phi = \tilde{S}_g(f * \phi)$ for $\phi \in \mathcal{D}(\mathbb{R})$. It is straightforward to check that $\tilde{S}_g(f * \phi) = (f * g) * \phi$, so we may write $g * f$ for $\bar{S}_g f$, in agreement with previous notation.

A Sequential Approach

XXII.2.2.31. Let $f \in \mathcal{D}'(\Omega)$, and let $\Lambda = (a, b) \subseteq_c \Omega$. Fix an m and an absolutely continuous F such that the restriction of f to Λ is $D^m F$. Choose a sequence (p_n) of polynomials converging uniformly to F on $[a, b]$ by the Weierstrass Approximation Theorem. Then the sequence $(p_n^{(m)})$ of m'th derivatives converges in $\mathcal{D}'(\Lambda)$ to the restriction of f. Thus f is represented on Λ by the sequence (p_n) and the number m, abbreviated $((p_n), m)$. (We could equally well use $\mathcal{C}^\infty$ functions instead of polynomials.) Using such an approximating sequence makes description of the derivative in $\mathcal{D}'(\Omega)$ easy: Df is represented by the same sequence (p_n) and the number $m + 1$.

XXII.2.2.32. Such a representing sequence is not unique, but if (q_n) is another representing sequence of polynomials (or $\mathcal{C}^\infty$ functions) for the same m, then $p_n - q_n \to 0$ uniformly on $[a, b]$. One can also have a representing sequence for a larger m by taking suitable antiderivatives (agreeing at some fixed point of $[a, b]$, say a). Thus, on Λ, f is represented by an equivalence class of sequences of polynomials (or $\mathcal{C}^\infty$ functions), where the equivalence relation is the one generated by these two types of equivalences.

We can for convenience consider more general representing sequences (f_n), where each f_n is $\mathcal{C}^\infty$ and (f_n) converges uniformly on $[a, b]$ to a continuous function f which is not necessarily absolutely continuous, and $k \geq 0$. For such a sequence, let F_n be the antiderivative of f_n vanishing at a, and $m = k + 1$; then $((F_n), m)$ is a representing sequence of the previous type.

XXII.2.2.33. Thus distributions on Ω can be defined as objects locally represented by pairs $((f_n), k)$, where (f_n) is a uniformly convergent sequence of $\mathcal{C}^\infty$ functions and $k \geq 0$. Two pairs $((f_n), k)$ and $((g_n), \ell)$ (with, say, $k \leq \ell$) locally represent the same distribution if the sequence $(f_n - g_n^{(\ell - k)})$ converges uniformly to 0. All the basic properties of distributions can be developed rather elegantly this way with modest prerequisites (e.g. Lebesgue integration is not needed). For example, if a distribution is represented locally by $((f_n), k)$, its distributional derivative is locally represented by $((f_n), k + 1)$. And the pairing with test

functions is also easily defined: if F is a distribution on Ω, and $\phi \in \mathcal{D}(\Omega)$, let $[a,b]$ be a closed bounded interval contained in Ω containing the support of ϕ, and let $((f_n),k)$ represent f on $[a,b]$; then the sequence

$$(-1)^k \int_a^b f_n(x)\phi^{(k)}(x)\,dx$$

converges to a number $\langle F,\phi\rangle$ which is readily shown (using integration by parts) to be independent of the choice of $[a,b]$ or the representing sequence and k. Note that only Riemann integration is used here.

Thus each distribution is locally a "k'th order limit of a uniformly convergent sequence of smooth functions" (cf. I.6.1.18.). This approach, which is strongly reminiscent of the construction of the real numbers via Cauchy sequences (III.5.2.), is described in detail in [AMS73].

XXII.2.2.34. The description can be streamlined: a distribution F is locally represented by a pair (f,k), where f is a continuous function and $k \geq 0$. Such a pair corresponds to a pair $((f_n),k)$, where (f_n) is a sequence of C^∞ functions converging uniformly to f [if $((g_n),k)$ is another local representation equivalent to $((f_n),k)$, then (f_n) and (g_n) have the same limit]. We have (f,k) equivalent to (g,ℓ) $(k \leq \ell)$ if $g \in \mathcal{C}^{\ell-k}$ and $f - g^{(\ell-k)}$ is a polynomial of degree less than k. If a distribution is locally represented by (f,k), its distributional derivative is locally represented by $(f,k+1)$, and if $\phi \in \mathcal{D}(\Omega)$, and the distribution F is locally represented on an $[a,b]$ by (f,k), we have

$$\langle F,\phi\rangle = (-1)^k \int_a^b f(x)\phi^k(x)\,dx\ .$$

This approach, developed in [], is clean but it is harder to derive some properties than in the sequence approach.

XXII.2.2.35. If $\Lambda_1 \subseteq \Lambda_2 \subseteq_c \Omega$, and (p_n) is a representing sequence for f with order m_1 on Λ_1, a compatible representing sequence for f on Λ_2 can be obtained. For if (q_n)

XXII.2.3. Distributions on $\mathbb{R}^n$

This section will largely be a repeat in greater generality of the previous two sections, which are of course the special case $n = 1$. Although most of the theory works identically, we have somewhat pedantically done the $n = 1$ case separately first, for a couple of reasons:

(i) The case $n = 1$ is a little easier to follow conceptually, and is also technically easier in some places.

(ii) There are a few results which are special (or at least much easier to state) in the case $n = 1$. For example, there are tricky issues with complicated (connected) regions which do not arise in the one-dimensional case, where we only need to work with intervals.

XXII.2.4. Tempered Distributions

XXII.2.4.1. It turns out to be important and useful to consider a somewhat larger class of test functions. We let $\mathcal{S}$ be the set of $\mathcal{C}^\infty$ functions from $\mathbb{R}^n$ to $\mathbb{R}$ (or $\mathbb{C}$) which do not necessarily have compact support, but which "vanish rapidly at infinity" (see () for the precise definition). Then $\mathcal{S}$ is a vector space with a natural topology, and we consider the space $\mathcal{S}'$ of continuous linear functionals on $\mathcal{S}$.

XXII.2.4.2. We have $\mathcal{D} \subseteq \mathcal{S}$. The relative $\mathcal{S}$-topology on $\mathcal{D}$ is not the same as the topology from XXII.2.1.12. (strictly weaker), but any continuous linear functional on $\mathcal{S}$ defines a continuous linear functional on $\mathcal{D}$, thus a generalized function or distribution. The distributions obtained are called *tempered distributions* since they are the distributions with controlled growth at infinity.

XXII.2.4.3. There are several good reasons for working with $\mathcal{S}$ (and $\mathcal{S}'$) rather than $\mathcal{D}$. Two of the most important are:

(i) The topology on $\mathcal{D}$ is not metrizable (not first countable), but the topology on $\mathcal{S}$ is metrizable, and in fact $\mathcal{S}$ is a Fréchet space.

(ii) The Fourier transform on $\mathbb{R}^n$ sends $\mathcal{S}$ into $\mathcal{S}$, and in fact sends $\mathcal{S}$ homeomorphically onto $\mathcal{S}$ (), and hence dualizes to a linear homeomorphism on $\mathcal{S}'$.

XXII.3. Sobolev Spaces

XXII.3.1. Definition of Sobolev Spaces

If Ω is an open set in $\mathbb{R}^n$ and $p \geq 1$, write $L^p(\Omega)$ for $L^p(\Omega, \mathcal{L}, \lambda)$.

XXII.3.1.1. Definition. Let Ω be an open set in $\mathbb{R}^n$, $m \in \mathbb{N} \cup \{0\}$, and $1 \leq p < \infty$. The *Sobolev space* $W^{m,p}(\Omega)$ is the set of all ordinary functions f such that f and all the distributional derivatives $D^\alpha f$ of order $\leq m$ are ordinary functions in $\mathcal{L}^p(\Omega)$, with functions agreeing a.e. identified. If $f \in W^{m,p}(\Omega)$, set

$$\|f\|_{m,p} = \left[\sum_{0 \leq |\alpha| \leq m} \|D^\alpha f\|_p^p \right]^{1/p} .$$

XXII.3.1.2. Since $L^p(\Omega) \subseteq L^1_{\text{loc}}(\Omega)$, the definition makes sense, and an element of $W^{m,p}(\Omega)$ is weakly differentiable of order at least m. To be really careful, we should define $\mathcal{W}^{m,p}(\Omega)$ to be the set of functions and $W^{m,p}(\Omega)$ the quotient by the functions which are 0 a.e., but this is a little messy since weak derivatives are only well defined up to a.e. equivalence, so we will adhere to the usual sloppiness and regard

$$W^{m,p}(\Omega) \subseteq L^p(\Omega) \subseteq L^1_{\text{loc}}(\Omega) \subseteq \mathcal{D}'(\Omega) .$$

It is easily checked that $\| \cdot \|_{m,p}$ is a norm on $W^{m,p}(\Omega)$. We have $W^{0,p}(\Omega) = L^p(\Omega)$; $W^{m,p}(\Omega) \subseteq W^{k,p}(\Omega)$ for any p if $0 \leq k \leq m$ (but note that the norms do not agree). If $f \in W^{m,p}(\Omega)$ and $|\alpha| \leq m$, then $D^\alpha f \in W^{m-|\alpha|,p}(\Omega)$ and $\|D^\alpha f\|_{m-|\alpha|,p} \leq \|f\|_{m,p}$, i.e. differential operators are bounded operators between Sobolev spaces.

XXII.3.1.3. If $n = 1$, the elements of $W^{m,p}(\Omega)$ ($m \geq 1$) are functions f which are $m-1$-times continuously differentiable, for which $f^{(m-1)}$ is locally absolutely continuous, and for which $f^{(k)} \in \mathcal{L}^p(\Omega)$ for $0 \leq k \leq m$ (we do not have to make any identifications in this case, so the elements are honest functions). But if $n > 1$, the situation is much more subtle (). If Ω is nice and m is sufficiently large ($m > n/p$), each element of $W^{m,p}(\Omega)$ is represented by a unique continuous function (), but this is not true for smaller m in general, and elements of $W^{m,p}(\Omega)$ do not have canonical representatives.

XXII.3.2. Sobolev Inequalities and Embedding Theorems

XXII.3.2.1. There is a collection of inequalities and embeddings of Sobolev spaces which are collectively called "Sobolev inequalities" or "the Sobolev embedding theorem," since they are mainly due to Sobolev [Sob63] but improved by Morrey [Mor40] and Gagliardo [Gag58]. The name *Sobolev inequalities* is probably better, since the essence of the results is the inequalities and the "embeddings" are a straightforward consequence, and the "embeddings" are actually not always embeddings (there can be a nontrivial kernel in some cases).

XXII.3.2.2. The inequalities only work on regions Ω which are "sufficiently nice"; the required niceness varies in different results. We will not carefully describe the various niceness conditions; cf. e.g. [AF03]. A reasonable sufficient condition for all results is that the boundary $\partial\Omega$ be *strongly locally Lipschitz*, which in the case where $\partial\Omega$ is compact means $\partial\Omega$ is locally the graph of a Lipschitz function. This condition is satisfied in many cases of interest:

Ω is bounded and convex.

Ω is a regular open set () and $\bar{\Omega}$ is a smooth manifold with compact boundary.

If $\partial\Omega$ is unbounded, the conditions are much more subtle, but are satisfied e.g. if Ω is a finite intersection of half-spaces, such as an open quadrant in $\mathbb{R}^2$.

The topology of Ω does not seem to enter in, and can be fairly complicated (but perhaps not too complicated). For example, an annulus in $\mathbb{R}^2$ or the interior or exterior of a solid smooth n-holed torus (even knotted) in $\mathbb{R}^3$ satisfies the niceness conditions.

XXII.4. Differential Equations and Distributions

Functional analysis methods and the theory of distributions play a fundamental role in the modern theory of differential equations, particularly PDEs. In this section, we describe some of the core features. Our treatment of this vast subject is far from comprehensive; a full treatment would include most of the theory of PDEs (a subject in which the author is not an expert).

When solving an equation (or solving a problem) in a particular space, it is often advantageous to work in a larger space with better completeness or compactness properties, and/or where the equation is better behaved. The most basic example is defining and working with the real numbers themselves (e.g. to find roots of polynomials). Another example is the use of currents and Hausdorff measure in geometric measure theory (). The idea is to show that the equation (problem) has a solution in the extended setting, and then argue that the solution actually is of the kind sought (i.e. is in the original space). In the case of differential equations (ordinary or partial), one tries to show that the equation has a "weak solution" in a space of generalized functions, and then via a regularity argument conclude that the weak solution is an actual solution.

XXII.4.1. Ordinary Differential Equations and Generalized Functions

XXII.4.2. Partial Differential Equations and Generalized Functions

XXII.4.3. Exercises

XXII.4.3.1. (a) Let $g \in \mathcal{D}'(\mathbb{R})$ with compact support. Show that $\tilde{S}_g$ (cf. ()) is continuous from $\mathcal{E}(\mathbb{R})$ to $\mathcal{E}(\mathbb{R})$.

(b) Let $g \in \mathcal{D}'(\mathbb{R})$ with compact support. Show that $\bar{S}_g$ () is continuous from $\mathcal{D}'(\mathbb{R})$ to $\mathcal{D}'(\mathbb{R})$, i.e. if $f_n \to f$ in $\mathcal{D}'(\mathbb{R})$, then $f_n * g \to f * g$ in $\mathcal{D}'(\mathbb{R})$.

(c) Let $f \in \mathcal{D}'(\mathbb{R})$, and (g_n) a sequence in $\mathcal{D}'(\mathbb{R})$ all of whose supports are contained in a fixed finite interval. Show that $f * g_n \to f * g$ in $\mathcal{D}'(\mathbb{R})$.

(d) What if in (c) the supports of the g_n are compact but not contained in a fixed bounded interval?

(e) Show that if $f, g, h \in \mathcal{D}'(\mathbb{R})$ and two of them have compact support, then $(f * g) * h = f * (g * h)$.

Chapter XXIII

Topological Groups and Abstract Harmonic Analysis

XXIII.1. Topological Groups

XXIII.1.1. Groups and Homomorphisms

To make this manuscript more self-contained, we give here a quick summary of the theory of groups. This topic more properly belongs in an algebra text, and there are many such texts which cover this vast and fundamental subject in far greater detail.

XXIII.1.1.1. DEFINITION. A *group* is a pair $(G, *)$, where G is a set and $*$ is a binary operation on G, i.e. a function $(x, y) \mapsto x * y$ from $G \times G$ to G, satisfying

(Ass) $*$ is *associative*: $x * (y * z) = (x * y) * z$ for all $x, y, z \in G$.

(Id) There is a *unit* or *identity* for $*$, an element $e \in G$ with $e * x = x * e = x$ for all $x \in G$.

(Inv) For each $x \in G$ there is an element $x^{-1} \in G$ with $x * x^{-1} = x^{-1} * x = e$.

A set with an associative binary operation is called a *semigroup*. A semigroup with an identity is sometimes called a *monoid*.

As usual, the operation $*$ is part of the definition of a group, so it is technically an abuse of terminology to make the common statement "Let G be a group." (We will nonetheless often abuse terminology in this way.)

XXIII.1.1.2. The operation in a group or semigroup is not necessarily commutative, i.e. the identity $x * y = y * x$ does not necessarily hold for all $x, y \in G$. A group whose operation is commutative is called a *commutative group* or *abelian group*, and similarly for semigroups.

XXIII.1.1.3. If $(G, *)$ is a group, the group operation is usually written multiplicatively, i.e. writing xy for $x * y$, and the identity e is often written 1. If $(G, *)$ is abelian, it is customary (but not universal) to write the operation as addition, writing $x + y$ for $x * y$. If additive notation is used, the identity e is written 0, and the inverse x^{-1} of an element x written $-x$. In many natural examples of groups, the group operation is really addition or multiplication, defined in usual ways. Semigroup notation is often similar. Because of associativity, for any x the elements xxx, $xxxx$, etc., are unambiguously defined; we denote by x^n ($n \in \mathbb{N}$) the element which is a product of n x's. In a group, x^n is also defined for $n < 0$ as the product of n copies of x^{-1}, and $x^0 = e$; then $x^n x^m = x^{n+m}$ for all $n, m \in \mathbb{Z}$. In additive notation, x^n is denoted nx.

XXIII.1.1.4. The operation in a group has both left and right cancellation: if $xz = yz$, then

$$x = x(zz^{-1}) = (xz)z^{-1} = (yz)z^{-1} = y(zz^{-1}) = y$$

and similarly if $xy = xz$, then $y = z$. In particular, the identity is unique: if $xy = x$ or $yx = x$ for some $x \in G$, then $y = e$ (and in particular $zy = yz = z$ for all $z \in G$); and inverses are unique: if $xy = e$ or $yx = e$, then $y = x^{-1}$ (and in particular $xy = yx = e$). We also have that $e^{-1} = e$ and that e is the unique idempotent in G, i.e. $xx = x$ if and only if $x = e$. Semigroups, even abelian monoids, do not have cancellation in general.

Homomorphisms

Homomorphisms (group-homomorphisms) are functions or maps between groups which respect the group operation:

XXIII.1.1.5. Definition. Let $(G, *_G)$ and $(H, *_H)$ be groups. A *group-homomorphism* (or just *homomorphism* if there is no possible confusion) is a function $\phi : G \to H$ such that

$$\phi(x *_G y) = \phi(x) *_H \phi(y)$$

for all $x, y \in G$. A *group-isomorphism* (or just *isomorphism* if there is no possible confusion) is a bijective group-homomorphism.

We usually just write the condition as $\phi(xy) = \phi(x)\phi(y)$, but it should be observed that the operations on the left and right are different unless the groups are the same.

The simple proof of the next result, and subsequent propositions, is left to the reader:

XXIII.1.1.6. Proposition. Let $\phi : G \to H$ be a homomorphism. Then

(i) $\phi(e_G) = e_H$.

(ii) $\phi(x^{-1}) = [\phi(x)]^{-1}$ for all $x \in G$.

(iii) If ϕ is an isomorphism, then $\phi^{-1} : H \to G$ is also an isomorphism (i.e. is a homomorphism).

An isomorphism is just an identification of the groups, which are abstractly identical. In a general homomorphism from G to H, elements of G are identified together in H in a very systematic way (XXIII.1.1.21.), but the essential character of the group operation is preserved.

"Homomorphism is a kind of *systematically abridged translation*. The original is not only translated into another language, but also abridged so that what results finally from translation and abbreviation is uniformly, systematically condensed into one-half or one-third or some other fraction of the original extension. Subtleties may be lost by such abridgement but everything that is in the original is represented by something in the translation, and, on a reduced scale, the relations are preserved."

G. Polya[1]

Subgroups and Cosets

XXIII.1.1.7. A subset of a group which is itself a group under the same operation is called a subgroup of the group. More precisely, if G is a group, then a subset H of G is a *subgroup* if

(i) H is *closed* under the group operation, i.e. if $x, y \in H$, then $xy \in H$.

(ii) The identity e_G is in H.

(iii) For all $x \in H$, x^{-1} is also in H.

(of course, (ii) is automatic from (i) and (iii)).

XXIII.1.1.8. Every group G has at least two subgroups: G itself and $\{e\}$. Other subgroups, which may or may not exist, are called *nontrivial subgroups*. Any intersection of subgroups is a subgroup. In particular, if $A \subseteq G$, then there is a smallest subgroup $\langle A \rangle$ of G containing A, the intersection of all subgroups of G containing A (there is at least one such subgroup, G), called the subgroup of G generated by A. This subgroup can be concretely described as the set of all finite products of powers, positive and negative, of elements of A (including the "empty product" e). In particular, if $A = \{x\}$, the subgroup $\langle A \rangle = \langle x \rangle$ is the set of all integer powers of x, called the *cyclic subgroup* of G generated by x. A cyclic subgroup is always abelian.

The next result generalizes XXIII.1.1.6.:

XXIII.1.1.9. Proposition. Let G and G' be groups, and $\phi : G \to G'$ a homomorphism.
(i) If H is a subgroup of G, then $\phi(H)$ is a subgroup of G'.
(ii) If H' is a subgroup of G', then $\phi^{-1}(H')$ is a subgroup of G.
In particular, the range of ϕ is a subgroup of G'.

XXIII.1.1.10. Definition. Let H be a subgroup of a group G, and $x \in G$. The *left coset* of H containing x is the subset $xH = \{xy : y \in H\}$ of G. The *right coset* is the set $Hx = \{yx : y \in H\}$.

Note that since $e \in H$, $x \in xH$ and $x \in Hx$ for any x. If G is abelian and the operation is written additively, cosets are normally written $x + H$ (there is no distinction between left and right cosets in the abelian case).

[1] "Generalization, Specialization, Analogy," *Induction and Analogy in Mathematics*; reprinted in *New Directions in the Philosophy of Mathematics*, ed. T. Tymoczko, p. 118.

XXIII.1.1.11. PROPOSITION. Let H be a subgroup of G. Then any two left cosets of H in G are either identical or disjoint. In fact, for any $x, y \in G$, the following are equivalent:

(i) $xH \cap yH \neq \emptyset$.

(ii) $xH = yH$.

(iii) $y^{-1}x \in H$.

An analogous statement holds for right cosets (with $y^{-1}x$ replaced by xy^{-1} in (iii)).

XXIII.1.1.12. Write G/H for the set of left cosets of H in G, and $H\backslash G$ for the set of right cosets. These sets are called *homogeneous spaces* and are naturally themselves groups in good cases (XXIII.1.1.20.).

Normal Subgroups

XXIII.1.1.13. If G is not abelian, the left and right cosets of a subgroup do not coincide in general (except for the coset $H = eH = He$ itself). But they do if H is "nicely situated" in G:

XXIII.1.1.14. PROPOSITION. Let H be a subgroup of a group G. The following are equivalent:

(i) Every left coset of H in G is also a right coset.

(ii) For every $x \in G$, $xH = Hx$.

(iii) For every $x \in G$ and $y \in H$, $x^{-1}yx \in H$ (i.e. $x^{-1}Hx = H$).

XXIII.1.1.15. DEFINITION. A subgroup H of a group G is *normal* (in G) if the conditions of XXIII.1.1.14. are satisfied.

XXIII.1.1.16. Note that normality of a subgroup depends on both G and H in general. The trivial subgroups G and $\{e\}$ are always normal. If G is abelian, every subgroup of G is normal.

Homomorphisms and Normal Subgroups

XXIII.1.1.17. PROPOSITION. Let $\phi : G \to G'$ be a group-homomorphism.
(i) If H' is a normal subgroup of G', then $\phi^{-1}(H')$ is a normal subgroup of G.
(ii) If ϕ is surjective and H is a normal subgroup of G, then $\phi(H)$ is a normal subgroup of G'.

The conclusion of (ii) is not necessarily true if ϕ is not surjective. In fact, $\phi(G)$ need not be a normal subgroup of G' in general.

XXIII.1.1.18. As a special case of (i), if $\phi : G \to G'$ is a homomorphism, define the *kernel* of ϕ to be

$$ker(\phi) = \phi^{-1}(\{e_{G'}\}) = \{x \in G : \phi(x) = e_{G'}\} \ .$$

Then $ker(\phi)$ is a normal subgroup of G.

Quotient Groups

XXIII.1.1.19. As a converse to XXIII.1.1.18., suppose H is a subgroup of G. Try to put a binary operation on the set of (left or right) cosets of H in G by $(xH)(yH) = xyH$. Then

XXIII.1.1.20. Proposition. (i) The operation $(xH)(yH) = xyH$ is well defined if and only if H is a normal subgroup of G.

(ii) If H is normal in G, this operation makes the set of cosets into a group, called the *quotient group* of G by H or G mod H, denoted G/H.

(iii) If H is normal in G, the map $x \mapsto xH$ from G to G/H is a (surjective) homomorphism with kernel H.

The First Isomorphism Theorem

The next result, although quite easy, is called the *First Isomorphism Theorem* of group theory (there are also a second, third, and sometimes fourth; see books on abstract algebra).

XXIII.1.1.21. Proposition. [First Isomorphism Theorem of Group Theory] Let G and G' be groups, H a normal subgroup of G, and $\phi : G \to G'$ a homomorphism.

(i) If $H \subseteq ker(\phi)$, then ϕ drops to a well-defined homomorphism $\bar{\phi} : G/H \to G'$ defined by $\bar{\phi}(xH) = \phi(x)$, which is injective if and only if $H = ker(\phi)$.

(ii) If $H = \ker(\phi)$, then $\phi(G)$ is isomorphic to G/H via $\bar{\phi}$. In particular, if ϕ is surjective, then $G' \cong G/H$ via $\bar{\phi}$.

XXIII.1.2. Topological Groups

XXIII.1.2.1. Definition. A *topological group* is a group G with a topology such that the group operation $(x, y) \mapsto xy$ is a continuous function from $G \times G$ to G (i.e. multiplication is jointly continuous on G), and inversion $x \mapsto x^{-1}$ is a continuous function from G to G.

XXIII.1.2.2. Note that both the group operation and the topology are parts of the definition of a topological group; thus it is technically incorrect to say "Let G be a topological group" (although we do it all the time). In fact, a group can have many different topologies making it into a topological group.

XXIII.1.2.3. EXAMPLES. (i) Let G be any group. Then G is a topological group with the discrete topology or the indiscrete topology.

A topological group with the indiscrete topology is not very interesting. However, a group with the discrete topology is much more interesting than might be at first anticipated. In fact, we can regard the theory of groups (with no topology) as a special case of the theory of topological groups by using the discrete topology (although this is logically circular if not done properly).

(ii) The additive group $\mathbb{R}$ (or $\mathbb{R}^n$) with its ordinary topology is a topological group. More generally, the additive group of any normed vector space is a topological group with the topology induced by the norm.

(iii) Let $GL_n(\mathbb{R})$ be the multiplicative group of invertible $n \times n$ matrices with real entries, with the relative topology as a subset of $M_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$. Then $GL_n(\mathbb{R})$ is a topological group.

(iv) Any subgroup of a topological group is a topological group with the relative topology.

(v) Groups of homeomorphisms of topological spaces can often be given natural topologies making them into topological groups, although the details can be subtle (cf. XXIII.1.7.8.–XXIII.1.7.14.).

These examples barely scratch the surface of the possibilities.

The condition defining a topological group can be streamlined:

XXIII.1.2.4. PROPOSITION. Let G be a group with a topology. Then G is a topological group under the given topology if and only if the map $\phi : G \times G \to G$ given by $\phi(x, y) = x^{-1}y$ is continuous.

PROOF: If G is a topological group, then ϕ is a composition of multiplication and inversion, hence continuous. Conversely, if ϕ is continuous, then

$$x \mapsto (x, e) \mapsto x^{-1}e = x^{-1}$$

is continuous, and hence

$$(x, y) \mapsto (x^{-1}, y) \mapsto (x^{-1})^{-1}y = xy$$

is continuous, so G is a topological group in the given topology. 👍

XXIII.1.2.5. Joint continuity of multiplication in a topological group implies separate continuity (). But separate continuity of multiplication does not imply joint continuity (Exercise XXIII.1.7.6.). There are situations where separate continuity does imply joint continuity ().

Here is a useful consequence of the joint continuity of multiplication and continuity of inversion:

XXIII.1.2.6. PROPOSITION. Let G be a topological group, I a directed set, and (x_i), (y_i) nets in G over I. Suppose $x_i y_i \to z$. Then

(i) If $x_i \to x$, then $y_i \to x^{-1}z$.

(ii) If $y_i \to y$, then $x_i \to zy^{-1}$.

PROOF: (i): We have

$$y_i = x^{-1}xx_i^{-1}(x_i y_i) \to x^{-1}xx^{-1}z \ .$$

The proof of (ii) is similar. 👍

Translations and Homogeneity

XXIII.1.2.7. DEFINITION. Let G be a topological group, and $g \in G$. The functions λ_g, ρ_g from G to G defined by

$$\lambda_g(x) = gx$$
$$\rho_g(x) = xg$$

are called *left translation* and *right translation* by g respectively.

XXIII.1.2.8. PROPOSITION. Let G be a topological group, and $g \in G$. Then the left and right translations λ_g and ρ_g are homeomorphisms from G to G. If $g \neq e$, these homeomorphisms have no fixed points.

PROOF: Translations are continuous by (separate) continuity of the group operation. Left translation λ_g has inverse $\lambda_{g^{-1}}$, hence is a bijection and a homeomorphism; similarly ρ_g has inverse $\rho_{g^{-1}}$. If $\lambda_g(x) = x$ for some $x \in G$, i.e. $gx = x$, then $g = (gx)x^{-1} = xx^{-1} = e$, and similarly for ρ_g. 👍

XXIII.1.2.9. COROLLARY. If G is a topological group, then the topology of G is *homogeneous*, i.e. if $x, y \in G$, then there is a homeomorphism from G to G sending x to y.

PROOF: $\lambda_{yx^{-1}}$ is such a homeomorphism. 👍

Thus a topological group "looks the same" around every point.

XXIII.1.2.10. Similarly, *inversion* ι, where $\iota(x) = x^{-1}$, is a continuous map from G to G which is its own inverse, i.e. ι is a homeomorphism.

The next result, although simple, appears quite bizarre and surprising at first glance, and is a dramatic illustration of the maxim that being open is not the opposite of being closed for subsets of a topological space. See Exercise XXIII.1.7.2. for a generalization.

XXIII.1.2.11. PROPOSITION. Let G be a topological group. Any open subgroup of G is closed.

PROOF: Let H be an open subgroup of G. Then, for any $g \in G$, the left coset gH is the image of H under the homeomorphism λ_g, and hence gH is also open. The complement of H in G is the union of the nonidentity left cosets of H in G, hence open. 👍

See Exercise XXIII.1.7.2. for an alternate proof.

Fundamental Systems of Neighborhoods

XXIII.1.2.12. If G is a topological group and U is a neighborhood of e, then for any $g \in G$ $gU = \lambda_g(U)$ and $Ug = \rho_g(U)$ are neighborhoods of g, and conversely if V is a neighborhood of g then $g^{-1}V$ and Vg^{-1} are neighborhoods of e. Thus to describe a topological group topology on a group G, it suffices to specify the neighborhoods of e.

XXIII.1.2.13. If A and B are subsets of G, write $AB = \{xy : x \in A, y \in B\}$, and $A^2 = AA$. If A, B, C are subsets of G, then $(AB)C = A(BC)$ by associativity; call this subset ABC. Products of more than three subsets can be defined similarly. A^n has an unambiguous meaning for any $n \in \mathbb{N}$, and $A^n A^m = A^{n+m}$.

If $A \subseteq G$, write $A^{-1} = \iota(A) = \{x^{-1} : x \in A\}$.

XXIII.1.2.14. Proposition. If G is a topological group, $A, B \subseteq G$, and one of A, B is open, then AB is open. If A is open, then A^{-1} is also open.

Proof: If B is open, then $aB = \lambda_a(B)$ is open for each $a \in A$, so $AB = \cup_{a \in A} aB$ is open. The proof if A is open is similar. $A^{-1} = \iota(A)$ is open if A is. 👍

XXIII.1.2.15. Proposition. Let G be a topological group. The collection $\mathcal{N}_e$ of neighborhoods of e has the following properties:

(i) If $U \in \mathcal{N}_e$, then there is a $V \in \mathcal{N}_e$ with $V^2 \subseteq U^\circ$.

(ii) There is a $W \in \mathcal{N}_e$ with $W = W^{-1}$ and $W \subseteq U^\circ$.

(iii) For every $g \in G$, there is a $V \in \mathcal{N}_e$ with $gVg^{-1} \subseteq U^\circ$.

Proof: (i): Let $\mu : G \times G \to G$ be multiplication, i.e. $\mu(x, y) = xy$. Then $\mu^{-1}(U^\circ)$ is an open neighborhood of (e, e) in $G \times G$, hence contains a neighborhood of the form $V_1 \times V_2$ where each V_j is in $\mathcal{N}_e$. Set $V = V_1 \cap V_2$.
(ii): Set $W = U^\circ \cap (U^\circ)^{-1}$.
(iii): Set $V = g^{-1}U^\circ g = \rho_g(\lambda_{g^{-1}}(U^\circ)) = \lambda_{g^{-1}}(\rho_g(U^\circ))$, which is in $\mathcal{N}_e$. 👍

XXIII.1.2.16. Corollary. If G is a topological group, and $\mathcal{F}$ is a local base at e for the topology (not necessarily consisting of open sets), then $\mathcal{F}$ has the following properties:

(i) If $U, V \in \mathcal{F}$, then there is a $W \in \mathcal{F}$ with $W \subseteq U \cap V$.

(ii) If $U \in \mathcal{F}$ and $g \in U$, then there is a $V \in \mathcal{F}$ with $Vg \subseteq U$.

(iii) If $U \in \mathcal{F}$, then there is a $V \in \mathcal{F}$ with $V^{-1}V \subseteq U$.

(iv) If $U \in \mathcal{F}$ and $g \in G$, then there is a $V \in \mathcal{F}$ with $gVg^{-1} \subseteq U$.

Such an $\mathcal{F}$ has additional properties such as:

(iiia) If $U \in \mathcal{F}$, there is a $V \in \mathcal{F}$ with $V^2 \subseteq U$.

(iiib) If $U \in \mathcal{F}$, there is a $V \in \mathcal{F}$ with $V^{-1} \subseteq U$.

In fact, (iiia) and (iiib) together are equivalent to (iii).

XXIII.1.2.17. DEFINITION. Let G be a group. A *fundamental system of neighborhoods of* e is a nonempty collection $\mathcal{F}$ of subsets of G, each containing e, satisfying (i)–(iv) of XXIII.1.2.16. (or equivalently (i), (ii), (iiia), (iiib), (iv)).

Note that in this definition, there is no specified topology on G, so the sets in $\mathcal{F}$ are not "neighborhoods" of e in G in the usual sense of the term. But the terminology is justified by:

XXIII.1.2.18. THEOREM. Let G be a group. If $\mathcal{F}$ is a fundamental system of neighborhoods of e in G, then there is a unique topology $\mathcal{T}$ on G making G a topological group, for which $\mathcal{F}$ is a local base at e.

XXIII.1.2.19. So a topological group topology on a group G can be specified by giving a fundamental system of neighborhoods of e, and every topological group topology on G arises in this manner. Specifying a topology via a fundamental system of neighborhoods of e is one of the most convenient ways of defining topologies on groups.

Products of Closed Sets

In contrast with XXIII.1.2.14., even a product of two closed sets in a topological group is not closed in general (). But it is if one of the sets is compact:

XXIII.1.2.20. PROPOSITION. Let G be a topological group, and A and B closed subsets of G. If one of A, B is compact, then AB is closed.

PROOF: Suppose A is compact. Let $z \in \overline{AB}$. Fix a net $(x_i y_i)$ in AB with $x_i y_i \to z$ (This can be done without any form of Choice, cf. ().) Passing to a subnet if necessary, we may assume $x_i \to x$ for some $x \in A$ by compactness. Then $y_i \to x^{-1}z$ by XXIII.1.2.6., and hence $x^{-1}z \in B$ since B is closed. Thus $z = x(x^{-1}z) \in AB$. The proof if B is compact is similar. 👍

Homogeneous Spaces and Quotients

XXIII.1.2.21. If G is a topological group and H a subgroup of G, not necessarily normal, we can form the left coset space G/H. Since G/H is a set-theoretic quotient of G, i.e. there is a natural quotient map $\pi : G \to G/H$, we may give G/H the quotient topology. G/H with this topology is called the *homogeneous space* of G by H (the terminology is justified by XXIII.1.2.22.(ii)).

XXIII.1.2.22. PROPOSITION. Let G be a topological group and H a subgroup of G. Let G/H be the homogeneous space. Then

(i) The quotient map $\pi : G \to G/H$ is continuous and open.

(ii) If $g \in G$, the map $\bar{\lambda}_g : G/H \to G/H$ defined by $\bar{\lambda}_g(xH) = gxH$ is a well-defined homeomorphism, and in particular the topology of G/H is homogeneous in the sense of XXIII.1.2.9.. The map $(g, xH) \mapsto \lambda_g(xH) = gxH$ is continuous from $G \times G/H$ to G/H.

PROOF: (i) The quotient map π is continuous by definition of the quotient topology. If U is open in G, then $\pi^{-1}(\pi(U)) = UH$, which is open in G by XXIII.1.2.14.; hence $\pi(U)$ is open in G/H by definition of the quotient topology.

(ii): If $xH = yH$, i.e. $y = xz$ for some $z \in H$, then $gy = gxz$, so $gy \in gxH$, i.e. $gyH = gxH$ and $\bar{\lambda}_g$ is well defined, and is a bijection since $\bar{\lambda}_g^{-1} = \bar{\lambda}_{g^{-1}}$. If $V \subseteq G/H$ is open, then $\pi^{-1}(V)$ is open in G, so $\lambda_{g^{-1}}(\pi^{-1}(V))$ is open in G since $\lambda_{g^{-1}}$ is a homeomorphism on G; thus $\bar{\lambda}_g^{-1}(V) = \pi(\lambda_{g^{-1}}(\pi^{-1}(V)))$ is open by (i). So $\bar{\lambda}_g$ is continuous. Since $\bar{\lambda}_g^{-1} = \bar{\lambda}_{g^{-1}}$ is also continuous, $\bar{\lambda}_g$ is a homeomorphism. For the last statement, if $x, y \in G$, then $\bar{\lambda}_{yx^{-1}}$ sends xH to yH, so G/H is homogeneous. 👍

XXIII.1.2.23. PROPOSITION. Let G be a topological group, and H a closed normal subgroup. Then

(i) G/H is a topological group in the quotient topology.

(ii) If G' is another topological group, and $\phi : G \to G'$ a continuous homomorphism whose kernel contains H, then the induced homomorphism $\psi : G/H \to G'$ () is continuous.

PROOF: (i): Let $\bar{V}$ be an open set in G/H, and let $V = \pi^{-1}(\bar{V})$. Then V is open in G. Since μ is continuous, $U = \mu^{-1}(V)$ is open in $G \times G$. The quotient map $\pi \times \pi$ from $G \times G$ to $(G/H) \times (G/H)$ is open, so $\bar{U} = (\pi \times \pi)(U)$ is open in $(G/H) \times (G/H)$, and $\bar{U} = \bar{\mu}^{-1}(\bar{V})$. Thus $\bar{\mu}$ is continuous. A similar argument shows that $\bar{\iota}$ is continuous on G/H. 👍

The homeomorphisms $\bar{\lambda}_g$ are an example of a group action:

XXIII.1.2.24. DEFINITION. Let G be a topological group and X a topological space. A *group action* of G on X is a continuous function $(g, x) \mapsto \alpha_g(x)$ from $G \times X$ to X such that $\alpha_e(x) = x$ and $\alpha_{gh}(x) = \alpha_g(\alpha_h(x))$ for all $x \in X$, $g, h \in G$.

XXIII.1.2.25. The point $\alpha_g(x)$ is often written $g \cdot x$. For fixed g, α_g is a continuous function from X to X with inverse $\alpha_{g^{-1}}$, i.e. α_g is a homeomorphism from X to X. The map $\alpha : G \to Homeo(X)$ is a group-homomorphism, which is continuous if $Homeo(X)$ is given the topology of pointwise convergence.

Separation Axioms

Topological groups automatically have nice separation properties:

XXIII.1.2.26. PROPOSITION. Let G be a topological group. The following are equivalent:

(i) G is T_0.

(ii) G is Hausdorff.

(iii) G is completely regular.

PROOF: (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) is trivial. Since the topology on G comes from a uniform structure (XXIII.1.3.4.), if G is T_0, it is completely regular, so (i) $\Rightarrow$ (iii). 👍

XXIII.1.2.27. There is a simple direct proof that a topological group is T_3, and hence that a T_0 topological group is Hausdorff; cf. Exercise XXIII.1.7.3..

XXIII.1.2.28. Corollary. A second countable T_0 topological group is metrizable.

One can do much better (XXIII.1.3.7.).

Closed Subgroups

XXIII.1.2.29. Proposition. Let G be a topological group, and H a subgroup. Then G/H is Hausdorff (completely regular) if and only if H is closed in G.

XXIII.1.2.30. Proposition. Let G be a topological group, and H a subgroup of G. Then $\bar{H}$ is also a subgroup of G. If H is a normal subgroup, so is $\bar{H}$.

XXIII.1.2.31. Thus $\overline{\{e\}}$ is the smallest closed subgroup of G. It is normal, and $G/\overline{\{e\}}$ is the largest Hausdorff quotient group of G. Any continuous homomorphism from G into a Hausdorff topological group factors through $G/\overline{\{e\}}$. The space $G/\overline{\{e\}}$ is the maximal T_0 quotient of G in the sense of XIII.8.3.10..

Products of Topological Groups

XXIII.1.2.32. Proposition. Any product of topological groups is a topological group with the product topology.

Inverse Limits

Connectedness

XXIII.1.2.33. Proposition. Let G be a topological group. If U is any open neighborhood of e in G, then the subgroup of G generated by U is open, and coincides with the subgroup of G generated by $\bar{U}$.

Proof: The subset $[U \cup (U^{-1})]^n$ of G is open for each n by XXIII.1.2.14.. The subgroup of G generated by U is $\cup_{n=1}^{\infty}[U \cup (U^{-1})]^n$. Since an open subgroup is closed (XXIII.1.2.11.), this subgroup contains $\bar{U}$. 👍

XXIII.1.2.34. Corollary. Let G be a connected topological group. Then every neighborhood of e generates G as a group.

Proof: Every open subgroup of a topological group is clopen, so a connected topological group cannot contain any proper open subgroups. Every neighborhood of e in G contains an open neighborhood of e. 👍

XXIII.1.2.35. PROPOSITION. Let G be a topological group and H a subgroup. If H and G/H are connected, then G is connected.

PROOF: Suppose H is connected and G is not connected. Let U and V be nonempty disjoint clopen sets in G with $U \cup V = G$. Since each left coset of H is connected, it is entirely contained in either U or V, i.e. U and V are each unions of left cosets of H. But then $\pi(U)$ and $\pi(V)$ are disjoint clopen subsets of G/H, so G/H is not connected. 👍

XXIII.1.2.36. PROPOSITION. Let G be a topological group. Then the connected component of the identity e is a closed normal subgroup H of G, and G/H is totally disconnected.

PROOF: Let H be the connected component of e in G. Then H is a closed subset of G. $H \times H$ is a connected subset of $G \times G$, and hence its image H^2 in G under multiplication is connected, and contains e; thus $H^2 \subseteq H$, i.e. H is closed under multiplication. Similarly, H^{-1} is connected and contains e, thus $H^{-1} \subseteq H$ and H is closed under inversion. So H is a subgroup of G. If $g \in G$, then gHg^{-1} is a connected subset of G, hence $gHg^{-1} \subseteq H$. Then $H = g^{-1}(gHg^{-1})g \subseteq g^{-1}Hg \subseteq H$, so $g^{-1}Hg = H$ and H is normal.

To see that G/H is totally disconnected, it suffices to show that the connected component of $\bar{e}$ in G/H is $\{\bar{e}\}$. Let K be the connected component of $\bar{e}$ in G/H, and $G' = \pi^{-1}(K)$. Then G' is a subgroup of G and $G'/H \cong K$ is connected, and since H is connected, G' is connected by XXIII.1.2.35.. Since $H \subseteq G'$, we have $G' = H$ so $K = \{\bar{e}\}$. 👍

XXIII.1.3. Uniformities on Topological Groups

Every topological group has a natural uniform structure (actually three natural uniform structures which may or may not coincide). In fact, these uniform structures are one of the models for general uniform structures ().

XXIII.1.3.1. DEFINITION. Let $(G, \mathcal{T})$ be a topological group.

(i) The *left uniform structure* $\mathcal{L}$ on G is the uniform structure generated by the sets

$$L_U = \{(x, y) \in G \times G : x^{-1}y \in U\}$$

as U ranges over all open neighborhoods of e in G.

(ii) The *right uniform structure* $\mathcal{R}$ on G is the uniform structure generated by the sets

$$R_U = \{(x, y) \in G \times G : xy^{-1} \in U\}$$

as U ranges over all open neighborhoods of e in G.

(iii) The *two-sided uniform structure* $\mathcal{U}$ on G is the uniform structure generated by the sets L_U and R_U as U ranges over all open neighborhoods of e in G.

XXIII.1.3.2. The uniform structures $\mathcal{L}$ and $\mathcal{R}$ coincide in many cases, e.g. if G is compact () or abelian, but not in general (). The uniform structure $\mathcal{U}$ coincides with $\mathcal{L}$ or $\mathcal{R}$ if and only if $\mathcal{L} = \mathcal{R}$.

XXIII.1.3.3. PROPOSITION. Let G be a topological group. The following are equivalent:

(i) $\mathcal{L} = \mathcal{R}$.

(ii) $\mathcal{L} = \mathcal{U}$.

(iii) $\mathcal{R} = \mathcal{U}$.

(iv) G has a fundamental system of neighborhoods of e invariant under all inner automorphisms, i.e. neighborhoods U such that $gUg^{-1} = U$ for all $g \in G$.

However, even though the uniform structures do not coincide in general, they all define the given topology on G:

XXIII.1.3.4. PROPOSITION. Let $(G, \mathcal{T})$ be a topological group. Then the topology on G defined by any of the three uniform structures $\mathcal{L}$, $\mathcal{R}$, $\mathcal{U}$ is $\mathcal{T}$.

XXIII.1.3.5. PROPOSITION. Let G be a topological group. Then

(i) The translations λ_g are uniformly continuous for $\mathcal{L}$ and $\mathcal{U}$, for every $g \in G$.

(ii) The translations ρ_g are uniformly continuous for $\mathcal{R}$ and $\mathcal{U}$, for every $g \in G$.

(iii) Inversion ι is uniformly continuous for $\mathcal{U}$.

Inversion is not uniformly continuous for $\mathcal{L}$ or $\mathcal{R}$ in general: inversion interchanges $\mathcal{L}$ and $\mathcal{R}$. In fact:

XXIII.1.3.6. PROPOSITION. Let G be a topological group. The following are equivalent:

(i) The translations λ_g are uniformly continuous for $\mathcal{R}$, for every $g \in G$.

(ii) The translations ρ_g are uniformly continuous for $\mathcal{L}$, for every $g \in G$.

(iii) Inversion ι is uniformly continuous for $\mathcal{L}$.

(iv) Inversion ι is uniformly continuous for $\mathcal{R}$.

(v) $\mathcal{L} = \mathcal{R} (= \mathcal{U})$.

Metrization

It is an immediate corollary of XXIII.1.3.4. and () that a first countable topological group is metrizable, since $\mathcal{L}$, $\mathcal{R}$, and $\mathcal{U}$ are all countably generated if G is first countable. There is a more precise result:

XXIII.1.3.7. Corollary. Every first countable Hausdorff topological group is metrizable under a left-invariant metric which generates the uniformity $\mathcal{L}$.

XXIII.1.3.8. If d is a left-invariant metric for G, then the metric $d'(g,h) = d(g^{-1}, h^{-1})$ is right-invariant and defines the topology of G (and the uniformity $\mathcal{R}$), so "left-invariant" can be replaced by "right-invariant" in XXIII.1.3.7..

However, G does not have a two-sided invariant metric in general; in fact, it has a two-sided invariant metric if and only if it satisfies the conditions of XXIII.1.3.6.. There is a metric giving the uniformity $\mathcal{U}$ if G is first countable, for example

$$d'(g,h) = d(g,h) + d(g^{-1}, h^{-1})$$

if d is left-invariant; but d' is neither left- or right-invariant unless $\mathcal{L} = \mathcal{R} = \mathcal{U}$.

Completion

XXIII.1.3.9. Definition. A topological group G is *complete* if it is complete with respect to the uniformity $\mathcal{U}$.

A group which is complete in this sense is often called *Raikov-complete*.

XXIII.1.3.10. Proposition. Let G be a complete topological group, and H a subgroup. Then

(i) H is complete if and only if it is closed in G.

(ii) If H is normal, then G/H is complete.

XXIII.1.3.11. Theorem. Let G be a topological group. Then there is a complete topological group $\bar{G}$ containing G as a dense subgroup. $\bar{G}$ is unique up to topological isomorphism which is the identity on G. $\bar{G}$ is called the *completion*, or *Raikov completion*, of G.

XXIII.1.3.12. A topological group G does not generally have a completion with respect to $\mathcal{L}$ or $\mathcal{R}$ (more precisely, the completion of G with respect to $\mathcal{L}$ or $\mathcal{R}$ as a uniform space, called the *Weil completion* of G, does not have a compatible structure as a topological group in general); in fact, it has such a completion (as a topological group) if and only if $\mathcal{L} = \mathcal{R}$.

XXIII.1.4. Locally Compact Topological Groups

Topological groups which are locally compact and Hausdorff are especially important, and have a lot of nice associated structure which allows analysis to be done on them.

XXIII.1.4.1. Definition. A *locally compact group* is a locally compact Hausdorff topological group, i.e. a topological group whose topology is locally compact and Hausdorff. A *compact group* is a compact Hausdorff topological group.

Thus the words "Hausdorff" and "topological" will be understood in the names "locally compact group" and "compact group."

XXIII.1.4.2. Examples. (i) The additive group $\mathbb{R}$, or $\mathbb{R}^n$, with its usual topology, is a locally compact group.

(ii) The group $GL_n(\mathbb{R})$ (XXIII.1.2.3.(iii)) is locally compact (in its usual topology).

(iii) $\mathbb{T}$ is a compact group (in its usual topology). Any product of copies of $\mathbb{T}$ is a compact goup.

(iv) Any group with the discrete topology is a locally compact group.

(v) Any closed subgroup of a locally compact group is locally compact.

(vi) The additive group of an infinite-dimensional normed vector space is not a locally compact group in the norm topology.

(vii) The additive group $\mathbb{Q}$ is not a locally compact group in its usual topology (the relative topology from $\mathbb{R}$). But it is in the discrete topology.

By homogeneity, we obtain:

XXIII.1.4.3. Proposition. A Hausdorff topological group G is locally compact if and only if there is a compact neighborhood of e in G.

XXIII.1.4.4. Proposition. Let G be a locally compact group. Then G has an open subgroup which is σ-compact. If G is connected, it is σ-compact.

Proof: Let U be a neighborhood of e in G with $\bar{U}$ compact. Let H be the subgroup of G generated by U. Then H is open and is the subgroup of G generated by $\bar{U}$ (XXIII.1.2.33.). If $V = U \cup U^{-1}$, then $\bar{V}$ is compact and generates H. We have that $\bar{V}^2$ is compact since it is the image of $\bar{V} \times \bar{V}$ under multiplication; by induction, $\bar{V}^n$ is compact for all n. Thus $H = \cup_{n=1}^{\infty} \bar{V}^n$ is σ-compact. 👍

XXIII.1.4.5. Corollary. Let G be a locally compact group. Then G partitions into clopen subsets, each of which is σ-compact and thus Lindelöf. So G is paracompact and hence normal.

Proof: The cosets of a σ-compact open subgroup give such a partition. 👍

XXIII.1.4.6. Corollary. Let G be a locally compact group. The following are equivalent:

(i) G is second countable.

(ii) G is first countable and σ-compact.

(iii) G is first countable and compactly generated.

XXIII.1.4.7. Corollary. If G is a first countable (equivalently, metrizable) locally compact group, then G has an open subgroup which is second countable.

XXIII.1.4.8. Proposition. Let G be a topological group, K a compact subset of G, and U an open neighborhood of K. Then there is an open neighborhood V of e in G such that $KV \subseteq U$.

Proof: For each $g \in K$, $ge \in U$, so by joint continuity there are neighborhoods W_g of g and V_g of e with $W_g V_g \subseteq U$. The W_g cover K, so there is a finite subcover $\{W_{g_1}, \ldots, W_{g_n}\}$. Set $V = V_{g_1} \cap \cdots \cap V_{g_n}$. If $g \in K$, then $g \in W_{g_k}$ for some k, so $gV \subseteq W_{g_k} V \subseteq W_{g_k} V_{g_k} \subseteq U$. Thus $KV \subseteq U$. 👍

Similarly, we have:

XXIII.1.4.9. Proposition. Let G be a topological group, K a compact subset of G, and U an open neighborhood of e. Then there is an open neighborhood V of e in G such that $g^{-1}Vg \subseteq U$ for all $g \in K$.

Proof: For each $g \in K$, $g^{-1}eg = e$, so by joint continuity there are neighborhoods W_g of g and V_g of e with $h^{-1}V_g h \subseteq U$ for all $h \in W_g$. The W_g cover K, so there is a finite subcover $\{W_{g_1}, \ldots, W_{g_n}\}$. Set $V = V_{g_1} \cap \cdots \cap V_{g_n}$. If $g \in K$, then $g \in W_{a_k}$ for some k, so $g^{-1}Vg \subseteq g^{-1}V_{g_k}g \subseteq U$. 👍

XXIII.1.4.10. Proposition. Let G be a (necessarily locally compact) group. If U is a compact open neighborhood of e in G, then U contains a compact open subgroup of G.

Proof: There is a neighborhood V of e with $V = V^{-1}$ and $UV \subseteq U$ (XXIII.1.4.8. with $K = U$). Then $V^2 \subseteq U$, and by induction $V^n \subseteq U$. Thus the subgroup of G generated by V, which is open, is contained in U. 👍

XXIII.1.4.11. THEOREM. Let G be a locally compact group. Then every neighborhood of e in G contains a compact subgroup H such that G/H is metrizable. If G is compactly generated, the subgroup H may be chosen normal.

PROOF: First let U be a compact neighborhood of e in G, and let G_0 be the subgroup of G generated by U. Then G_0 is an open subgroup. If H is a subgroup of G_0, then G/H is a separated union of cosets of G_0/H; hence G/H is metrizable if G_0/H is metrizable. Thus it suffices to show the last statement, i.e. that if G is generated by a compact symmetric neighborhood U of e, then for any neighborhood V of e, G has a compact normal subgroup H contained in V with G/H metrizable (first countable).

Fix a neighborhood V, which we may take to be compact and symmetric. Set $V_1 = V$. Generate a decreasing sequence (V_n) of symmetric compact neighborhoods of e such that $g^{-1}V_{n+1}g \subseteq V_n$ for all $g \in U$ (XXIII.1.4.9.) and such that $V_{n+1}^2 \subseteq V_n$. Set $H = \cap_n V_n$. Then H is a compact subgroup of G, and $g^{-1}Hg \subseteq H$ for all $g \in U$; since U generates G, H is normal in G. If $\pi : G \to G/H$ is the quotient map, the sets $(\pi(V_n))$ are compact neighborhoods of the identity which form a local neighborhood base by XIII.11.1.13.. Thus G/H is first countable. 👍

XXIII.1.4.12. PROPOSITION. Every locally compact group is complete.

XXIII.1.4.13. PROPOSITION. Let G be a locally compact group and H a closed subgroup. Then G/H is locally compact in the quotient topology. So if H is normal, G/H is a locally compact group.

The next result is closely related to the Open Mapping Theorem for Banach spaces (), with a similar proof based on the Baire Category Theorem. This result is a special case of XXIII.1.6.2..

XXIII.1.4.14. THEOREM. [OPEN MAPPING THEOREM FOR LOCALLY COMPACT GROUPS] Let G and G' be locally compact groups, and ϕ a continuous group homomorphism from G onto G'. If G is σ-compact (so G' is also σ-compact), then ϕ is an open mapping, and thus G' is topologically isomorphic to $G/\ker(\phi)$.

As in the Banach space case (), we also get a Closed Graph Theorem for locally compact groups:

XXIII.1.4.15. THEOREM. [CLOSED GRAPH THEOREM FOR LOCALLY COMPACT GROUPS] Let G and G' be σ-compact locally compact groups, and $\phi : G \to G'$ a homomorphism. If the graph $\Gamma(\phi)$ is a closed subset of $G \times G'$, then ϕ is continuous.

PROOF: $G \times G'$ is locally compact and σ-compact, and $\Gamma(\phi)$ is a closed subgroup, hence also locally compact and σ-compact. The map $\pi_G : (x, \phi(x)) \mapsto x$ is a continuous bijective homomorphism from $\Gamma(\phi)$ onto G, hence a topological isomorphism by XXIII.1.4.14.; thus the inverse map ψ is continuous, and hence $\phi = \pi_{G'} \circ \psi$ is continuous. 👍

XXIII.1.4.16. Corollary. Let G be a group, and let $\mathcal{T}$ be a Hausdorff topology on G, not necessarily a topological group topology. Then there is at most one topology on G stronger than $\mathcal{T}$ making it into a σ-compact locally compact group.

Proof: Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be topologies stronger than $\mathcal{T}$ making G into a locally compact σ-compact group, and let ϕ be the identity map from $(G, \mathcal{T}_1)$ to $(G, \mathcal{T}_2)$. The graph of ϕ is the diagonal in $G \times G$. The product topology $\mathcal{T}_1 \times \mathcal{T}_2$ on $G \times G$ is stronger than the topology $\mathcal{T} \times \mathcal{T}$. Since $\mathcal{T}$ is Hausdorff, the diagonal is closed in $G \times G$ for the topology $\mathcal{T} \times \mathcal{T}$, and hence also closed in the topology $\mathcal{T}_1 \times \mathcal{T}_2$. Thus ϕ has closed graph and is therefore continuous, i.e. $\mathcal{T}_1$ is stronger than $\mathcal{T}_2$. The same argument with $\mathcal{T}_1$ and $\mathcal{T}_2$ interchanged shows that $\mathcal{T}_2$ is stronger than $\mathcal{T}_1$. 👍

XXIII.1.4.17. XXIII.1.4.14.–XXIII.1.4.16. are false in general without "σ-compact": the additive group $\mathbb{R}$ is a locally compact group in either the ordinary topology or the discrete topology. The $\mathcal{T}$ in XXIII.1.4.16. is also necessary in general if G is uncountable: $\mathbb{R}$ has discontinuous group-automorphisms (using AC).

XXIII.1.4.18. XXIII.1.4.16. does *not* say that a group has at most one σ-compact Hausdorff topology making it into a topological group. In fact, there is an abelian group G which has $2^{\aleph_0}$ mutually nonequivalent topological group topologies, each of which is compact and metrizable [Fuc59] (where two topological group topologies $\mathcal{T}_1$ and $\mathcal{T}_2$ on a group G are *equivalent* if $(G, \mathcal{T}_1)$ and $(G, \mathcal{T}_2)$ are isomorphic as topological groups). ✋

However, we do have:

XXIII.1.4.19. Corollary. The only locally compact group topology on a countable group is the discrete topology.

Any closed subgroup of a locally compact group is locally compact (). Conversely:

XXIII.1.4.20. Theorem. Let G be a Hausdorff topological group, and H a subgroup which is locally compact in the relative topology. Then H is a closed subgroup of G.

XXIII.1.5. Polish Groups and Baire Groups

XXIII.1.5.1. Definition. A *Polish group* is a topological group which is a Polish space (), i.e. homeomorphic to a separable complete metric space.

XXIII.1.5.2. There is a large overlap between the class of locally compact groups and the class of Polish groups: the overlap consists precisely of the second countable locally compact groups (XXIII.1.4.6.). The theory of Polish groups is to a large extent similar and parallel to the theory of locally compact groups. Except for one little thing: Polish groups which are not locally compact do not have Haar measure, so the most crucial tool for analysis is lacking. Thus the theory of Polish groups is not as satisfactory overall as that of locally compact groups.

There are many important topological groups which are Polish but not locally compact, for example:

XXIII.1.5.3. Examples. (i) The additive group of any separable Banach space is a Polish group.

(ii) The multiplicative group of invertible elements in any separable Banach algebra is a Polish group in the relative norm topology (),().

(iii) The group of unitary operators on a separable Hilbert space is a Polish group under the strong or weak operator topologies (which coincide on this group; cf. ()).

(iv) The group of homeomorphisms of a compact metric space is a Polish group in the topology of uniform convergence ().

(v) The group of isometries of a separable complete metric space is a Polish group in the topology of pointwise convergence (which coincides with the topology of u.c. convergence; cf. ()).

(vi) Any closed subgroup of a Polish group is a Polish group.

(vii) A countable product of Polish groups is a Polish group.

XXIII.1.5.4. There is one sense in which Polish groups are nicer than locally compact groups: there is a universal Polish group $\tilde{G}$, a Polish group with the property that every Polish group is topologically isomorphic to a closed subgroup of $\tilde{G}$. In fact, the group of homeomorphisms of the Hilbert cube with the topology of uniform convergence is such a universal group ([Usp86]; cf. [Kec95, 9.18]).

XXIII.1.6. Open Mappings and Automatic Continuity

In this section, we collect together some miscellaneous results about when surjective homomorphisms of topological groups are automatically open mappings, and when group operations and group actions are automatically continuous.

We first give simple criteria for when homomorphisms are continuous or open:

XXIII.1.6.1. Proposition. Let G and G' be topological groups, and $\phi : G \to G'$ a group-homomorphism. Then:

(i) If ϕ is continuous at one point, then it is continuous everywhere.

(ii) ϕ is an open mapping if and only if, for every open neighborhood U of e, $\phi(U)$ has nonempty interior (i.e. contains a nonempty open set).

Proof: (i): Suppose ϕ is continuous at g, i.e. if $g_i \to g$, then $\phi(g_i) \to \phi(g)$. If $x_i, x \in G$ and $x_i \to x$, then $gx^{-1}x_i \to g$, so

$$\phi(g)\phi(x)^{-1}\phi(x_i) = \phi(gx^{-1}x_i) \to \phi(g)$$

and thus

$$\phi(x_i) = \phi(x)\phi(g)^{-1}\phi(g)\phi(x)^{-1}\phi(x_i) \to \phi(x)\phi(g)^{-1}\phi(g) = \phi(x) \ .$$



The first main result is a general version of XXIII.1.4.14.:

XXIII.1.6.2. THEOREM. [OPEN MAPPING THEOREM FOR σ-COMPACT GROUPS] Let G and G' be topological groups, with G' a Baire group, and ϕ a continuous group homomorphism from G onto G'. If G is σ-compact (so G' is also σ-compact), then ϕ is an open mapping, and thus G' is topologically isomorphic to $G/\ker(\phi)$.

PROOF: The homomorphism ϕ factors topologically through $G/\ker(\phi)$, so we may assume without loss of generality that ϕ is a bijection. Let $G = \cup_{n=1}^{\infty} K_n$, with K_n compact. Then $G' = \cup_{n=1}^{\infty} \phi(K_n)$, and $\phi(K_n)$ is compact and hence closed in G' since G' is Hausdorff. By the Baire Category Theorem, some $\phi(K_n)$ has nonempty interior V. The restriction of ϕ to K_n is a homeomorphism, so ϕ^{-1} is continuous on V since V is open. Since V is nonempty, ϕ^{-1} is continuous (XXIII.1.6.1.(i)), i.e. ϕ is an open map. 👍

XXIII.1.6.3. Note that no hypothesis on G beyond σ-compactness is needed, not even an assumption that G is Hausdorff. This proof is attributed to H. GLÖCKNER, who observed that the local compactness hypothesis on G in many references is not necessary.

XXIII.1.6.4. COROLLARY. Let G be a countable group. Then the only Polish topological group topology on G is the discrete topology.

XXIII.1.7. Exercises

XXIII.1.7.1. Let G be the additive group of real numbers. Give G the Sorgenfrey topology (XIII.7.10.14.). Show that the group operation (addition) is jointly continuous (i.e. continuous from $G \times G$ to G), but that inversion (negation) is not continuous. Thus $(\mathbb{R}, +)$ with this topology is not a topological group.

XXIII.1.7.2. Let G be a topological group, and A a subset of G. Show that the closure of A is contained in AV for any neighborhood V of e, and in particular $\bar{U} \subseteq U^2$ if U is a neighborhood of e in G. [If $g \in \bar{A}$, then $gV^{-1} \cap A \neq \emptyset$.]

XXIII.1.7.3. (a) Use XXIII.1.7.2. and XXIII.1.2.15.(i) to give a direct proof that a topological group is T_3 (cf. XIII.7.5.1.).

(b) Give a direct proof that a T_0 topological group is Hausdorff.

XXIII.1.7.4. (a) Let G be a topological group and H a subgroup. If H contains a neighborhood of e, show that H is open. Thus the hypothesis in XXIII.1.2.33. that U be an open neighborhood can be relaxed to just requiring that U be a neighborhood.

(b) Show that a closed subgroup of a topological group is either open or nowhere dense.

XXIII.1.7.5. Let G be a topological group, and U an open neighborhood of e in G. Set $W = \cup_{n=1}^{\infty} U^n$.

(a) Show that W is a clopen subset of G.

(b) W need not be a subgroup of G. [Consider $\{0, 1\} \subseteq \mathbb{Z}$ with the discrete topology.]

(c) If G is connected, then $W = G$.

XXIII.1.7.6. How much of the theory of topological groups requires joint continuity of multiplication? That is, if we change the definition of a topological group G to only require continuity of inversion and separate continuity of multiplication (i.e. that ι and λ_g, ρ_g are continuous for all $g \in G$), how much of the theory persists? [Consider the case where G is an infinite group with the finite complement topology.]

XXIII.1.7.7. For a topological space X to have a group structure making it a topological group, i.e. for X to be homeomorphic to a topological group, it is necessary that X be homogeneous. But this is far from sufficient, even if X is compact and metrizable.

(a) The Hilbert cube $\mathbb{H}$ is homogeneous (), but not homeomorphic to a topological group. [Every continuous function from $\mathbb{H}$ to $\mathbb{H}$ has a fixed point ().]

(b) S^n for even $n > 0$ is not homeomorphic to a topological group. [Use ().]

In fact, S^n is homeomorphic to a topological group if and only if n is 0, 1, or 3 (), ().

Groups of Homeomorphisms

If X is a topological space, the group $\mathrm{Homeo}(X)$ of homeomorphisms of X can be given various topologies, e.g. the topology of pointwise convergence or the compact-open topology (which is the topology of u.c. convergence if X is metrizable ()). $\mathrm{Homeo}(X)$ is sometimes, but not always, a topological group under these topologies. The next several problems explore when $\mathrm{Homeo}(X)$, or various natural subgroups, is or is not a topological group.

XXIII.1.7.8. Let X be a Hausdorff space. Give $\mathrm{Homeo}(X)$ the compact-open topology. Show that multiplication is jointly continuous.

XXIII.1.7.9. Let X be a compact Hausdorff space. Show that $\mathrm{Homeo}(X)$ is a topological group under the compact-open topology (which is the topology of uniform convergence).

XXIII.1.7.10. [Are46a] Let X be the space $\mathbb{N}_0 \cup \{\frac{1}{n} : n \in \mathbb{N}\}$, with its relative topology from $\mathbb{R}$. Then X is a locally compact metrizable space. Show that inversion is not continuous in $\mathrm{Homeo}(X)$. [Consider ϕ_n, where $\phi_n(n) = \frac{1}{n}$, $\phi_n(m) = m - 1$ if $m > n$, $\phi_n\left(\frac{1}{m}\right) = \frac{1}{m+1}$ if $m \geq n$, and $\phi_n(x) = x$ otherwise.]

It is shown in [Are46a] that if X is locally compact and locally connected, then inversion in $\mathrm{Homeo}(X)$ is continuous in the compact-open topology, and hence $\mathrm{Homeo}(X)$ is a topological group in this topology.

XXIII.1.7.11. Let X be a topological space. Give $\mathrm{Homeo}(X)$ the topology of pointwise convergence. Then multiplication is separately continuous.

XXIII.1.7.12. Give $\mathrm{Homeo}(\mathbb{R}^2)$ the topology of pointwise convergence. Show that multiplication is not jointly continuous, and that inversion is not continuous.

XXIII.1.7.13. Let $X = \mathbb{R}$ or $[0, 1]$. Show that the topology of pointwise convergence on $\mathrm{Homeo}(X)$ coincides with the compact-open topology (topology of u.c. convergence), and that $\mathrm{Homeo}(X)$ is a topological group under this topology.

XXIII.1.7.14. Let X be a metric space. Show that the group $\mathrm{Isom}(X)$ of isometries (isometric bijections) of X is a topological group in the topology of pointwise convergence.

XXIII.1.7.15. Group Topologies on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. If G is a group and H a (Hausdorff) topological group, and $\phi : G \to H$ is an injective homomorphism, G can be identified with $\phi(G)$ and given the relative topology, which makes G into a (Hausdorff) topological group. In particular, if H is a (Hausdorff) topological group and x is an element of H of infinite order, the map $n \mapsto x^n$ from $\mathbb{Z}$ to H can be used to give a (Hausdorff) topological group topology on $\mathbb{Z}$. This gives an enormous variety of nonequivalent (XXIII.1.4.18.) group topologies on $\mathbb{Z}$.

(a) Let S be a nonempty subset of $\mathbb{R}$. If $G = \mathbb{Z}, \mathbb{Q}$, or $\mathbb{R}$, map G into $\prod_{s\in\mathcal{S}} \mathbb{T}$ by sending $x \in G$ to the point $\phi_S(x)$ with s-coordinate $e^{2\pi isx}$. Show that if $G = \mathbb{Z}$ or $\mathbb{Q}$, then ϕ_S is injective unless S is contained in a cyclic subgroup of $(\mathbb{Q}, +)$, and if $G = \mathbb{R}$, then ϕ_S is injective unless S is contained in a cyclic subgroup of $(\mathbb{R}, +)$. In these cases, give G the relative topology $\mathcal{T}_S$. The topology $\mathcal{T}_S$ is metrizable (first countable) if and only if S is countable.

(b) Let $\mathcal{B}$ be a Hamel basis for $\mathbb{R}$ over $\mathbb{Q}$ consisting of irrational numbers. If S_1, S_2 are distinct nonempty subsets of $\mathcal{B}$ (not singletons if $G = \mathbb{R}$), show that the topologies $\mathcal{T}_{S_1}$ and $\mathcal{T}_{S_2}$ are nonequivalent. Conclude that there are $2^{2^{\aleph_0}}$ mutually nonequivalent Hausdorff topologies on G making $(G, +)$ a topological group. These topologies are all totally bounded, i.e. the completion is compact. These topologies are all weaker than the ordinary topology on G.

(c) The topologies $\mathcal{T}_S$ are by no means all the Hausdorff topological group topologies on $(\mathbb{Q}, +)$: for example, the discrete topology, the relative topology from $\mathbb{R}$, and the $\boldsymbol{n}$-adic topologies (III.5.3.9.) are not among them. (The completions with respect to these topologies are locally compact but noncompact.) But there are exactly $2^{2^{\aleph_0}}$ distinct topologies on $\mathbb{Q}$. The only locally compact or Polish topological group topology on $(\mathbb{Q}, +)$ is the discrete topology. Similar statements are true for $(\mathbb{Z}, +)$.

(d) If G is any infinite abelian group of cardinality κ, then G has 2^{2^κ} mutually nonequivalent Hausdorff topological group topologies [Pod77]. In particular, $(\mathbb{R}, +)$ has $2^{2^{2^{\aleph_0}}}$ mutually nonequivalent Hausdorff topological group topologies. Using that $(\mathbb{R}, +)$ is isomorphic to the direct sum of $2^{\aleph_0}$ copies of $(\mathbb{Q}, +)$, construct from (b) for $\mathbb{Q}$ a family of $2^{2^{2^{\aleph_0}}}$ mutually nonequivalent totally bounded Hausdorff topological group topologies on $(\mathbb{R}, +)$. The additive group of any separable Banach space is also isomorphic to $\mathbb{R}$.

XXIII.2. Haar Measure

The most important tool for analysis on locally compact groups is Haar measure. Every locally compact group has an essentially unique translation-invariant measure called Haar measure. Haar measure generalizes Lebesgue measure on $\mathbb{R}$ or $\mathbb{R}^n$ and is of similar use and importance. And the fact that topological groups which are not locally compact do not have a similar translation-invariant measure is a serious obstacle to their study.

XXIII.2.1. Definition and Existence

XXIII.2.1.1. Definition. Let G be a Hausdorff topological group. A *left Haar measure* on G is a measure μ defined on the Baire sets in G, with the following properties:

(i) $\mu(U) > 0$ for every nonempty open Baire set U in G.

(ii) $\mu(K) < \infty$ for every compact G_δ K in G.

(iii) μ is *left invariant*: $\mu(\lambda_g(A)) = \mu(A)$ for every $g \in G$ and every Baire set A in G.

A *right Haar measure* on G is a measure μ on the Baire sets in G satisfying (i), (ii), and

(iii′) μ is *right invariant*: $\mu(\rho_g(A)) = \mu(A)$ for every $g \in G$ and every Baire set A in G.

If G is abelian, there is no difference between a left Haar measure and a right Haar measure on G. However, if G is not abelian a left Haar measure need not be a right Haar measure, and conversely (XXIII.2.2.6.(c)). If G is second countable, the Baire sets are the same as the Borel sets in G, so (i) holds for all open sets and (ii) holds for all compact sets in G.

Conditions (i) and (ii) are nondegeneracy conditions: without (i) the zero measure would be a left and right Haar measure, and without (ii) the measure in which every nonempty set has infinite measure would be a left and right Haar measure.

XXIII.2.1.2. Examples. (i) Lebesgue measure is a (left and right) Haar measure on $\mathbb{R}$ or $\mathbb{R}^n$ with addition.

(ii) A (left or right) Haar measure on $\mathbb{T}$ can be described in several equivalent ways. One simple way is to identify it with $[0, 1)$ by $t \leftrightarrow e^{2\pi it}$ and transfer Lebesgue measure on $[0, 1)$ to $\mathbb{T}$. One could also take arc length measure on $\mathbb{T}$ (or 1-dimensional Hausdorff measure) to obtain the same measure up to a factor of 2π.

(iii) Counting measure is a (left or right) Haar measure on any discrete group.

(iv) Any positive scalar multiple of a left Haar measure on G is a left Haar measure, and similarly for right Haar measures.

The main theorem is:

XXIII.2.1.3. Theorem. Let G be a locally compact group. Then there is a left Haar measure on G, which is unique up to positive scalar multiple. There is also a right Haar measure on G, unique up to a positive scalar multiple. Conversely, any (Hausdorff) topological group with a left or right Haar measure is locally compact.

XXIII.2.2. Haar Integrals

By the results of (), there is an entirely equivalent formulation of Haar measure as an integral or as a positive linear functional.

XXIII.2.2.1. DEFINITION. Let G be a locally compact group. A *left Haar integral* on G is a nonzero positive linear functional I on $C_c(G)$ which is *left invariant*: $I(g \cdot \phi) = I(\phi)$ for all $g \in G$, $\phi \in C_c(G)$, where $(g \cdot \phi)(x) = \phi(gx)$, i.e. $g \cdot \phi = \phi \circ \lambda_g$. A *right Haar integral* is a nonzero positive linear functional I on $C_c(G)$ which is *right invariant*: $I(\phi \cdot g) = I(\phi)$ for all $g \in G$, $\phi \in C_c(G)$, where $(\phi \cdot g)(x) = \phi(xg)$, i.e. $\phi \cdot g = \phi \circ \rho_g$.

XXIII.2.2.2. By the Riesz Representation Theorem (), a positive linear functional on $C_c(G)$ is given by integration with respect to a Baire measure. So if I is a left Haar integral on G, then

$$I(\phi) = \int_G \phi \, d\mu$$

for a Baire measure μ.

XXIII.2.2.3. PROPOSITION. If I is a left Haar integral on G, then μ is a left Haar measure on G, and similarly for a right Haar integral.

Conversely:

XXIII.2.2.4. PROPOSITION. If μ is a left Haar measure on G, then I defined by $I(\phi) = \int_G \phi \, d\mu$ is a left Haar integral on G, and similarly for a right Haar measure.

Thus Theorem XXIII.2.1.3. is equivalent to:

XXIII.2.2.5. THEOREM. Let G be a locally compact group. Then there is a left Haar integral on G, which is unique up to positive scalar multiple. There is also a right Haar integral on G, unique up to a positive scalar multiple.

The Haar integral is easier and more elementary to describe and find than the Haar measure in many cases:

XXIII.2.2.6. EXAMPLES. (i) For $\mathbb{R}$ with addition, the Haar integral (left and right) is simply given by Riemann integration: if $\phi \in C_c(\mathbb{R})$,

$$I(\phi) = \int_{-\infty}^{+\infty} \phi(t) \, dt \ .$$

This is not really an improper Riemann integral since ϕ has compact support: it is an ordinary Riemann (or even Cauchy) integral over a finite interval, which varies from function to function.

(ii) The (normalized) Haar integral on $\mathbb{T}$ is also given by Riemann integration: if $\phi \in C_c(\mathbb{T}) = C(\mathbb{T})$, then

$$I(\phi) = \int_0^1 \phi(e^{2\pi i t})\, dt \ .$$

(iii) Here is perhaps the simplest example of a locally compact group for which left and right Haar measure (or integral) are different: the real "$ax + b$ group." Let

$$G = \left\{ \begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} : a, b \in \mathbb{R}, a > 0 \right\} \ .$$

G is a subgroup of $GL(2, \mathbb{R})$, i.e. is a group under matrix multiplication. G can be identified as a set with the open right half-plane $\mathbb{R}_+ \times \mathbb{R}$ in $\mathbb{R}^2$ by

$$\begin{bmatrix} a & b \\ 0 & 1 \end{bmatrix} \longleftrightarrow (a, b)$$

and given the topology from this identification (which agrees with the subspace topology from $GL(2, \mathbb{R})$ regarded as an open subset of $M_2(\mathbb{R}) \cong \mathbb{R}^4$). From this point of view, the group operations are given by

$$(a, b)(x, y) = (ax, ay + b)$$

$$(a, b)^{-1} = \left(\frac{1}{a}, -\frac{b}{a}\right)$$

(cf. ()); the identity is $(1, 0)$. Since these operations are continuous, G is a locally compact group (it is even a Lie group).

If $g = (a, b) \in G$, left multiplication by g is an affine map $(x, y) \mapsto (ax, ay + b)$ which is differentiable with constant derivative $\begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ and Jacobian a^2.

Consider the linear functional I on G given for $\phi \in C_c(G)$ by

$$I(\phi) = \int_G \frac{\phi(x, y)}{x^2}\, dA = \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(x, y)}{x^2}\, dy\, dx$$

(since the support of ϕ is a compact subset of the open right half-plane, the integral is only really over a sufficiently large closed rectangle on which $\frac{1}{x^2}$ is bounded, hence is a proper Riemann integral). Then I is a nonzero positive linear functional on $C_c(G)$. If $g = (a, b) \in G$, then by the change-of-variables formula, writing $u = ax$, $v = ay + b$, we have

$$I(g \cdot \phi) = \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(ax, ay + b)}{x^2}\, dy\, dx = \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(ax, ay + b)}{a^2 x^2} a^2\, dy\, dx$$

$$= \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(u, v)}{u^2}\, dv\, du = I(\phi)$$

(the integrals are actually over suitably large compact regions in G; cf. XXIII.2.9.3.). Thus I is a left Haar integral on G. The corresponding left Haar measure is given by

$$\mu(A) = \int_A \frac{1}{x^2}\, d\lambda(x, y)$$

for A a Borel set in G, where λ is Lebesgue measure on $\mathbb{R}^2$.

Similarly, right multiplication by $g = (a, b)$ is given by $(x, y) \mapsto (xa, xb + y)$, with derivative $\begin{bmatrix} a & 0 \\ b & 1 \end{bmatrix}$ and Jacobian a. Thus the functional

$$I_r(\phi) = \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi(x, y)}{x} \, dy \, dx$$

is a right Haar integral on G, with right Haar measure

$$\mu(A) = \int_A \frac{1}{x} \, d\lambda(x, y)$$

for A Borel. Thus left and right Haar measures on G are different (XXIII.2.9.3.).

G is called the "$ax + b$ group" since it is isomorphic to the group of all order-preserving invertible affine transformations of $\mathbb{R}$, i.e. transformations of the form $f(x) = ax + b$, $a > 0$. One can similarly take the subgroup

$$\left\{ \begin{bmatrix} A & b \\ 0 & 1 \end{bmatrix} : A \in GL_n(\mathbb{R}), b \in \mathbb{R}^n \right\}$$

of $GL_{n+1}(\mathbb{R})$ of affine transformations of $\mathbb{R}^n$; restrict to such matrices where $det(A) > 0$ to get the connected component of the identity (orientation-preserving affine transformations), or to $A \in SO(n)$ to get the orientation-preserving affine isometries (rigid motions, cf. III.12.1.6.) of $\mathbb{R}^n$. There is a similar "$ax + b$ group" for any vector space over any field, or even modules over rings; some of these have topologies making them interesting topological groups (see e.g. [Bla83]).

XXIII.2.3. Invariant Measures for Group Actions

XXIII.2.4. Properties of Haar Measure

XXIII.2.5. The Modular Function

XXIII.2.6. Measures on Homogeneous Spaces

XXIII.2.7. Extending Haar Measure

XXIII.2.8. Measurable Groups

XXIII.2.9. Exercises

XXIII.2.9.1. Let G be a locally compact group.

(a) Show that there is a neighborhood of e in G which is a compact G_δ, which contains a neighborhood of e which is an open Baire set. [Use that G is completely regular.]

(b) Show that if μ is a left Haar measure on G, then there is a nonempty open Baire set V in G with $\mu(V) < \infty$.

(c) Conversely, if μ is a left invariant Baire measure on G for which there is a nonempty open Baire set V with $\mu(V) < \infty$, then $\mu(K) < \infty$ for every compact G_δ K. [K can be covered by finitely many left translates of V.]

Thus condition (ii) in the definition of a left Haar measure can be replaced by

(ii′) There is a nonempty open Baire set V in G with $\mu(V) < \infty$.

The same change can be made in the definition of a right Haar measure.

XXIII.2.9.2. Here is a warmup for Example XXIII.2.2.6.(c). Let $G = \mathbb{R}_+$ be the positive real numbers with multiplication.

(a) Show that the linear functional on $C_c(G)$ given for $\phi \in C_c(G)$ by

$$I(\phi) = \int_0^\infty \frac{\phi(x)}{x}\, dx$$

is (left and right) invariant, hence is a Haar integral. Write down the corresponding Haar measure.

(b) Show the result of (a) alternately by noting that $x \mapsto e^x$ is a topological isomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}_+, \cdot)$ and transferring the Haar integral on $\mathbb{R}$ to $\mathbb{R}_+$ using the change-of-variables formula.

XXIII.2.9.3. (a) In example XXIII.2.2.6.(c), give a more careful argument about regions of integration in the application of the change-of-variables formula in showing I is left invariant.

(b) Show by example that I is not right invariant, i.e. that the left and right Haar integrals for G are truly different. (Caution: this involves more than just showing that $I \neq I_r$.)

XXIII.2.9.4. (a) Let $G = GL(2, \mathbb{R})$. Identify G with an open set in $\mathbb{R}^4$ by

$$\begin{bmatrix} x & y \\ z & w \end{bmatrix} \longleftrightarrow (x, y, z, w)\ .$$

As in XXIII.2.2.6.(c), show that the linear functional defined for $\phi \in C_c(G)$ by

$$I(\phi) = \int_G \frac{\phi(x, y, z, w)}{(xw - yz)^2}\, d\lambda(x, y, z, w)$$

(the integral is actually a Riemann integral over a suitable compact region) is both a left and right Haar integral for G. Note that the denominator is the square of the determinant.

(b) Show similarly that if $G = GL(n, \mathbb{R})$, identified with an open set in $\mathbb{R}^{n^2}$, the left and right Haar integral is given by

$$I(\phi) = \int_G \frac{\phi(x_{ij})}{|Det(x_{ij})|^n}\, d\lambda(x_{ij})$$

where (x_{ij}) symbolically denotes the n^2-tuple $(x_{11}, \ldots, x_{nn})$, or the corresponding matrix. (Note that the case $n = 1$ is XXIII.2.9.2..)

XXIII.3. Group Algebras and Unitary Representations

One of the most important applications of the existence of Haar measure on locally compact groups is that certain Banach algebras can be associated to any locally compact group which have the "same" representation theory as the group. These Banach algebras are a crucial tool in showing that locally compact groups have a nice representation theory, and in developing a theory of analysis on locally compact groups (*abstract harmonic analysis*).

XXIII.3.1. The Group Algebra

The simplest algebra, the prototype for the others, is the group algebra.

XXIII.3.1.1. DEFINITION. Let G be a group (with no topology, or the discrete topology). The *group algebra* of G is the set $\mathbb{C}G$ of formal finite complex linear combinations

$$\mathbb{C}G = \left\{ f = \sum_{x \in G} \alpha_x \cdot x \right\}$$

of elements of G, where $\alpha_x \in \mathbb{C}$ for each x and only finitely many α_x are nonzero. The sum and product are defined in the obvious way:

$$\left(\sum_{x \in G} \alpha_x \cdot x \right) + \left(\sum_{x \in G} \beta_x \cdot x \right) = \sum_{x \in G} (\alpha_x + \beta_x) \cdot x$$

$$\left(\sum_{x \in G} \alpha_x \cdot x \right) \left(\sum_{x \in G} \beta_x \cdot x \right) = \sum_{y \in G} \sum_{z \in G} (\alpha_y \beta_z) \cdot yz = \sum_{x \in G} \left(\sum_{yz = x} \alpha_y \beta_z \right) \cdot x = \sum_{x \in G} \left(\sum_{y \in G} \alpha_y \beta_{y^{-1}x} \right) \cdot x$$

(all the sums are really finite sums, so rearrangement and grouping is legitimate).

XXIII.3.1.2. $\mathbb{C}G$ is clearly an algebra over $\mathbb{C}$, with unit $\sum_{x \in G} \alpha_x \cdot x$, where $\alpha_e = 1$ and $\alpha_x = 0$ for $x \neq e$. $\mathbb{C}G$ is commutative if and only if G is abelian.

XXIII.3.1.3. $\mathbb{C}G$ is perhaps better thought of as the set of all functions from G to $\mathbb{C}$ with finite support, identifying $f = \sum_{x \in G} \alpha_x \cdot x$ with the function $f(x) = \alpha_x$. The addition then is just usual addition of functions, and multiplication is "convolution":

$$(f * g)(x) = \sum_{y \in G} f(y) g(y^{-1}x) \ .$$

The unit is then χ_e. There is a natural subgroup of the group of units of $\mathbb{C}G$ isomorphic to G: $x \mapsto \chi_x$.

Representations

In its most general form, a *representation* of a group G on a set X is a homomorphism from G to the group of bijections of X (i.e. representing G as a group of transformations of X). If X has some structure (e.g. linear and/or topological), it is normally required that representations preserve the structure. Most commonly, X is a vector space and the bijections are required to be linear:

XXIII.3.1.4. DEFINITION. Let G be a group and V a vector space over $\mathbb{F}$. A *representation* of G on V is a homomorphism from G to the group of invertible linear transformations of V.

We will only consider the case $\mathbb{F} = \mathbb{C}$, i.e. representations on complex vector spaces.

Similarly, we may define a representation of an algebra:

XXIII.3.1.5. DEFINITION. Let A be a (unital) algebra over $\mathbb{F}$ and V a vector space over $\mathbb{F}$. A *representation* of A on V is a (unital) homomorphism from A to $\mathcal{L}(V)$.

We will again only consider the case $\mathbb{F} = \mathbb{C}$. In this section, we will only consider unital representations; later nonunital algebras and representations will be used, in which case there are some additional technicalities needed.

XXIII.3.1.6. Note that in XXIII.3.1.4. and XXIII.3.1.5., the representation is not required to be injective. An injective representation is called *faithful*.

One of the most important uses of the group algebra is in studying representations of the group:

XXIII.3.1.7. PROPOSITION. Let G be a group and V a complex vector space. There is a natural one-one correspondence between representations of G on V and (unital) representations of $\mathbb{C}G$ on V: if G is regarded as a subgroup of the group of units of $\mathbb{C}G$, then any unital representation of $\mathbb{C}G$ on V restricts to a representation of G on V, and every representation of G on V extends uniquely to a representation of $\mathbb{C}G$ on V. The two are inverse processes.

PROOF: This is almost obvious. If π is a representation of G on V, extend π to $\mathbb{C}G$ by setting

$$\pi\left(\sum_{x\in G} \alpha_x \cdot x\right) = \sum_{x\in G} \alpha_x \pi(x)$$

(the sums are really finite sums). The algebra homomorphism property of the extended π is routine to check, as are the rest of the assertions. 👍

XXIII.3.1.8. EXAMPLES. (i) Let G be any group. The representation ι of G on $\mathbb{C}$ with $\iota(x) = 1$ for all $x \in G$ is called the *trivial representation* of G. The corresponding representation of $\mathbb{C}G$ is the homomorphism $\iota : \mathbb{C}G \to \mathbb{C}$ given by $\iota(f) = f(e)$.

(ii) If G is any group, there are two important representations of G on $\mathbb{C}G$. For $x \in G$, define $\lambda(x) \in \mathcal{L}(\mathbb{C}G)$ by

$$[\lambda(x)f](y) = f(x^{-1}y) \ .$$

This λ is called the *left regular representation* of G. Similarly, define $\rho(x) \in \mathcal{L}(\mathbb{C}G)$ by

$$[\rho(x)f](y) = f(yx) \ .$$

This ρ is the *right regular representation* of G. (Putting inverses in the right places in these formulas to make them representations can be a delicate matter.) The corresponding representations of $\mathbb{C}G$ on $\mathbb{C}G$ are by left and right convolution: $\lambda(f)g = f * g$ and $\rho(f)g = g * \tilde{f}$, where $\tilde{f}(x) = f(x^{-1})$.

XXIII.3.1.9. There is an extensive algebraic theory of group representations, especially of finite groups, which we will not describe since it is really algebra, not analysis. Some of this theory will be subsumed in the theory of unitary representations of locally compact groups, described below.

The *-Algebra Structure of the Group Algebra

The group algebra $\mathbb{C}G$ has a natural structure as a *-algebra:

XXIII.3.1.10. DEFINITION. Let G be a group, with group algebra $\mathbb{C}G$. Define an involution on $\mathbb{C}G$ by

$$f^*(x) = \overline{f(x^{-1})} \ .$$

XXIII.3.1.11. The following properties of the involution are easily established for all $f, g \in \mathbb{C}G$, $\alpha \in \mathbb{C}$:

(i) $(f+g)^* = f^* + g^*$

(ii) $(\alpha f)^* = \bar{\alpha} f^*$

(iii) $(f * g)^* = g^* * f^*$

(iv) $(f^*)^* = f$.

[Property (iii) is a graphic example of the overuse of $*$ in functional analysis!]

With this involution, $\mathbb{C}G$ becomes a *-algebra in the sense of (). The elements of G, regarded as units in $\mathbb{C}G$, are "unitaries": $\chi_{x^{-1}} = (\chi_x)^{-1} = (\chi_x)^*$ for any $x \in G$.

*-Representations of *-Algebras

XXIII.3.1.12. DEFINITION. Let A be a (unital) *-algebra and V a complex inner product space. A **-representation* of A on V is a *-homomorphism π from A to $\mathcal{B}(V)$ with its involution from the inner product, i.e. π is an algebra-homomorphism with $\pi(x^*) = \pi(x)^*$ for all $f \in A$.

We usually only consider *-representations of *-algebras on Hilbert spaces, but occasionally we will want to use *-representations on incomplete inner product spaces. The additional generality is inconsequential, since any bounded operator on a normed vector space extends uniquely to the completion (). We rarely consider representations of *-algebras which are not *-representations, and occasionally "representation" will be taken to mean "*-representation" when there is no possibility of confusion (but we will try to be careful, if sometimes pedantic!)

Unitary Representations and *-Representations

We have the following analog of XXIII.3.1.4.:

XXIII.3.1.13. DEFINITION. Let G be a group and V a complex inner product space. A *unitary representation* of G is a homomorphism from G to the group $\mathcal{U}(V)$ of unitary operators on V ().

Again, we usually only consider unitary representations of groups on Hilbert spaces, but occasionally we will want to use unitary representations on incomplete inner product spaces.

XXIII.3.1.14. Unitary representations of groups are especially important for two reasons: (1) they do indeed arise naturally in many applications, and (2) unlike the theory of general representations which is largely intractable for infinite groups (general representations essentially reduce to unitary representations for finite groups), an extensive and manageable theory of unitary representations of general groups, and even of locally compact topological groups, can be developed.

Then we have the following important analogy of XXIII.3.1.7., whose proof is straightforward:

XXIII.3.1.15. PROPOSITION. Let G be a group and V a complex inner product space. There is a natural one-one correspondence between unitary representations of G on V and (unital) *-representations of $\mathbb{C}G$ on V: if G is regarded as a subgroup of the group of units of $\mathbb{C}G$, then any unital *-representation of $\mathbb{C}G$ on V restricts to a unitary representation of G on V, and every unitary representation of G on V extends uniquely to a *-representation of $\mathbb{C}G$ on V. The two are inverse processes.

XXIII.3.1.16. A faithful *-representation of $\mathbb{C}G$ evidently corresponds to a faithful unitary representation of G. But a faithful unitary representation of G does not necessarily correspond to a faithful *-representation of $\mathbb{C}G$. For example, if G is any subgroup of the circle group $\mathbb{T}$ (e.g. any cyclic group), the identity map on G gives a faithful one-dimensional unitary representation of G, but the corresponding *-representation of $\mathbb{C}G$ cannot be faithful if G is nontrivial. For this and many other reasons (e.g. ()) it is important to consider nonfaithful representations.

XXIII.3.2. The L^1 Algebra and the C*-Algebra of a Locally Compact Group

For (nondiscrete) topological groups, the group algebra is an inadequate object because it does not take the topology into consideration. Even for infinite discrete groups, it is useful to use a completion of the group algebra. In this section we show how to remedy both shortcomings for locally compact groups, ultimately associating to each locally compact group a group C*-algebra with the same representation theory.

XXIII.3.2.1. If G is a locally compact group, the most direct analog of the group algebra is $C_c(G)$, the set of complex-valued continuous functions on G with compact support. Note that if G has the discrete topology, then $C_c(G)$ is just $\mathbb{C}G$.

$C_c(G)$ is an algebra with pointwise multiplication, but we instead want to use convolution. To define convolution we use Haar measure. Fix a left Haar measure λ on G, and define

$$(f * g)(x) = \int_G f(y)g(y^{-1}x)\, d\lambda(y)$$

for $f, g \in C_c(G)$.

XXIII.3.2.2. PROPOSITION. Convolution is an associative binary operation of $C_c(G)$ making $C_c(G)$ into an algebra over $\mathbb{C}$.

XXIII.3.2.3. The definition of $C_c(G)$ as an algebra depends on the choice of left Haar measure λ. But if another left Haar measure μ is chosen, μ is just a scalar multiple of λ and an isomorphic algebra is obtained using μ (). If G is discrete and λ is chosen to be counting measure on G, then the multiplication so defined on $C_c(G) = \mathbb{C}G$ is just the standard convolution on $\mathbb{C}G$ defined in XXIII.3.1.1..

XXIII.3.2.4. **Caution:** It is common in the literature on representation theory to say that a locally compact group G is "separable" if $L^1(G)$ (or, equivalently, $C^*(G)$) is separable, i.e. if G is second countable. However, this use of the term "separable" can be misleading, since a locally compact group, even a compact abelian group, which is separable in the topological sense of having a countable dense subset is not necessarily "separable" in this sense (a product of $2^{\aleph_0}$ circle groups or nontrivial finite groups is a counterexample; cf. XIII.6.3.3.). Thus we will avoid using the term "separable" in any sense for topological groups.

XXIII.4. Lie Groups

A Lie group is a group which is a smooth manifold and whose multiplication and inversion are smooth maps. Since a smooth manifold has an underlying structure as a topological space and smooth functions are continuous, a Lie group has an underlying structure as a topological group (although it is technically not quite correct to say that a Lie group "is" a topological group), which is locally compact. Many of the most familiar and important topological groups are Lie groups. It turns out that every topological group whose topology is locally Euclidean has a unique structure as a Lie group. The theory of Lie groups bridges geometry, topology, linear and abstract algebra, and analysis, and impacts much of modern mathematics.

Lie groups are named in honor of Sophus Lie; note that "Lie" is pronounced "Lee."

> "In recent times the view becomes more and more prevalent that many branches of mathematics are nothing but the theory of invariants of special groups."
>
> *Sophus Lie*[2]

(This view was widely held in the late 19th/early 20th century, but has tempered since and only applies literally to some parts of modern mathematics, although it is still a good guiding principle in sufficiently general form.)

XXIII.4.1. The Classical Groups

XXIII.4.2. Lie Groups and Lie Algebras

XXIII.4.3. Locally Euclidean Groups

XXIII.4.3.1. Definition. A *locally Euclidean group* is a topological group whose topology is locally Euclidean (XIII.20.1.1.).

XXIII.4.3.2. A locally Euclidean space is T_1 (XIII.20.1.2.(viii)), and thus a locally Euclidean group G is Hausdorff, hence locally compact. By XIII.20.1.2.(vi), XXIII.1.2.36., and XXIII.1.4.4., the connected component of the identity is an open normal subgroup of G which is σ-compact. Thus, as a topological space, a locally Euclidean group is a topological manifold (and in particular metrizable; cf. XXIII.1.3.7.).

XXIII.4.3.3. Definition. Let G be a topological group. Then G has *no small subgroups*, or is an *NSS group*, if there is a neighborhood of the identity which contains no nontrivial subgroup of G.

An NSS group is Hausdorff, since $\overline{\{e\}}$ is contained in every neighborhood of e (Exercise XXIII.1.7.2.). Any Lie group is an NSS group (). By XXIII.1.4.11., a locally compact NSS group is metrizable.

XXIII.4.3.4. Hilbert asked in his 1900 problem list (#5) whether every locally Euclidean group is a Lie group (i.e. has a Lie group structure; Hilbert did not state the problem quite so cleanly). This proved to be a difficult problem and was solved in 1952 by A. Gleason, D. Montgomery, and L. Zippin ([Gle52], [MZ52]); special cases had been established by von Neumann (compact groups [vN33]), Pontryagin (abelian groups []), and Chevalley (solvable groups [Che41]). The proof of the general theorem was simplified by H. Yamabe ([Yam53a], [Yam53b]), although it is still deep and difficult. It is one of the remarkable results of mathematics, often described as "getting something for nothing."

[2] [Mor93, 1752].

XXIII.4.3.5. THEOREM. Let G be a locally compact group. The following are equivalent:

(i) G is locally Euclidean.

(ii) G is an NSS group.

(iii) G has a unique structure as a Lie group.

The proof is far beyond the scope of this book. See e.g. [MZ74] or [Kap71].

"[This result] shows that you can't have a little bit of orderliness without a lot of orderliness."

A. Gleason[3]

The implication (ii) ⇒ (i) of XXIII.4.3.5. has a generalization due to GLEASON and YAMABE:

XXIII.4.3.6. THEOREM. Let G be a locally compact group. Then G has an open subgroup H such that, if U is any neighborhood of e in G, there is a compact normal subgroup K of H, contained in U, such that H/K is locally Euclidean (and thus has a unique structure as a Lie group).

The K cannot be chosen normal in G in general (). But since H/K is a clopen subset of the homogeneous space G/K, it follows that G/K is locally Euclidean as a topological space, hence a manifold.

[3] [Yan02, p. 144].

XXIII.5. Locally Compact Abelian Groups and Pontryagin Duality

Locally compact abelian groups have special properties which make them an especially attractive class of objects to study, with important applications. There is a beautiful duality theory which leads to, among other things, a general theory of Fourier transform. Analysis on locally compact abelian groups, including the general Fourier transform, is called *harmonic analysis*.

XXIII.5.1. The Dual Group

To each locally compact abelian group G there is associated a dual group $\hat{G}$ which is also a locally compact abelian group. The major result, the Pontryagin Duality Theorem (), will be that the dual group $\hat{\hat{G}}$ of $\hat{G}$ is topologically isomorphic to G, so there is a true duality between G and $\hat{G}$. In this subsection, we define the dual group and examine some of its properties.

XXIII.5.1.1. Definition. Let G be a locally compact abelian group, and $\mathbb{T}$ the circle group (). A *character* of G is a continuous homomorphism $\chi : G \to \mathbb{T}$. If χ_1, χ_2 are characters, define the *product* $\chi_1\chi_2$ by $(\chi_1\chi_2)(g) = \chi_1(g)\chi_2(g)$ for $g \in G$, where the multiplication is in $\mathbb{T}$ (i.e. the pointwise product as complex-valued functions). Let $\hat{G}$ be the set of all characters of G, with the topology of uniform convergence on compact subsets of G (compact-open topology, ()). $\hat{G}$ is called the *dual group*, or *Pontryagin dual group*, of G.

The terminology is justified by the following result:

XXIII.5.1.2. Theorem. Let G be a locally compact abelian group. Then $\hat{G}$ is a locally compact abelian group under the product and topology of XXIII.5.1.1..

Proof: First note that $\hat{G}$ is an abelian group under the product: the product is obviously associative and commutative, the constant function 1 is an identity, and if $\chi \in \hat{G}$, then $\chi^{-1} \in \hat{G}$ defined by $\chi^{-1}(g) = [\chi(g)]^{-1} = \overline{\chi(g)}$ is an inverse for χ.

To show $\hat{G}$ is a topological group,

It remains to show that $\hat{G}$ is locally compact. 👍

XXIII.5.1.3. Examples. (i) Let $G = \mathbb{Z}$ with the discrete topology. Any homomorphism χ from $\mathbb{Z}$ to $\mathbb{T}$ is continuous, and completely determined by $\chi(1)$, which can be any element of $\mathbb{T}$; thus $\hat{\mathbb{Z}}$ can be identified as a set with $\mathbb{T}$. The product of two characters is just the product in $\mathbb{T}$ under this identification, so $\hat{\mathbb{Z}} \cong \mathbb{T}$ as a group. The topology on $\hat{\mathbb{Z}}$ is the topology of pointwise convergence since $\mathbb{Z}$ has the discrete topology. If (χ_j) is a net in $\hat{\mathbb{Z}}$, then $\chi_j \to \chi$ pointwise if and only if $\chi_j(1) \to \chi(1)$. Thus the topology on $\hat{Z}$ is just the usual topology on $\mathbb{T}$ under the identification, i.e. the identification $\chi \mapsto \chi(1)$ is a topological isomorphism from $\hat{\mathbb{Z}}$ onto $\mathbb{T}$.

(ii) Let $G = \mathbb{R}$ with the usual topology. If χ is a continuous homomorphism from $\mathbb{R}$ to $\mathbb{T}$, then χ lifts to a continuous function $\tilde{\chi} : \mathbb{R} \to \mathbb{R}$ (), and it is easily checked that $\tilde{\chi}$ is a homomorphism. But the only continuous homomorphisms from $\mathbb{R}$ to $\mathbb{R}$ are multiplication by a constant: if $\phi : \mathbb{R} \to \mathbb{R}$ is a continuous homomorphism,

then ϕ is uniquely determined on $\mathbb{Q}$ by $\phi(1) = \lambda$, and by continuity we must have $\phi(x) = \lambda x$ for all $x \in \mathbb{R}$. Thus, if $\tilde{\chi}(x) = \lambda x$ for fixed $\lambda \in \mathbb{R}$, we must have $\chi(x) = e^{2\pi i \lambda x}$, and we can identify χ with λ, i.e. call this character χ_λ. The map $\lambda \mapsto \chi_\lambda$ is a group-isomorphism from $\mathbb{R}$ onto $\hat{\mathbb{R}}$.

To show that this is a topological isomorphism, suppose $\chi_{\lambda_j} \to \chi_\lambda$ uniformly on compact sets. Then in particular

(iii) Let $G = \mathbb{T}$. Suppose χ is a continuous homomorphism from $\mathbb{T}$ to $\mathbb{T}$. Compose χ with the homomorphism $\phi : \mathbb{R} \to \mathbb{T}$ with $\phi(x) = e^{2\pi i x}$ to get a character $\bar{\chi}$ of $\mathbb{R}$, which must be of the form $\bar{\chi}(x) = e^{2\pi i \lambda x}$ for some $\lambda \in \mathbb{R}$, so we have $\chi(e^{2\pi i x}) = e^{2\pi i \lambda x}$. This function is well defined only if $\lambda \in \mathbb{Z}$. So a continuous homomorphism from $\mathbb{T}$ to $\mathbb{T}$ can only be of the form $\chi_n(\xi) = \xi^n$ for some fixed $n \in \mathbb{Z}$ (), so $\hat{\mathbb{T}} \cong \mathbb{Z}$ as groups. Since $\mathbb{T}$ is compact, the topology on $\hat{\mathbb{T}}$ is the topology of uniform convergence. But no χ_n for $n \neq 0$ can be uniformly close to the constant function 1, so the topology on $\hat{\mathbb{T}}$ is the discrete topology, i.e. $\hat{\mathbb{T}} \cong \mathbb{Z}$ with the discrete topology.

(iv) Let $G = \mathbb{Z}_n$ with the discrete topology. Then any $\chi \in \hat{G}$ is again completely determined by $\chi(1)$. But since 1 has order n in G, $\chi(1)$ must have order n in $\mathbb{T}$, i.e. $\chi(1)$ must be an n'th root of unity. Any n'th root of unity defines a character, and the characters multiply like the roots of unity, so we have $\hat{\mathbb{Z}}_n \cong \mathbb{Z}_n$, and the topology on $\hat{\mathbb{Z}}_n$ is obviously also the discrete topology. But there is no *natural* isomorphism between $\hat{\mathbb{Z}}_n$ and $\mathbb{Z}_n$: any isomorphism requires making a choice of generators for the groups.

XXIII.5.1.4. The map $G \mapsto \hat{G}$ is contravariantly functorial: if H is a locally compact abelian group and $\phi : G \to H$ is a continuous homomorphism, then there is an induced map $\hat{\phi} : \hat{H} \to \hat{G}$ defined by $\hat{\phi}(\chi) = \chi \circ \phi$. It is easily checked that $\hat{\phi}$ is also a continuous homomorphism, i.e. $G \mapsto \hat{G}$ is a contravariant functor from the category of locally compact abelian groups and continuous homomorphisms to itself.

XXIII.5.1.5. The characters of a locally compact abelian group G are just the one-dimensional unitary representations () of G. Every one-dimensional representation is obviously irreducible, and every irreducible unitary representation of an abelian group is one-dimensional (), so $\hat{G}$ is just the set of irreducible representations of G. Two one-dimensional representations are equivalent () if and only if they are identical as homomorphisms to $\mathbb{T}$, so $\hat{G}$ is just the set of equivalence classes of irreducible representations of G, i.e. the notation is consistent with the notation of (). But if G is abelian, $\hat{G}$ has the additional natural structure of a (topological) group; $\hat{G}$ for nonabelian G has no such structure.

XXIII.5.1.6. By (), every unitary representation of G defines representations of $L^1(G)$ and $C^*(G)$, and conversely. Thus, if G is locally compact abelian, each character of G defines a (bounded) *-homomorphism from $L^1(G)$ or $C^*(G)$ to $\mathbb{C}$, and these are precisely all the irreducible representations of these group algebras. Thus $\hat{G} \cong \widehat{L^1(G)} = \widehat{C^*(G)}$. The topology on $\hat{G}$ is precisely the hull-kernel topology on $\widehat{C^*(G)} \cong \mathrm{Prim}(C^*(G))$.

Since irreducible representations are separating (), we obtain:

XXIII.5.1.7. Proposition. Let G be a locally compact abelian group, and $g \in G$, $g \neq e$. Then there is a character χ of G with $\chi(g) \neq 1$.

XXIII.5.1.8. DEFINITION. If G is a locally compact abelian group and H is a closed subgroup, the *annihilator* of H in $\hat{G}$ is

$$H^\perp = \{\chi \in \hat{G} : \chi(h) = 1 \text{ for all } h \in H\} .$$

An element of $H^\perp$ defines a character of G/H in the obvious way. Conversely, every character of G/H can be regarded as a character of G whose kernel contains H, i.e. an element of $H^\perp$.

XXIII.5.1.9. PROPOSITION. Let G be a locally compact abelian group and H a closed subgroup. Then the above correspondence is a topological isomorphism between $H^\perp$ and $\widehat{G/H}$.

Pontryagin Duality

XXIII.5.1.10. Let G be a locally compact abelian group, with dual group $\hat{G}$. There is a natural homomorphism $\rho : G \to \hat{\hat{G}}$ given by $[\rho(g)](\chi) = \chi(g)$. This map is trivially injective, and almost trivially continuous.

The main result is:

XXIII.5.1.11. THEOREM. [PONTRYAGIN DUALITY] Let G be a locally compact abelian group, with dual group $\hat{G}$. The map ρ is a topological isomorphism from G onto $\hat{\hat{G}}$. Pontryagin dual gives a contravariant category isomorphism of period two on the category of locally compact abelian groups and continuous homomorphisms.

XXIII.5.1.12. In particular, we have a Galois-type correspondence between closed subgroups of G and closed subgroups of $\hat{G}$, with a subgroup of G corresponding to a quotient group of $\hat{G}$ and vice versa: the correspondence is $H \subseteq G \leftrightarrow \hat{H} = \hat{G}/(H^\perp)$.

Finer Structure

XXIII.5.1.13. THEOREM. Let G be a locally compact abelian group. Then

(i) $\hat{G}$ is second countable if and only if G is second countable.

(ii) $\hat{G}$ is compact [resp. discrete] if and only if G is discrete [resp. compact].

If G is discrete (so $\hat{G}$ is compact), there are further topological properties of $\hat{G}$ which reflect algebraic properties of G:

XXIII.5.1.14. PROPOSITION. Let G be a discrete abelian group, and $\hat{G}$ the compact dual group. Then

(i) $\hat{G}$ is connected if and only if G is torsion-free.

(ii) $\hat{G}$ is totally disconnected (zero-dimensional) if and only if G is a torsion group.

XXIII.5.1.15. This result does not hold in general if G is not discrete; there are many counterexamples. The best counterexample is perhaps the additive group $\mathbb{Q}_p$ of p-adic numbers (): it is locally compact, self-dual, torsion-free, and totally disconnected. The only implications which hold for general locally compact G are ($\hat{G}$ connected) $\Rightarrow$ (G torsion-free) [if H is a nontrivial finite subgroup of G, then $H \cong \hat{H} \cong \hat{G}/H^{\perp}$ so $\hat{G}$ has a nontrivial discrete quotient] and (G torsion) $\Rightarrow$ ($\hat{G}$ totally disconnected) [if $\hat{G}$ is not totally disconnected, it has a nontrivial closed connected subgroup H; thus $\hat{H} \cong G/H^{\perp}$ is torsion-free, so G has a nontrivial torsion-free quotient].

XXIII.5.2. Exercises

XXIII.5.2.1. (a) Let G be an abelian group with the discrete topology. Show that the topological dimension (covering dimension *dim*) of the compact group $\hat{G}$ equals the rank () of G.

(b) A torsion-free abelian group of finite rank is countable. Hence a connected compact abelian group of finite topological dimension is metrizable.

(c) If H is a connected compact abelian group of topological dimension 1, then H is either a circle or a solenoid (). [Apply (a) to $\hat{H}$.]

(d) The rank 2 torsion-free abelian groups, i.e. the subgroups of $\mathbb{Q}^2$ containing $\mathbb{Z}^2$, are "unclassifiable" (in a technically precise sense, cf. [Tho03]). Thus the connected compact abelian groups of topological dimension 2 (all of which are metrizable) are "unclassifiable" in the same sense. In particular, they can be far more complicated than products of two circles and/or solenoids. The classification problem for rank n becomes strictly more complex as n increases.

XXIII.5.2.2. **Almost Periodic Functions on $\mathbb{R}$.** Let $C_b(\mathbb{R})$ be the C*-algebra of bounded complex-valued continuous functions on $\mathbb{R}$, with supremum norm (topology of uniform convergence). A function $f : \mathbb{R} \to \mathbb{C}$ is *periodic* if there is a positive number p, called a *period* for f, such that $f(t+p) = f(t)$ for all $t \in \mathbb{R}$.

(a) If f is a continuous periodic function, show that f is necessarily bounded, i.e. in $C_b(\mathbb{R})$, and if f is not constant, there is a smallest period p_0 for f, and all other periods for f are of the form np_0 for some $n \in \mathbb{N}$.

(b) A sum of two periodic functions f and g is periodic if and only if their periods $p_0(f)$ and $p_0(g)$ are *commensurable*, i.e. $\frac{p_0(f)}{p_0(g)} \in \mathbb{Q}$. Thus a linear combination of periodic functions is not necessarily periodic.

(c) Let $AP(\mathbb{R})$ be the closure in $C_b(\mathbb{R})$ of the set of linear combinations of continuous periodic functions. The functions in $AP(\mathbb{R})$ are called *almost periodic functions*. Show that $AP(\mathbb{R})$ is closed under multiplication, and is thus a unital C*-subalgebra of $C_b(\mathbb{R})$.

(d) Another intrinsic description of almost periodic functions can be given. If $\lambda \in \mathbb{R}$, define $T_\lambda : C_b(\mathbb{R}) \to C_b(\mathbb{R})$ by $[T_\lambda f](t) = f(t-\lambda)$. Then T_λ is an isometric linear operator (algebra automorphism) on $C_b(\lambda)$,

and $\lambda \mapsto T_\lambda$ is a continuous homomorphism of the additive group $\mathbb{R}$ into the automorphism group of $C_b(\mathbb{R})$. For $f \in C_b(\mathbb{R})$ set

$$\Lambda(f) = \{T_\lambda f : \lambda \in \mathbb{R}\} \subseteq C_b(\mathbb{R}) \ .$$

The set $\Lambda(f)$ is a continuous image of $\mathbb{R}$. Show that $\Lambda(f)$ is compact (and homeomorphic to $\mathbb{T}$) if and only if f is periodic, and $\Lambda(f)$ is precompact if and only if $f \in AP(\mathbb{R})$.

(e) Let B be the dual space $\widehat{AP(\mathbb{R})} \cong \mathrm{Prim}(AP(\mathbb{R}))$ of $AP(\mathbb{R})$ (), so B is a compact Hausdorff space and $AP(\mathbb{R}) \cong C(B)$. There is a natural map of $\mathbb{R}$ into B: since $AP(\mathbb{R})$ is a unital C*-subalgebra of $C_b(\mathbb{R})$, there is a surjective map from $\widehat{C_b(\mathbb{R})} \cong \beta\mathbb{R}$ () to B, and $\mathbb{R}$ embeds naturally as a dense subset of $\beta\mathbb{R}$. Let $\tilde{\mathbb{R}}$ be the image of $\mathbb{R}$ in B. Then $\tilde{\mathbb{R}}$ is dense in B.

(f) The map $\mathbb{R} \to \tilde{\mathbb{R}}$ can be described simply and concretely: if $\lambda \in \mathbb{R}$, define χ_λ by $\chi_\lambda(f) = f(\lambda)$. Then χ_λ is a character on $C_b(\mathbb{R})$, and by restriction on $AP(\mathbb{R})$. In particular, the map $\mathbb{R} \to \tilde{\mathbb{R}}$ is injective, i.e. $\mathbb{R}$ "embeds" in B as a dense subspace. B is called the *Bohr compactification* of $\mathbb{R}$.

(g) The C*-algebra $AP(\mathbb{R})$ is nonseparable: for $\lambda \in \mathbb{R}$, let $\phi_\lambda(t) = e^{i\lambda t}$. Each ϕ_λ is periodic, hence $\phi_\lambda \in AP(\mathbb{R})$ for all $\lambda \in \mathbb{R}$, and $\|\phi_\lambda - \phi_\mu\| = 1$ if $\lambda, \nu \in \mathbb{R}$, $\lambda \neq \mu$. Thus B is not metrizable. Linear combinations of the ϕ_λ are uniformly dense in $AP(\mathbb{R})$.

(h) Show that the map from $\mathbb{R}$ to $\tilde{\mathbb{R}} \subseteq B$, although one-to-one and continuous, is not a homeomorphism, i.e. the relative topology on $\tilde{\mathbb{R}}$ is strictly weaker than the ordinary topology on $\mathbb{R}$ (and not first countable or locally compact). However, the two topologies have the same convergent sequences, i.e. if (λ_n) is a sequence in $\mathbb{R}$, then $\lambda_n \to \lambda$ in $\mathbb{R}$ if and only if $\chi_{\lambda_n} \to \chi_\lambda$ in $\tilde{\mathbb{R}}$. (Thus calling $\mathbb{R} \mapsto \tilde{\mathbb{R}}$ an "embedding" must be interpreted properly, and the Bohr compactification is not a compactification of $\mathbb{R}$ in the sense of XIII.11.9.1.; cf. XIII.11.9.2..)

(i) Show that B can be made into a compact abelian topological group. Specifically, B can be identified with $\widehat{\mathbb{R}_d}$, the dual group of the additive group $\mathbb{R}_d$ of real numbers with the discrete topology. This gives another description of the embedding of $\mathbb{R}$ into B: the dual of the identity map from $\mathbb{R}_d$ to $\mathbb{R}$, a continuous homomorphism, embeds $\tilde{\mathbb{R}} \cong \mathbb{R}$ into $\widehat{\mathbb{R}_d} \cong B$, i.e. $\tilde{\mathbb{R}}$ is the set of characters of $\mathbb{R}_d$ which are continuous when $\mathbb{R}$ is given its ordinary topology (in particular, $\tilde{\mathbb{R}}$ is a subgroup of B, and the embedding $\mathbb{R} \to \tilde{\mathbb{R}}$ is a group-homomorphism). It also gives another proof that B is not metrizable, since $\mathbb{R}_d$ is an uncountable discrete group. In fact, $\mathbb{R}_d \cong \mathbb{Q}^{\mathfrak{c}}$, a direct sum of $\mathfrak{c} = 2^{\aleph_0}$ copies of the additive group $\mathbb{Q}$, so B is topologically isomorphic to the product of $\mathfrak{c}$ copies of the solenoid $\hat{\mathbb{Q}}$. The copy $\tilde{\mathbb{R}}$ of $\mathbb{R}$ in B "winds around" densely in B. Since $\tilde{\mathbb{R}}$ is connected (or since $\mathbb{R}_d$ is torsion-free), B is connected (but far from path-connected).

(j) There is an interesting inner product which can be put on $AP(\mathbb{R})$. Show that

$$\langle f, g \rangle = \lim_{b \to +\infty} \left[\frac{1}{b} \int_0^b f(t)\overline{g(t)}\, dt \right] = \lim_{b \to +\infty} \left[\frac{1}{2b} \int_{-b}^b f(t)\overline{g(t)}\, dt \right]$$

for $f, g \in AP(\mathbb{R})$ gives a well-defined inner product. The set $\{\phi_\lambda : \lambda \in \mathbb{R}\}$ from (g) is an uncountable orthonormal set for this inner product. The inner product gives a strictly weaker topology than the uniform topology on $AP(\mathbb{R})$ from the supremum norm, and is not complete: for example, any periodic Lebesgue measurable and locally square-integrable function on $\mathbb{R}$ is in the completion. The completion can be identified with $L^2(B, \mu)$, where μ is Haar measure on B.

(k) The same construction can be done for any locally compact but noncompact abelian group G (e.g. any discrete abelian group): let $\hat{G}_d$ be the dual group $\hat{G}$ with the discrete topology, and then G embeds as a dense subgroup of the compact abelian group $B_G = \widehat{\hat{G}_d}$. In particular, the Bohr compactification of $\mathbb{R}^n$ can be formed this way. Since $\hat{\mathbb{R}}_d^n \cong \mathbb{R}_d^n$ is isomorphic to $\mathbb{R}_d$ (both are isomorphic to $\mathbb{Q}^{\mathfrak{c}}$), $B_{\mathbb{R}^n}$ is topologically isomorphic to $B_{\mathbb{R}} = B$, although not in a way that respects the embeddings of $\mathbb{R}^n$ and $\mathbb{R}$.

Chapter XXIV

Addenda

XXIV.1. Mathematical Names

One problem not often discussed, which every student and mathematician faces, is learning how to pronounce names of mathematicians when citing or describing their work. I know I have struggled with this on occasion, beginning with my student days, and judging from mispronunciations I have heard from other mathematicians I am hardly alone. Thus I have included this guide as a public service. The information here is correct to the best of my knowledge, but I make no guarantees, and I welcome corrections. I have generally omitted names whose pronunciation should be obvious to English speakers. There are a number of websites such as http://pronouncemath.blogspot.com which may be helpful, but none should be regarded as authoritative.

Since neither I nor most potential readers are professional linguists, I have written the pronunciations using English syllables instead of linguistic symbols. The pronunciations are only approximations, and reflect commonly accepted pronunciations by English-speaking mathematicians rather than exact pronunciations in the original language. Of course, languages vary significantly, particularly in the pronunciation of vowels and r's, and also l's, n's, and other consonants; stress patterns also vary. It is probably better for a mathematician speaking in English to use an English pronunciation instead of an exact original pronunciation anyway.

The English syllables and character combinations are slightly ambiguous in places, but should be pronounced in the usual English way; thus "ch" is always pronounced as in "chair". One sound not present in English is the German *ch* or Russian x, pronounced like an English "h" but farther back in the throat. This sound is written "kh".

The stressed syllable is written in capital letters.

Spellings of names not originally written in the Latin alphabet are also an issue. This arises most critically with Russian names, but can also occur with Arabic, Chinese, or other names. (It is potentially an issue with Indian and Japanese names, but these seem to have standard and universally-used transliterations into the Latin alphabet.)

Russian names are not always transliterated consistently, and the standard English, French, and German transliteration schemes are rather different; and since much mathematical literature has historically been written in German or French, French- and German-style transliterations have come into common use even in English texts. Thus, for example, the name of the well-known Russian mathematician Тихонов, who originally published his papers in German, is frequently transliterated TYCHONOFF instead of the proper English transliteration TIKHONOV (this difference is exacerbated by the fact that a final consonant like в in Russian is normally pronounced unvoiced). An initial *E* in Russian is usually pronounced "yeh", so should be properly transliterated *Ye* in this case (the name of the former Russian president Ельцин is transliterated YELTSIN), but Егоров is usually transliterated EGOROV or EGOROFF, not YEGOROV. We have given common spelling variations in the list. Another difficulty is that some Russian mathematicians, especially Jewish ones, have names which were originally German which were transliterated into Russian; there is a question whether these should be retransliterated into English or restored to their original German spellings. For example, Бернштейн (Sergei) is usually written BERNSTEIN, and Урысон is usually URYSOHN. Наймарк is sometimes spelled NAIMARK and sometimes NEUMARK. Yet another gray area concerns Russian mathematicians who emigrate and begin spelling their own names a certain way in the Latin alphabet. The prime example for this book is BESICOVITCH (his spelling). We will regard this example as a voluntary name change and use his spelling.

One last note of caution, more relevant for contemporary names than historical ones: many Americans (and Canadians) are of European or Asian ancestry, and their names are not always pronounced or spelled in the ancestral way (this is even true of BLACKADAR, which was originally BLACKADDER, although not of BBC fame!) There are particularly many German-Americans, Italian-Americans, Polish-Americans, and Chinese-Americans, and Jewish Americans whose names are of German, Polish, or Russian origin. (There are also a

great many Americans with Hispanic names, but unfortunately this group is seriously underrepresented in the world of mathematics.)

Here are some common general pronunciation conventions for some common languages:

German: *a* is normally pronounced "ah", never "eh" unless it has an umlaut (*ä*). Other vowels which can have umlauts are *ö* and *ü*, which are hard to pronounce for an English speaker; *ö* sounds something like "er", and *ü* is something like a deep "ee" with rounded lips. The vowels *ä, ö, ü* are sometimes written *ae, oe, ue* respectively.

Double vowels are just pronounced as longer versions of the corresponding single vowel: *aa* is pronounced "ah", *ee* is pronounced roughly like the name of the English letter *a*, and *oo* is pronounced "oh", not English "oo". (The combination *eh* is pronounced the same as *ee*, leading to pun jokes among German students concerning the word root *lehr*, "teach", and *leer*, "empty". *ah* and *oh* are similarly pronounced like *aa* and *oo* respectively.) A final *e* is always pronounced, normally as "uh" and not "ee".

Diphthongs are pronounced as follows: *au* as "ow", not "aw"; *eu* as "oy", not "oo" or "you"; *ei* as the name of the English letter *i*; *ie* as "ee".

The letter *v* is usually pronounced "f"; the letter *w* is usually pronounced roughly like English *v*, but a little softer with rounded lips. *z* is pronounced "ts". The combination *th* is pronounced "t" (the sound "th", voiced or unvoiced, does not occur in German). These letters also sometimes occur consecutively at the break between parts of a compound word, in which case they are pronounced separately; BEETHOVEN is an example (although this name is of Flemish origin). *st* is often, but not always, pronounced "sht". *ch* is pronounced "kh". In *sh*, which occurs rarely, the letters are pronounced separately; the sound "sh" is spelled *sch*. Thus *Monatsheft* is pronounced "MOE-nahts-heft", not "MOE-naht-sheft". The combination *sh* usually only occurs, as here, in compound words where the first part ends in *s* and the second part begins with *h*. This applies to names such as PRINGSHEIM.

Russian: *a* (а) is pronounced "ah". *e* (е) in initial position or following another vowel is pronounced "yeh"; after a consonant, some trace of the "y" usually remains, but it mainly disappears into the preceding consonant (which is technically pronounced soft). There is another letter *e* (э) which is pronounced without the "y", but this letter mostly appears in foreign words and rarely if ever in Russian names. Sometimes in a stressed position е is pronounced "yaw" (е pronounced this way should be written ё, but the "umlaut" is typically omitted). For example, the names of the former Soviet leaders Хрущев and Горбачев are properly pronounced "khrooshch-OFF" and "gar-ba-CHOFF" respectively; the most notable instance among mathematical names is in Чебышев. When this occurs in the list, it is noted in the pronunciation guide. The letter *o* (о) is pronounced "aw", or "uh" (sometimes "ah") in an unstressed position. The letter *u* (у) is pronounced "oo", not "uh"; the soft form ю is pronounced "you". The letter *i* (и) is pronounced "ee". There is another vowel ы, the hard version of и, transliterated *y*, which is pronounced similarly to the German *ü* but farther back in the throat. The vowel я, "ya", is the soft form of а.

There are no diphthongs in Russian: consecutive vowels are pronounced separately. Thus, for example, Воеводский is pronounced "vaw-yeh-VAWD-skee", and the chemist Менделеев is pronounced "men-del-YAY-yeff".

Most Russian consonants have a hard and soft form. The hard form is the "default", and consonants become soft only when followed by a soft vowel (я, е, и, ю) or a soft sign (ь), which is not pronounced. The soft sign is sometimes transliterated with an apostrophe (as in OL'SHANSKII), but we have avoided using this. There is also a "hard sign" or "separative sign" ъ, which occurs only rarely in modern Russian and which now usually has the same pronunciation effect as ь. The letter й, which occurs only after vowels, is pronounced with a slight "y" sound and usually transliterated *i*.

As mentioned earlier, some consonants at the end of words, or before other unvoiced consonants, are pronounced unvoiced.

Polish: Polish names are frequently mispronounced badly in English, and some have acquired widely accepted English pronunciations quite different from the Polish ones. This is mainly due to the fact that the Polish spelling conventions for many sounds are different than the English ones. As a result, some Polish names look "unpronounceable" to English speakers not familiar with the Polish conventions. But a little familiarity makes them relatively easy to pronounce, at least approximately.

In Polish, *w* is pronounced "v"; *ł* is pronounced "w", *c* is pronounced "ts" (except when followed by *i* or *z*), and *j* is pronounced "y".

The combination *sz* is pronounced "sh", *cz* is pronounced "ch", and *rz* is pronounced "zh". These are hard sounds. The accented letters *ś*, *ć*, *ź* (*si*, *ci*, *zi* before vowels) are soft forms of *sz*, *cz*, and *rz* respectively; *ż* is pronounced like *rz*.

The vowel *ó* is pronounced "oo", the same as *u*. There are also nasal vowels and a nasal *n* (*ń*), but these are perhaps too arcane to worry about here.

For example, the common first name *Jerzy* is pronounced "YEH-zhee", not "Jersey". (The Czech version of this name, *Jiři*, is a name I admit I cannot pronounce correctly, although I know how.)

In Polish words, including names, the stress is almost always on the next-to-last syllable.

Some Poles (mainly Jews, but not exclusively) have names which are originally German but which are written with Polish spelling conventions. Notable examples are the names *Szulc* and *Sznajder*, which are the same (and pronounced the same) as the German *Schulz* and *Schneider* respectively. Among mathematicians, the names Aronszajn and Szpilrajn come to mind.

Hungarian: Hungarian names also have the reputation of being hard for English speakers to pronounce, and the reason is again differences in spelling conventions. The Hungarian spelling conventions are in some instances the direct opposites of the Polish ones.

The vowels *a*, *e*, *i*, *o*, *ö*, *u*, *ü* are pronounced similarly to the corresponding German vowels. An accent denotes a long vowel; *ő* and *ű* are the long forms of *ö* and *ü* respectively.

Hungarian consonants and combinations are as follows: *c* is pronounced "ts", *cs* is "ch", *gy* is a soft "j" (actually similar to the British pronunciation of "due"), *j* is like "y"; *ly*, *ny*, *ty* are soft versions of *l*, *n*, *t*; *s* is pronounced "sh" and *sz* is pronounced "s"; and *zs* is "zh" (hence *dzs* is a hard "j").

The stress in Hungarian words is on the first syllable. Note that an accent on a vowel does *not* indicate stress.

French: French names truly are hard for many English speakers to pronounce correctly, although most English speakers know how to make a decent approximation since they are relatively familiar.

Italian:

Japanese: The syllables of Japanese words are all pronounced with roughly the same stress. However, some syllables are partially swallowed. For example, the name of the city *Fukushima*, the site of the recent nuclear disasters, is pronounced something like "FKOO-shi-muh" (at least it sounds like this to me).

Chinese: I will not attempt to describe the proper pronunciation of Chinese names, since it is far beyond my expertise. However, I will recall a conversation I had with a Chinese mathematician, about the Chinese name *Ng* (which is usually pronounced "ing" in English, which is at least better than "nig".) I asked the proper pronunciation, and he replied, "Mm?" I thought he had not understood what I said, so I repeated the question, and he again replied, "Mm?"

Of course, there are many other languages, each with its own peculiarities. For example, consider the name of the eminent, and recently tragically deceased, Danish mathematician Uffe Haagerup. Danes pronounce

his last name something like “OR-up”, but somewhat less distinctly. English-speaking mathematicians usually say “HOG-er-up”, which Uffe (“OOF-uh”) didn’t seem to mind too much.

Abel	AH-bull	*abelian* pronounced “uh-BEEL-yan”
Alaoglu	AL-uh-loo	Approximate; “g” very soft (Turkish)
Aleksandrov	ah-leks-AHN-druff	
Alexandrov	See Aleksandrov	
Aronszajn	AR-un-shine	
Arzelà	ahr-ZAY-la	
Ascoli	as-COAL-ee	s is unvoiced
Baire	BEAR	
Banach	BAHN-akh	Often pronounced BAHN-ock
Bernoulli	bear-NEW-lee	
Bernstein	BEARN-shtine	
Blackadar	BLACK-uh-dahr	
Bolzano	bol-TSAHN-o	Italian name, but Bohemian
Borel	burr-ELL	
Bourbaki	bour-bah-KEE	
Bożejko	bo-ZHAY-ko	
Brouwer	BROW-er	
Carathéodory	car-uh-thay-uh-DOOR-ee	“th” is unvoiced
Cartan	car-TAHN	n is nasal
Cauchy	KOH-shee	
Cavalieri	cav-ul-YAIR-ee	
Čech	CHEKH	As in “Czech”
Cesàro	chuh-SAH-row	Italian
Chebyshev	cheb-uh-SHOFF	often cheb-uh-SHEFF
d’Alembert	dahl-em-BEAR	
Daniell	dan-YELL	
Darboux	dahr-BOO	
Dedekind	DAY-duh-kind	“kind” as in “kin”
DeMoivre	duh-MWAVR	
Denjoy	don-ZHWAH	n is nasal
Desargues	days-ARG	
Descartes	day-CART	
Dirac	durr-ACK	
Dirichlet	DEER-ikh-lay	often DEER-ish-lay. See note
Egorov	ye-GOR-uff	Not EGG-o-roff!
Erdős	AIR-dosh	AIR-dush is more correct
Euclid	YOU-klid	
Euler	OY-ler	
Fatou	fat-TOO	
Fejér	FAY-er	FAY-air is closer to correct
Fermat	fair-MAH	
Fibonacci	fee-bun-AH-chee	Often fib-un-AH-chee
Fourier	FOUR-yay	
Fraenkel	FRENK-ul	
Fréchet	fresh-A	A as the name of the letter a
Frege	FRAY-guh	
Frobenius	fro-BEAN-ee-us	
Fubini	foo-BEE-nee	

Galois	gal-WAH	
Gauss	GOWSS	
Goursat	gor-SAH	
Haar	HAHR	
Hadamard	ad-uh-MAHR	
Hausdorff	HOUSE-dorf	
Heine	HIGH-nuh	
Hermite	air-MEET	*Hermitian* pronounced her-MEE-shun
Hilbert	HILL-burt	
Hölder	HEL-dur	Approximate; sometimes HULL-der
Jacobi	ya-KOE-bee	Actually German, of Jewish ancestry Sometimes ja-KOE-bee; *Jacobian* pronounced ja-KOE-bee-un
Jordan	zhor-DAHN	N is nasal
Khinchin	KHEEN-cheen	Often KHIN-chin
Kolmogorov	kul-muh-GOR-uff	
Kovalevskaya	ko-va-LYEV-skah-ya	
Krein	KRINE	
Kronecker	KRON-eck-ur	
Kuratowski	kur-a-TOV-skee	
Lagrange	la-GRAHNGE	Not la-GREHNGE!
Laplace	la-PLAHS	Final "s" unvoiced
Laurent	lor-ON	N is nasal
Lebesgue	le-BEG	Second syllable is long.
Legendre	le-ZHAHN-der	
Leibniz	LIBE-nits	
Levi-Civita	LEH-vee-CHIV-ee-tah	
Lévy	LAY-vee	
l'Hôpital	low-pee-TAHL	Often pronounced LOW-pit-ahl
l'Hospital	See l'Hôpital	
Lie	LEE	
Lindelöf	LIN-del-off	Finnish name
Lindemann	LIN-duh-mahn	
Liouville	LEE-oo-vill	
Lipschitz	LIP-shits	No delicate way to write it!
Łoś	WAUSH	
Lusin	See Luzin	French transliteration?
Luzin	LOO-zeen	Often LOO-zin
Marczewski	mar-CHEV-skee	Same person as Szpilrajn
Markov	MARK-uff	
Milman	MIL-mun	
Minkowski	min-KOV-skee	
Mycielski	mi-CHEL-skee	
Nagy	See Sz.-Nagy	
Neumann	NOY-mahn	
Nevanlinna	NEV-ann-lin-uh	
Nikodým	NICK-uh-deem	Polish mathematician, Czech name
Noether	NER-ter	Approximate; first "r" not really there
Ostrogradskii	ah-strah-GRAHD-skee	

Pascal	pass-KAL	
Peano	pay-AH-no	not “Piano”!
Perron	pair-RONE	See note
Picard	pee-CAR	
Plancherel	plahn-sher-EL	n is nasal
Poincaré	pwan-car-AY	
Poisson	pwah-SAWN	N is nasal
Pontryagin	pon-tree-AH-geen	g is hard (actually three syllables pon-TRYAH-geen)
Pringsheim	PRINGZ-hime	“hime” as in “time”
Prüfer	PREE-fer	approximate; often mispronounced PROOF-er
Pythagoras	pith-AG-or-us	
Raabe	RAH-buh	
Radon	RA-don	Not RAY-don
Ramanujan	ra-MAHN-oo-jun	
Riemann	REE-mahn	
Riesz	REESE	
Rolle	RAHL	
Schwartz (J., L.)	SHWARTS	
Schwarz (Hermann)	SHVARTS	Usually pronounced SHWARTS
Shilov	SHEE-luff	
Sierpiński	share-PEEN-skee	n is technically nasal
Sinai	SEEN-eye	
Smulian	SHMOOL-ee-un	S should be written Š
Sobolev	SO-buh-leff	SO-buh-lyeff is more accurate
Souslin	See Suslin	French transliteration
Stieltjes	STEELT-yes	Dutch
Suslin	SOOS-leen	Often SOOS-lin; not SUSS-lin!
Sylow	SEE-low	Norwegian, name apparently unrelated to *Shilov*
Sz.-Nagy	NAHDZH	DZH is very soft
Szegő	SAY-go	SAY-ger is closer to correct
Szpilrajn	SHPEEL-rine	Same person as Marczewski
Tchebysheff	See Chebyshev	
Tietze	TEET-suh	not TEET-see or TEET-zee!
Tikhonov	TEEKH-uh-noff	Often pronounced TICK-uh-noff
Toeplitz	TOE-plits	Correct pronunciation closer to TER-plits
Tonelli	ton-EL-lee	
Tychonoff	See Tikhonov	
Ulam	OO-lahm	
Urysohn	oor-uh-SAWN	not YOUR-i-sun
van Rooij	van ROY	
Vitali	vee-TAHL-ee	
Volterra	voal-TAIR-uh	
von Neumann	fun NOY-mun	
Weierstrass	VYE-er-shtrahss	

Weil (André)	VAY	As in "way" (see note)
Weyl (Hermann)	VILE	No reflection on character!
Wiener	VEE-nur	
Wroński	VRON-skee	often RON-skee; *Wronskian* pronounced RON-skee-un
Yegorov	See Egorov	
Zermelo	tser-MEL-o	See note

Notes:

Dirichlet These pronunciations are almost universally used, but there is some doubt about their correctness. Dirichlet was German, but his family originally came from the town of Richelette in the French-speaking part of Belgium. His full surname was *Lejeune Dirichlet*, originally *le Jeune de Richelette*. Even with a French pronunciation the *t* would be pronounced in this name. The logical pronunciation of *Dirichlet* should be something like deer-ish-LET. I have been unable to determine how he himself pronounced his name, although sources suggest he pronounced it with silent *t*.

Perron Perron was German, but the origin of his surname is obscure.

Weil/Weyl André Weil is sometimes called "der andre Weyl" (thanks to P. Muhly for this information!)

Zermelo The origin of this name is obscure. Zermelo, who was German (and not Jewish), told people who asked that it was a contraction of *Walzermelodie* (cf. [Rei70, p. 98]). More serious speculations on an East Prussian origin of the name, apparently also originating with Zermelo himself, can be found in [Ebb07, p. 1].

I must say I feel fortunate to be a native English speaker, not just because English has become the international language of mathematics (and many other subjects), but also because I am very happy I didn't have to learn the language as an adult. It seems to me like a difficult and irregular language, particularly because of the use of articles (very hard for nonnative speakers to get right) and the tenuous relationship between spelling and pronunciation for many words, along with other irregularities. There are other grammar rules which every native speaker knows but don't know they know, such as:

> "[English adjectives] absolutely have to be in this order: opinion-size-age-shape-colour-origin-material-purpose Noun. So you can have a lovely little old rectangular green French silver whittling knife. But if you mess with that order in the slightest you'll sound like a maniac."
>
> *Mark Forsyth*[1]

and I admire the ability of so many of my foreign colleagues to master the language quite well. Although it is sometimes said that "the international language of mathematics is bad English," I find most of it surprisingly good (although there is a story about a letter of recommendation written by a mathematician who was not a native English speaker, saying that the candidate "exposes himself well in the classroom.")

> "English is weird. It can be understood through tough, thorough thought though."
>
> *Unknown*

We could say the same thing about analysis!

[1] *The Elements of Eloquence: How to Turn the Perfect English Phrase.*

"Why English Is Hard to Learn"

We'll begin with *box*; the plural is *boxes*,
But the plural of ox is *oxen*, not *oxes*.
One fowl is a *goose*, and two are called *geese*,
Yet the plural of *moose* is never called *meese*.

You may find a lone *mouse* or a house full of *mice*;
But the plural of *house* is *houses*, not *hice*.
The plural of *man* is always *men*,
But the plural of *pan* is never *pen*.

If I speak of a *foot*, and you show me two *feet*,
And I give you a *book*, would a pair be a *beek*?
If one is a *tooth* and a whole set are *teeth*,
Why shouldn't two *booths* be called *beeth*?

If the singular's *this* and the plural is *these*,
Should the plural of *kiss* be ever called *keese*?

We speak of a *brother* and also of *brethren*,
But though we say *mother*, we never say *methren*.
Then the masculine pronouns are *he*, *his*, and *him*;
But imagine the feminine . . . *she*, *shis*, and *shim*!



— Anonymous

XXIV.1.1. The Greek Alphabet

One thing every mathematics student must learn is the Greek alphabet, or at least most of it. Greek letters are frequently used as mathematical symbols. Many upper-case Greek letters are essentially identical to Latin letters (although not necessarily the equivalent Latin letter!), and these upper-case Greek letters are rarely if ever used as mathematical symbols. The other upper-case letters are commonly used, but the lower-case letters are more common in mathematics. The only lower-case Greek letter I have never seen used as a mathematical symbol is omicron, although upsilon is rarely used since it is hard to distinguish from nu (or especially from Latin v).

Upper Case	Lower Case	Name	Latin Equivalent
A	α	alpha	a
B	β	beta	b
Γ	γ	gamma	g
Δ	δ	delta	d
E	ϵ	epsilon	e (short)
Z	ζ	zeta	z
H	η	eta	e (long)
Θ	θ	theta	th
I	ι	iota	i
K	κ	kappa	k
Λ	λ	lambda	l
M	μ	mu	m
N	ν	nu	n
Ξ	ξ	xi	ks
O	o	omicron	o (short)
Π	π	pi	p
P	ρ	rho	r
Σ	σ	sigma	s
T	τ	tau	t
Y	υ	upsilon	u
Φ	ϕ or φ	phi	ph (f)
X	χ	chi	kh
Ψ	ψ	psi	ps
Ω	ω	omega	o (long)

The name "omicron" is properly pronounced O-mi-kron, not OM-i-kron, since it means "o micron" ("little o"). The name "omega" is "o mega" ("big o").

The name "phi" is sometimes pronounced FEE and sometimes FY, and there does not seem to be an accepted correct pronunciation. The name "Psi" is SIGH, sometimes PSIGH, and "Xi" is KSEE. Pronunciations of other letter names are fairly evident.

XXIV.2. Origin of Terms and Notation

I have collected here my best information of when various standard mathematical terms were first introduced. Corrections and additions are solicited. Information is from [Caj93], http://jeff560.tripod.com/mathword.html, and http://jeff560.tripod.com/mathsym.html unless otherwise specified.

Terms

Many of these terms were introduced in languages other than English, in appropriate form. We have given the English form or translation.



Figure XXIV.1: A bonesetter in Croatia

algebra: originally *al-jabr*, Arabic for "restoring" (referring to adding or subtracting the same quantity from both sides of an equation), later meaning "bonesetter" (i.e. a restorer of broken bones, cf. Figure XXIV.1), Al-Khwarizmi, 825. Not used in the modern sense until much later.

algebraic topology: Lefschetz, 1936 [Jam99, p. 570]. Previously usually called "combinatorial topology" (not exactly the same thing; Emmy Noether was instrumental in the transition from "combinatorial topology" to algebraic topology in the late 1920s [Jam99, p. 564-565]).

algorithm: from Al-Khwarizmi, whose book (825) introduced "Arabic" numbers (actually originating in India) and modern calculation methods to Europe. "The method of Al-Khwarizmi" referred to solving a problem by calculation.

Banach space: Fréchet, 1928. Banach spaces were defined by Banach in 1920; he called them *espace du type (B)*. It is unclear what the B referred to.

calculus: The term, from the Latin meaning "small pebble," was used to refer to calculation (word of same origin) since Roman times. The modern usage of *calculus* (*differential calculus*) was introduced by LEIBNIZ around 1680. NEWTON first used the term in 1691. The original meaning survives in some other mathematical terms such as *predicate calculus*, *functional calculus*, and *calculus of residues*.

compact (topology): FRÉCHET, 1906 (not quite in the modern sense).

complex number: GAUSS. He used the term to resolve the ambiguity in the use of *imaginary number* to refer both to nonreal complex numbers and to pure imaginary numbers (an ambiguity which persists to a degree even today), and because GAUSS disliked the term "imaginary," which he considered misleading.

continued fraction: WALLIS, *Opera Mathematica I*, 1695 (published at age 79!) [Kli72, p. 255].

corollary: from Latin *corolla*, a small wreath; the English word dates from the 14th century and has been used in the modern mathematical sense (a theorem which is an immediate or easy consequence of another theorem) at least since 1669.

definite integral: LAPLACE, *Memoire sur les suites*, 1779.

derivative: LAGRANGE, *Théorie des fonctions analytiques*, 1797 [Gra10]. LAGRANGE'S term (which he used as early as 1772) was *fonction derivée*. Derivatives were previously described as differentials, fluxions, etc.; the general notion of a function was just appearing in LAGRANGE'S time.

eigenvalue: A hybrid German-English word, apparently first used by A. S. EDDINGTON in 1927. Many mathematicians have rebelled against the term and used alternates, but *eigenvalue* is now firmly established.

eigenvector: Apparently first used by R. BRAUER and H. WEYL, *Spinors in n dimensions*, 1935.

ergodic: coined by BOLTZMANN in 1884.

field (algebra): DEDEKIND used *Zahlenkörper* in 1858. The English *field* was first used by E. H. MOORE in 1893.

functional analysis: PAUL LÉVY, *Leçons de l'analyse fonctionelle*, 1922.

geometry: From Greek "measuring the earth." Used in ancient Greece before PLATO.

harmonic analysis: W. THOMPSON (LORD KELVIN), 1867.

Hilbert space: ERHARD SCHMIDT, *Über die Auflösung linearer Gleichungen mit unendlich vielen Unbekannten*, 1908 (for ℓ^2). According to [Jah03, p. 385], the name "Hilbert space" was first used in 1908 by A. SCHÖNFLIES. Abstract Hilbert spaces were defined by VON NEUMANN in 1928 [vN30].

homeomorphism: POINCARÉ, *Analysis Situs*, 1892 (not quite in the modern sense). The modern usage was standard by 1930.

homology: Used by POINCARÉ in *Analysis Situs*, 1895, but not clearly in the modern sense [Jam99, p. 85]. The word was used in a different sense by JAMES STIRLING in 1749.

homotopy: DEHN and HEEGAARD, *Analysis Situs*, 1907 [Jam99, p. 88]. The concept of homotopy, or continuous deformation, had been around a long time.

imaginary number: DESCARTES, *La Geometrie*, 1637. Such numbers had long been considered.

integral: JAKOB BERNOULLI, *Acta eruditorium*, 1690. JOHANN BERNOULLI disputed this (he disputed many things) and claimed originality for the term. JOHANN BERNOULLI first used the term *integral calculus*, and convinced LEIBNIZ to use it too.

kernel (integral equations): FREDHOLM, *Sur une classe des équations fonctionelles*, 1903.

kernel (algebra): PONTRYAGIN, *Topological Groups*, 1939?.

linear algebra: the term was used by BOMBELLI in about 1550, but the modern usage as the subject of vector spaces and linear transformations only dates from about 1950.

logarithm: JOHN NAPIER, 1614.

manifold: The term was used to denote affine subspaces of $\mathbb{R}^n$ long before the modern definition of manifold was given. First used in something approaching the modern sense by RIEMANN. The "correct" modern definition was nailed down by WHITNEY in the 1930s.

mathematics: from Greek *manthanein*, "to learn." The term *mathematike tekhne*, "mathematical science," already existed in ancient Greece. The word *mathematics* gradually shifted from a plural noun to a singular one (at least in American English; it is somewhat ambiguous in British English, where it is abbreviated "maths," and still plural in some other languages).

matrix: SYLVESTER, 1848. Matrix notation had been recently introduced by CAYLEY. Determinants (of arrays) were introduced by CAUCHY. The older meaning of *matrix* was "the place of origin of something else," and SYLVESTER apparently regarded mathematical matrices as the place of origin of determinants.

parallelepiped: EUCLID. The only controversy is about pronunciation: the accent is most commonly on the next-to-last syllable, although it is sometimes put on the previous syllable (more correct in Greek?) Some authors questionably spell this word *parallelopiped*, evidently with accent on the next-to-last syllable.

pathological: Used in medicine since the 17th century. The first documented use in mathematics is in MURRAY and VON NEUMANN, *On Rings of Operators*, 1936 (referring to non-type-I factors).

pure mathematics: FRANCIS BACON, *The Advancement of Learning*, 1605.

real number: DESCARTES, *La Geometrie*, 1637.

ring: DEDEKIND.

scalar: WILLIAM ROWAN HAMILTON, *On Quaternions*, 1844.

set: Apparently first used (German *Menge*) by BOLZANO, *Wissenschaftslehre*, 1837 [Fra66, p. 2].

topology: J. LISTING, 1847 [Jam99]. The term did not come into general use until the twentieth century; *analysis situs* (introduced by LEIBNIZ, used by POINCARÉ) predominated.

transcendental number: LEIBNIZ, 1704 [Rib00b]. Transcendental numbers were not proved to exist until 1844 (by LIOUVILLE).

trigonometry: BARTHOLOMAEUS PITISCUS, 1595.

vector: Used in astronomy to denote a *radius vector*. First used to denote a point in $\mathbb{R}^3$ by WILLIAM ROWAN HAMILTON in 1844. Only gradually became used to denote a point in a general vector space.

vector space: PEANO, *Formulario mathematico*, 1895.

Notation

a^n (exponent): DESCARTES, 1637 (natural number exponents); NEWTON, 1676 (negative and fractional exponents). Exponents were first used (with different notation) by NICOLE ORESME about 1360.

$|\cdot|$ (absolute value): evidently introduced by WEIERSTRASS.

$\binom{\cdot}{\cdot}$ (binomial coefficient): ANDREAS VON ETTINGSHAUSEN, 1826. EULER used this symbol, but with a fraction bar in the middle.

$\mathbb{C}$ (complex numbers): NATHAN JACOBSON, 1939.

. (decimal point): Unclear. Used by Bartholomaeus Pitiscus in 1612. Popularized by Napier. Decimals had long been used in the Middle East, and ancient Babylonians used a base 60 version. Simon Stevin used decimals, with different notation, in 1585, and, according to [Ste15], the decimal point evolved from Stevin's notation.

$\div$ (division): Johann Rahn, *Teutsche Algebra*, 1659.

δ (limit): Cauchy (for *différence*). Johann Bernoulli used δ for a difference of functions in 1706.

e ($2.718\cdots$): Euler, 1727. The reason for e is obscure, but it probably was not because of his name.

$\emptyset$ (empty set): Bourbaki, 1939. Introduced by Weil from the Norwegian alphabet.

ϵ (limit): Cauchy (for *erreur*).

$=$ (equals): Robert Recorde, *The Whetstone of Witte*, 1557. Developed independently in Bologna about the same time.

$f(x)$ (function): Euler, *Commentarii Academiae Scientarium Petropolitanae*, 1734.

$f : X \to Y$ (function): Hurewicz, 1940.

f' (derivative): Lagrange, *Théorie des fonctions analytiques*, 1797 [Gra10]. The notations $\frac{dy}{dx}$ (Leibniz) and $\dot{x}$ (Newton) previously predominated.

! (factorial): Christian Kramp, *Elements d'arithmétique universelle*, 1808. The alternate notation $\llcorner$, still occasionally seen, is actually newer (Thomas Jarrett, 1827).

$\frac{a}{b}$ (fraction): Fraction bar introduced by Arabs; fractions were used in ancient Greece and by Hindus. Fraction notation first used in Europe by Fibonacci in *Liber Abbaci* [Ste15].

$>$ (greater than): Thomas Harriott, *Artis Analyticae Praxis ad Aequationes Algebraicas Resolvendas*, 1631. Wallis used the first version of $\geq$ in 1670.

i ($\sqrt{-1}$): Euler, 1777. Complex numbers, especially square roots of negative real numbers, had long been considered, although they were still "imaginary."

∞ (infinity): Wallis, *De sectionibus conicis*, 1655.

$\lfloor \cdot \rfloor$ (integer part): Kenneth Iverson, 1962. Gauss defined and used this function and denoted it $[\cdot]$.

$\int$ (integral): Leibniz, 1675. The definite integral notation $\int_a^b$ was first used by Fourier in 1822.

$\cap$ (intersection): Peano, *Formulario mathematico*, 1895.

$<$ (less than): Thomas Harriott, *Artis Analyticae Praxis ad Aequationes Algebraicas Resolvendas*, 1631. Wallis used the first version of $\leq$ in 1670.

$\lim_{x \to a}$ (limit): John Leathem, *Volume and Surface Integrals Used in Physics*, 1905. Weierstrass used $\lim_{x=a}$.

$\in$ (member): Peano, *Formulario mathematico*, 1895.

$-$ (minus): Johannes Widmann, 1489.

$\mathbb{N}$ (natural numbers): Dedekind used N in *Was sind und was sollen die Zahlen*, 1888.

O (big-O): Paul Bachmann, *Analytische Zahlentheorie*, 1894; but see o.

o (little-o): first apparently used in the modern sense by Landau in 1909 (O and o are sometimes called *Landau symbols*). But o was used by Newton (not quite in the same sense, but to represent a negligibly small quantity; cf. [**?**, p. 204] and V.7.).

∂ (partial): Condorcet, 1768. $\frac{\partial u}{\partial x}$ first used by Legendre, 1786. This symbol is an italic Cyrillic d, and this may be the origin of the mathematical symbol (possibly through Euler?)

π ($3.14159\cdots$): William Jones, 1706 (for *perimeter*). Usage popularized by Euler. The number itself has been known and studied since antiquity.

$+$ (plus): Nicole Oresme, 1360.

$\mathbb{Q}$ (rational numbers): Bourbaki introduced Q in the 1930s, evidently for *Quotient*.

$\mathbb{R}$ (real numbers): Dedekind used $\mathbf{R}$ in *Stetigkeit und irrationale Zahlen*, 1872 (he used R for the rational numbers). The real numbers as a named set were rarely referred to before this. Use of "blackboard bold" for $\mathbb{R}$ and other sets of numbers, which is still not universal, is of uncertain origin.

$/$ (division): Thomas Twinin, 1718. May have evolved from fraction notation, which dates much earlier.

$\{\cdot\}$ (set): Zermelo, *Untersuchungen über die Grundlage der Mengenlehre*, 1907.

$\surd$ (radical): Christoff Rudolff, *Die Coss*, 1525. Descartes (1637) added the vinculum. The notation $\sqrt[n]{}$ was suggested by Albert Girard in 1629 and may have been first used by Rolle in *Traité d'Algèbre* in 1690.

$\sum$ (sum): Euler, *Institutiones calculi differentialis*, 1755.

$\times$ (multiplication): William Oughtred, 1631.

$\cup$ (union): Peano, *Formulario mathematico*, 1895.

$\mathbb{Z}$ (integers): Edmund Landau introduced $\bar{\mathfrak{Z}}$ (evidently for *Zahl*) in *Grundlagen der Analysis*, 1930. Bourbaki used Z.

XXIV.3. Miscellany

XXIV.3.1. Age and Mathematics

> "The mathematician lives long and lives young; the wings of the soul do not early drop off, nor do its pores become clogged with the earthly particles blown from the dusty highways of vulgar life."
>
> *J. Sylvester*[2]

> "Very few people do anything creative after the age of thirty-five. The reason is that very few people do anything creative before the age of thirty-five."
>
> *Joel Hildebrand*[3]

Weierstrass is a notable counterexample to the widely held belief that mathematics is exclusively an enterprise for the young. He never earned a doctorate (he was eventually awarded an honorary one), did not have a university position until age 41, and did most of his work on the foundations of analysis over the next twenty years after that. He did significant mathematical research well into his sixties and perhaps early seventies (he published the Weierstrass Approximation Theorem () at age 70), and lectured regularly about his work into his seventies. (He did do important mathematical work in his twenties and thirties while he was employed as a high-school [gymnasium] teacher, but this work was not published or known in the mathematical community until much later.)

Hausdorff did not do much mathematics until he was in his forties (he previously wrote two books on philosophy under the name Paul Mongré).

John Wallis (1616-1703) is perhaps an even more notable example. He continued to produce and write significant mathematics into his eighties, and retained his chair at Oxford until his death at age (almost) 87. (He apparently did not learn or do much serious mathematics until *after* his Oxford appointment at age 33, although he had about as much previous mathematics education as existed in England at the time.) Although he was 27 years older than Newton, he remained active and productive for several years after Newton had ceased to do creative mathematics. When Wallis was 81, Leibniz invited him to Hannover to instruct the Germans on cryptography (he had been England's Royal cryptographer for decades); Wallis declined for strategic reasons, not because he was physically incapable.

Wallis's contemporary Thomas Hobbes (1588-1679), a formidable philosopher but a lightweight mathematician, with whom Wallis had a longstanding polemical debate, lived to age 91. Newton lived to age 84 and remained intellectually sharp almost to the end, but produced little new mathematics after a serious illness at age 50, although he did finally publish part of the work he had done decades earlier.

> "... no old Men (excepting Dr. Wallis) love Mathematicks."
>
> *Isaac Newton*[4]

William Oughtred (1574-1660), one of the most important English mathematicians before Wallis, lived to age 86.

[2] [?, p. 405].
[3] R. Byrne, *The Other 637 Best Things Anybody Ever Said*, 616.
[4] William Whiston, *Memoirs*, p. 315-316; cf. Westfall, *Never at Rest*, p. 139.

Leopold Vietoris died in 2002, two months short of his 111th birthday. He was the oldest person in Austria at the time of his death (and the oldest documented Austrian man ever). He published his last research paper at the age of 103. Sergei Nikolskii (1905-2012) also published research papers until age 101, and Jacques Dixmier (1924-), who is still alive at this writing, recently published a paper at age 100. I am not aware of any other mathematician who published a research work after his/her 100th birthday.

"If you live to a hundred you have it made because very few people die past the age of a hundred."

George Burns[5]

Other notably long-lived mathematicians:

Dirk Struik (1894-2000) 106 years

Henri Cartan (1904-2008) 104 years

Boris Bukreev (1859-1962) 103 years

Notably short-lived mathematicians:

Évariste Galois (1811-1832) 20 years

Jacques Herbrand (1908-1931) 23 years

Mikhail Suslin (1894-1919) 24 years

Pavel Urysohn (1898-1924) 26 years

Niels Henrik Abel (1802-1829) 26 years

Frank Ramsey (1903-1930) 26 years

Rufus Bowen (1947-1978) 31 years

[5] R. Byrne, *The Other 637 Best Things Anybody Ever Said*, 585.

XXIV.3.2. The Mathematical Organ

It was fashionable at one time for phrenologists to make studies of the skulls of various classes of people, including mathematicians. The following is adapted from *The Mathematician as a Type*, by William Story.

It has been claimed that a very large number of mathematicians show a striking physiognomic peculiarity called the "mathematical organ." Franz Joseph Gall, the originator of phrenology, describes it thus:

> "The outer part of the roof of the orbita is depressed in such a manner that the upper edge of the hollow of the eve retains its natural arch only on the inner half, and that the outer half becomes a straight line, which runs sloping downward to its outer end. In consequence of this, the outer part of the upper lid falls and covers the eye, more than usually. Still more decided is the character when the lateral part of the orbita is pushed outward, so that the angular process of the frontal bone (the outer upper angle of the edge of the roof) projects sidewise over the front part of the temple."

Story continues:

> "I know a great part of the living mathematicians personally [and] I have studied the busts, and painted and engraved portraits of many others. In all, without exception, I have found the organ described. Consider the portrait of young Colburn; in him the outer part of the orbita is depressed and pushed out in such a way that this peculiarity did not escape the authors of the first notices about the young man in the American newspapers. Consider the portraits of Kepler, Newton, Leibnitz, Huyghens, Descartes, Euler, Roberval, Bernoulli, Lagrange, de la Place, Lalande, Herschell, Olbers, Bessel, Monge, Carnot, etc."

P. J. Möbius, a psychiatrist and disciple of Gall, and grandson of August Ferdinand Möbius, who published a book on *The Talent for Mathematics* (he also published a pamphlet "The Physiological Mental Weakness of Woman" which tarnished his reputation), says:

> "Gall's statements about the mathematical organ are entirely correct, but I have three remarks to make: 1. Nature varies the forms more than might appear from Gall's description. 2. The mathematical organ is not equally developed on both sides, but as a rule more strongly on the left. 3. The mathematical organ consists in part of a thickening of the soft parts. ... In fact, we find these modifications in almost all great mathematicians. However, sometimes the ridge running downward and backward from the lateral end of the eyebrows is more prominent and sometimes the slope of the outer edge of the orbita. Often the eye stands normally open, often the outer half of the lid hangs down so that the eye is half closed and seems to slant. If we observe the face from in front, it also strikes us that the distance between the outer angle of the eye and the contour of the face is strikingly great. Sometimes the space between the outer angle of the eye and the end of the eyebrow is particularly broad, the distance between the eye and the edge of the forehead magnified. In such cases there is usually neither a distinct ridge nor a slope; the origin is then, so to say, built upward. In individual cases the angular process is uniformly broadened with indefinite outlines, in others it projects with a sharp edge. The description is inaccurate in respect to the variety of nature. A definition that should be in accord with all variations is hard to be found. It would be, perhaps, best to say the mathematical organ consists in an abnormal formation of the angular process that amounts to magnification of the space inclosed by the process. We can say that on no two heads is the organ formed in exactly

> the same way. ... For the important fact that the mathematical organ is to be found prevalently on the left, I need only refer to portraits. In the case of only moderate mathematical ability a distinct modification is usually to be found only on the left. But at times there is scarcely a perceptible difference between the left and the right, for example, in Leibnitz. At times, again, the right as well as the left is abnormally formed, but the one is entirely different from the other, for example, in Gauss. Here, also, the variations are many and it is neither practicable nor advisable to describe all particular cases in words. ... Even in considering portraits one comes to the thought that the prominences are to be referred in part to an abnormally strong development of the skin. If we examine the living, we can convince ourselves by feeling that a considerable hyperplasia of the soft part exists. We feel distinctly that the angular process is indeed uncommonly strongly developed, but also that the skin, more definitely the hypodermal tissue, is thickened. The skin with an abundant cushion of fat forms a loose sack laid around the angular process. Very frequently we see also strikingly heavy eyebrows."

With this information, you will undoubtedly be able to recognize the next mathematician you meet on the street, and even judge the quality of this mathematician at a glance. And mathematics hiring and grad admissions committees will have a much easier job – candidates will not even have to fill out applications, just send a photo or preferably a bust. (Also note that the author has "strikingly heavy eyebrows," although he lacks an "abundant cushion of fat forming a loose sack laid around the angular process," making him a good but not great mathematician.)

XXIV.3.3. Mathematics and Mental Illness

This is a sensitive subject that I discuss with some hesitation. Mathematicians are often portrayed stereotypically in films and television as social misfits of borderline sanity. There are well-known examples of mathematicians who have suffered from mental illness, such as CANTOR, GÖDEL, and NASH, and other mathematicians whose behavior is at least quite odd by conventional standards (e.g. GROTHENDIECK, PERELMAN). And then there is TED KACZYNSKI. Popular books, particularly about logic, sometimes seem to suggest that sustained intense thought about mathematics and logic can, and typically will, drive a person insane.

An extreme (and fortunately not so well known) example was ANDRÉ BLOCH, who spent most of his life in an asylum for the criminally insane after killing his brother and two other family members out of fear that they were showing symptoms of hereditary mental illness present in his family history. While in the asylum he did a substantial amount of solid mathematics, particularly in Complex Analysis, and regularly corresponded with leading mathematicians, most of whom apparently did not know he was institutionalized. (He did do odd things such as date all his letters April 1 no matter when they were written.)

But mathematicians in fact appear to be no more, or less, susceptible to mental illness than any other group of people; in any category of people of substantial size, some percentage can be expected to have mental illness. For example, one can easily identify at least as many professional athletes as mathematicians who have had mental health problems, not even counting athletes who suffer brain injuries. Studies have shown that schizophrenics typically have below-average aptitude in mathematics.

Autism (which is not really mental illness) is possibly a different matter. There is evidence of positive correlation between certain types of autism (e.g. Asperger's) and some types of mathematical ability, at least the ability to focus and concentrate on the mental process needed to do mathematics. It is likely that some prominent mathematicians (possibly including NEWTON) have been mildly autistic (see [**?**]). But the vast majority of mathematicians are in no way autistic.

It is true that many mathematicians are free thinkers who have an unconventional view of the world; one might even expect that the ability to transcend conventional ways of thinking is an important, even necessary, part of being successful in mathematics. And many mathematicians have rather low regard for social customs and norms, although many more are quite conventional socially. Part of the reason many mathematicians are somewhat unconventional may be that they can feel rather free to be so, due to the extreme degree of tolerance in the world of mathematics: mathematicians, probably more than almost any other social group, judge each other to a great extent simply on the quality of their work and ideas and not on social conformity.

> "No one ever made a difference by being like everyone else."
>
> *P. T. Barnum*[6]

It is also not accurate to characterize mathematicians as misfits who are loners or uncomfortable around others and in social situations. There are some such mathematicians, just as there are such people in most other walks of life, and mathematics may indeed be a comfortable refuge for such a person with sufficient mathematical talent; but most mathematicians are pretty adept and comfortable in social situations (VON NEUMANN was reputedly the life of every party he attended, and he is far from unique among mathematicians

[6] Line from the 2017 film *The Greatest Showman.* There is no documentation that BARNUM actually said it.

in this respect). And mathematicians tend to have broad interests and knowledge, and are not at all single-minded (except when they get working intensively on a mathematical problem, or some other project – mathematicians are very good at focusing their attention when necessary); you will find that mathematicians are often very interesting to talk to about almost any subject.

> "[M]athematicians can be funny. They are the kings of logic, which gives them an extremely discriminating sense of the illogical."
>
> *Alex Bellos*[7]

[7] [?, p. xi].

Standard Notation in this Book

We will use the following standard (and sometimes not so standard) symbols to denote commonly used mathematical structures. In a few cases, the same letter is used to represent more than one structure; hopefully no confusion will result.

B^n	Closed unit ball in $\mathbb{R}^n$
$\mathbb{C}$	Field of complex numbers
$\mathbb{D}$	(1) Ring of dyadic rational numbers () (2) Open unit disk in $\mathbb{R}^2$
$\mathbb{F}$	A field, usually $\mathbb{R}$ or $\mathbb{C}$
$\mathbb{H}$	(1) Division ring of quaternions () (2) Hilbert cube $\mathbb{I}^{\mathbb{N}}$
$\mathbb{I}$	Closed unit interval $[0, 1]$ in $\mathbb{R}$
$\mathbb{J}$	Set of irrational numbers in $\mathbb{R}$
$\mathbb{K}$	C*-Algebra of compact operators on ℓ^2
$\mathbb{M}_n$	C*-algebra of $n \times n$ matrices over $\mathbb{C}$
$\mathbb{N}$	Natural numbers (positive integers)
$\mathbb{Q}$	Field of rational numbers
$\mathbb{R}$	Field of real numbers
$\mathbb{R}^n$	Euclidean space of dimension n (set of n-tuples from $\mathbb{R}$)
S^n	Unit sphere in $\mathbb{R}^{n+1}$
$\mathbb{T}$	Topological group of complex numbers of modulus 1
$\mathbb{Z}$	Ring of integers
$\mathbb{Z}_n$	Additive group or ring of integers mod n
$\mathbb{Z}_p$	(1) Additive group or field of integers mod p (p prime) (2) Ring of p-adic integers ()

The symbol 👍 denotes the end of a proof, and ✋ denotes "caution" or, at the end of an argument, "this is not really a proof as it stands." Other hand symbols occasionally used should have obvious meaning.

XXIV.4. Literature Review

"People who like this sort of thing will find this the sort of thing they like."

Abraham Lincoln[8]

We will give brief reviews here of some of the important books on real analysis and measure theory, and some other topics discussed in this book. In spite of any negative or qualifying comments I make here, almost all of these books are worthwhile to read and study. I have not yet included nearly all the books I am aware of that should be on the list, and I have not included any of the books that should be on the list that I am not aware of.

I generally try to write my own review before looking at any other reviews, and in most cases I have not read any other reviews; in cases where I have and other reviews differ substantially from mine or make additional points, I sometimes add comments to that effect (with attribution).

Undergraduate Analysis

Browder, *Mathematical Analysis: An Introduction* [Bro96]

This nice book covers undergraduate analysis, both one-variable and multivariable, as well as basic measure theory, and ends with some elementary differential geometry (manifolds and differential forms). Textbooks written by first-rate mathematicians often have a certain quality not found in some other books; the author's overall approach, insights, and perspectives are quite illuminating. This book covers a lot of ground in a readable way, although some topics covered are not treated in depth.

The author is one of three brothers who are or were all eminent mathematicians. Their father was Earl Browder, a well-known American communist.

Fichtenholz, *Differential and Integral Calculus* [Fic89]

In Russian. This classic three-volume treatise on undergraduate Real Analysis was used as a text by generations of students in the Soviet Union. Far more than just a calculus text. A German translation is available, but not an English one.

Fichtenholz, *Foundations of Mathematical Analysis* [Fc57]

In Russian. This two-volume set is a reworking of parts of *Differential and Integral Calculus*, and is aimed at a slightly higher (post-calculus) audience. A generally careful and readable exposition, although it suffers from excessive hand-waving in places (a criticism which can probably be made of most books on this list, including my own). Contains some material and viewpoints not readily found in Western texts. Unfortunately, no translation is available.

Godement, *Analysis I* [God04]

A rather unique treatment of basic analysis, written in an unusual (for a mathematics book) narrative style, with lots of digressions to apparently unrelated subjects, many of which are connected with the author's obsession with militarism and nuclear weapons (while I admire his dedication on these matters, and mostly agree with his sentiments, the comments about them in the personal preface would have sufficed). Not all of his other digressions are as lofty, including some rather shockingly crude sexism on p. 9. On the mathematical

[8]Book review; cf. R. Byrne, *The Third and Possibly the Best 637 Best Things Anybody Ever Said*, 514.

front, his frequent commentaries on the nature and history of the mathematics make interesting reading and are quite informative. Some of the pronouncements are dogmatic and even narrow-minded, for example the statements on p. 382 concerning a French textbook that defines e^{ix} to be $\cos x + i \sin x$:

> "It would be difficult to imagine a more aberrant conception of mathematics: Euler's most famous formula degenerating into a pure and simple *notation*, and furthermore incomprehensible! ... It is difficult to imagine that mathematicians, even "applied", could ever have advocated it."

(see III.6.2.14.; of course, I didn't leave it at that! cf. XII.1.1.8.(iii), XII.9.9.10.. To be fair, the author does not really suggest that e^{ix} has any intrinsic meaning, as some unsophisticated references seem to; the meaning is carefully defined over a number of pages before this comment.)

The mathematics in this book is more carefully and precisely presented than might be apparent from the narrative presentation, although I did find a few errors, obscurities, and misprints (as I do in almost every book). It is pure mathematics, emphasizing concepts over calculations and with almost no mention of applications.

The author's admonition that "One does not learn analysis or anything else from one single book; there is neither Bible, nor Gospel nor Koran in Mathematics" should be well taken. This is an interesting and useful book which can help give part of the big picture of analysis.

This book is Volume I of a four-volume sequence. I am not yet prepared to comment on Volumes II-IV.

Körner, *A Companion to Analysis* [Kör04]

An insightful and entertainingly written account of undergraduate analysis. Treatment of the topics covered is mostly excellent. Although the mathematics is "honest" and carefully done, it is unsystematic and somewhat informal in places, and not all essential topics are covered, which make the book rather unsuitable as a text. But for its intended purpose as a "companion", it is excellent.

Krantz, *Real Analysis and Foundations* [Kra05]

Like "little Rudin," bridges undergraduate- and graduate-level analysis. I am not terribly impressed with this book, although it has some good features: it is generally pretty solid and complete, and gives a good introduction and motivation for interesting and important applications. But the exposition and organization are of uneven quality, and there are more errors than there should be in a book like this, especially a second edition. (Krantz is very prolific as a writer, but this comes at a price: his writing tends to suffer from a lack of polish and attention to detail, although much of what he writes is pretty good.) He claims that the writing style of this book, unlike older books, speaks to modern-day students in their own language, but I fail to see much difference in style between this book and older books; maybe he simply means the substantial number of statements in the text specifically addressed to the reader, which is indeed rather unusual writing style for a mathematics text but which I don't imagine makes much practical difference to most readers.

The "preface" is primarily a masterpiece of self-promotion. It also contains the ridiculous assertion (written in 2005) that "real analysis has not appreciably changed in 150 years," overlooking such fundamental work as that of Cantor, Dedekind, Lebesgue, ..., as well as most of the foundational work of Weierstrass (he did not yet even have a university position in 1855) and almost all of topology and functional analysis (and much more). Even the Riemann integral had just barely been developed and was not yet published or widely known. At least half of the present book, and a substantial portion of Krantz's book, cover real analysis unknown, or at least not properly understood, in 1855.

If he had said "30 years" instead of "150 years," the statement would have been more defensible, and more in line with what he was claiming in the preface to the first edition: that some other authors of recent

analysis books have said that the subject *has* changed in the last 30 years. But this is a straw man; I am not aware of any author who has made such an assertion.

All in all, this book does not live up to the author's hype.

The above comments were written about the second edition. A third edition has recently appeared, which has been revised into more of a traditional undergraduate real analysis text: the chapters on Lebesgue measure, wavelets, and differential forms have been removed. A chapter on functional analysis has been added (complete with introductory text directly contradicting the "150 year" statement). The remaining text is a little more refined, but the general comments above still apply. A quick and cursory perusal of the third edition turned up some sample howlers: every metric space is separable (section 13.2 problem 17; he says that this is "fairly tricky to prove" – I'll say!); every bounded operator on L^2 is "essentially" a multiplication operator by the Spectral Theorem (section 14.4 problem 9); and in section 14.6 problem 6, a "norm" is defined which is not a norm (this topological vector space $\mathcal{C}^\infty([a,b])$ is in fact not normable, although it is completely metrizable with a metric defined by a formula similar to the "norm" formula, cf. XX.4.3.17.). There are also a number of inaccurate internal references, which is strange since accurate internal references are so easy in LaTeX. Judging by the new preface, the author still has a pretty high opinion of himself and his book. (I should note that the prefaces to some of Krantz's other books, such as [Kra15], have a far different tone and create a different picture of the author, who I must emphasize I do not know personally.)

Now there is a fourth edition. The slimming down of topics that began in the third edition has continued: for example, the chapter on functional analysis which appeared (with all its difficulties) in the third edition has disappeared just as suddenly. The author says he doesn't want people to think this is a "high-level" text (I don't think there is much danger of that; he also says he wants the text "to be the generational touchstone for the subject and the go-to text for developing young scientists" – a laudable but pretentious goal which is not likely either), but he still claims that this book is more rigorous and at the same time more readable than anything else since Little Rudin. But I don't see either more rigor or more readability than, say, in [Pug15] (or even as much in this case).

For example, the discussion on p. 1 of the inadequacy of $\mathbb{Q}$ for analysis is logical mush: the claim is that $\mathbb{Q}$ is "not closed under limits" (apparently meaning limits of sequences) and must be expanded to a number system which is. But this argument presupposes that we already have a (perhaps still ill-defined) larger number system in which limits can be taken: if $\mathbb{Q}$ is the only number system we have, it is tautologically closed under limits of convergent sequences. And if we want limits of all sequences, $\mathbb{R}$ is also not closed. He tries to assert that every sequence of real numbers which "should" converge actually does. This is evidently an informal reference to the idea of a Cauchy sequence, a somewhat subtle notion which would certainly not leap to mind for someone (other than a Cauchy) who hadn't seen it before. Many people would say the sequence $(1, 2, 3, \dots)$ "should" converge to $+\infty$, but $+\infty \notin \mathbb{R}$, so not every sequence in $\mathbb{R}$ which "should" converge does, i.e. $\mathbb{R}$ is not closed under limits either (e.g. in the extended real line). I'm not saying it is easy to give an elementary yet mathematically and logically honest introduction to the notion of completeness, but one can certainly do a lot better than the discussion in the book.

On p. 25, he proves that a sequence in $\mathbb{C}$ converges if and only if it is Cauchy (after proving this for real sequences) by making the "observation" that a sequence in $\mathbb{C}$ converges (resp. is Cauchy) if and only if the sequences of real and imaginary parts converge (resp. are Cauchy). There is no previous discussion of anything like this. In fact, both "observations" are propositions which need to be proved. Of course, the proofs are very easy, but they are not trivial, especially for someone who has just seen the definitions of the concepts; at the very least, these should be expressly designated as exercises for the reader, not just "observations."

On p. 290, it is asserted that "a moment's thought" shows that an open spherical shell cannot be written

as a disjoint union of open balls. Maybe so. But another moment's thought reveals that this is actually quite nontrivial to prove (and I'm not sure why the question is even relevant to the discussion on that page).

In fact, almost every page I looked at contained something I could quibble with, although many of the criticisms are petty ones concerning the author's writing style. For example, on p. 19, he says, "If a sequence does not converge then we frequently say that it *diverges*." Frequently? When do we not? (OK, I guess we don't always actually *say* it diverges, but we *could*.) Why not just say: "A sequence *diverges* if it does not converge."?

I guess *someone* (other than the author) must like this book, since it has been through four editions.

Pugh, *Real Mathematical Analysis* [Pug15]

This is overall an excellent book, and is one of the best references for learning and understanding real analysis. The author's explanations and insights on most topics are outstanding. The book is mathematically honest and generally rigorous. However, many arguments are written a little informally, with some routine details glossed over; thus a student using this as a textbook might have trouble developing the skill of writing complete rigorous analysis proofs if the book's proofs are taken as a model (to be sure, there are plenty of proofs in the book written out fully enough to meet anyone's standards).

Rudin, *Principles of Mathematical Analysis* [Rud76]

"Little Rudin" or "Baby Rudin." One of the most widely-used analysis books for a long time, with good reason. Covers pre-measure-theory analysis with a last chapter on Lebesgue measure.

Schumacher, *Closer and Closer* [**?**]

A leisurely and non-stressful introduction to undergraduate real analysis. This is a pretty good book, despite its problematic title (cf. IV.1.3.11.); it is nicely written, the mathematics is generally carefully done and is not "dumbed down," and a surprising amount of material is covered. The discussion of the Implicit Function Theorem is especially good. The treatment is mostly mathematically "honest" (I.8.), although some subtleties, e.g. the Axiom of Choice, are not mentioned. One mistake I noted is the wrong definition of countable compactness (p. 141), confusing *countably compact* with *Lindelöf*. The approach is an "active learning" one, based on the Socratic method: some arguments are given in great detail, but many are left as exercises for the reader, often with little hint as to how to proceed. Such an approach is certainly defensible pedagogically. This is a good supplementary text for a reader desiring a gentle but reasonably thorough treatment of the subject.

Stromberg, *An Introduction to Classical Real Analysis* [Str81]

This is a terrific book (although I would not describe the writing as a literary masterpiece). It is thorough and detailed, but also has much helpful explanatory material. The first few chapters cover standard (one-variable) undergraduate analysis at a similar level to most undergraduate texts, but the pace and depth increase when integration is reached. Riemann integration is not treated, except in passing after the fact; instead, the Riesz-Daniell approach to Lebesgue integration is followed. Some complex analysis is interwoven and integrated into the development at appropriate points. The exercises in this book are especially notable. This is not only a textbook (probably most suited to an honors-level analysis course), but also a great reference work even for mature analysts.

Wade, *An Introduction to Analysis* [**?**]

This is a decent book on undergraduate real analysis, covering both the one-variable and multivariable theories, but there are many small things I would wish to do differently. The fourth edition has a number of improvements over the third edition (although the deletion of the chapter on differential forms is not one of them), but the fourth edition seems "cheaper" than the third in both typesetting and binding. The author also did not do a very good job proofreading this edition before publication. There is a long list of typographical errors from the first printing on his website which have been corrected in later printings, but a significant number of typographical and mathematical errors remain.

Historical Calculus Texts

A selection of historical and/or obsolete calculus texts. These are interesting to peruse to see how exposition of the subject has evolved, if for no other reason.

Courant, *Differential and Integral Calculus* [CJ99]–[CJ89]

I only recently looked carefully at this two-volume book. I was somewhat concerned that it would be a cookbook-type treatment lacking in rigor after reading the introduction (as well as some comments in [CR79] on the nature of mathematics), which begins:

> "Although there is no lack of textbooks on the differential and integral calculus, the beginner will have difficulty in finding a book which leads him straight to the heart of the subject and gives him the power to apply it intelligently. He refuses to be bored by diffuseness and general statements which convey nothing to him, and will not tolerate a pedantry which makes no distinction between the essential and the non-essential, and which, for the sake of a systematic set of axioms, deliberately conceals the facts to which the growth of the subject is due."

My concern was unfounded. This was the best calculus/elementary real analysis textbook written up to its time of publication (1927 for the German edition, 1934 for the English edition), and is still a good treatment, although probably not suitable as a textbook today (but still worth reading as a supplemental text). The theme is to develop differentiation and integration simultaneously as inverse processes, an approach which is both conceptually and historically valuable. It is far more carefully written than, say, [Gri21] (which preceded it by only a few years), and is a model of maintaining mathematical honesty along with an appropriate level of detail in the exposition. In fact, I have found almost nothing substantial to quibble with. The author is not hesitant to express opinions on various points, which make for interesting reading (cf. V.3.1.26.). The English editions differ somewhat from the original German edition.

The review by R. Boas in the *Mathematical Reviews* is somewhat less complimentary, referring to the books as "museum pieces." The review also states that at least one fallacious proof from the 1927 German edition has not been corrected in later editions, but does not say what it is. I have found one seriously inadequate (not exactly wrong) proof, the argument on p. 291-292 of volume 2, that a curve with a continuous parametrization which is continuously differentiable except at finitely many points has Jordan content zero. The reduction in the argument is incorrect, and such a curve need not even be rectifiable (e.g. XIX.4.5.2.). But the difficulty can be avoided by first enclosing each singular point in a small square and then reducing subintervals away from the singularities as in the argument. (The result is also not stated precisely enough; the parametrization must be on a closed bounded interval.)

Edwards, *An Elementary Treatise on The Differential Calculus*

This was evidently a popular calculus text in the late nineteenth century, since it went through at least three editions; my copy is a later printing of the third edition from 1906.

This text is almost entirely mechanical; mathematical rigor is almost nonexistent. The most egregious gap is that there is, as far as I can determine, no suggestion that limits, and thus derivatives (called *differential coefficients*), do not always exist. Calculus is pretty easy if you don't worry about the existence of limits! (It is, of course, a remarkable and significant fact that they mostly do, at least if you stick to combinations of elementary functions, and you can get a long way by mere formula manipulation, although you can get into occasional trouble even here.)

For example, here is the entire treatment of the Implicit Function Theorem:

> "If we have
> $$\phi(x, y) = 0$$
> then
> $$\phi(x + h, y + k) = 0.$$
> Hence
> $$\frac{\phi(x + h, y + k) - \phi(x, y + k)}{h} + \frac{\phi(x, y + k) - \phi(x, y)}{k} \cdot \frac{k}{h} = 0;$$
> which, when h and k are indefinitely diminished, becomes (as explained in Art. 159 [total differential, similarly nonrigorous])
> $$\frac{\partial \phi}{\partial x} + \frac{\partial \phi}{\partial y} \cdot \frac{dy}{dx} = 0,$$
> or
> $$\frac{dy}{dx} = -\frac{\frac{\partial \phi}{\partial x}}{\frac{\partial \phi}{\partial y}} \ .$$
> This is a *very useful formula* for the determination of $\frac{dy}{dx}$ in cases in which the relation between x and y is *an implicit one*, of which the solution for y in terms of x is inconvenient or impossible."

The definition of limit in Article 10 is a fairly careful (narrative) version of the usual definition of an ordinary limit of a numerical function, not valid for an "infinite" limit (and he declares that the expression is "ultimately" equal to the limit, not really wrong if the meaning is defined to be this, but possibly misleading). But a much more expansive notion of limit seems to be in the author's mind: the first rather curious example of limit is

> "If an equilateral polygon be inscribed in any closed curve, and the sides of the polygon be decreased indefinitely and at the same time increased in number indefinitely, the polygon continually approximates to the form of the curve, *and ultimately differs from it in area by less than any assignable magnitude*, and the curve is said to be the *limit* of the polygon inscribed in it."

[There is an assertion here that needs to be proved, maybe with some restriction on the curve like rectifiability; and what does "continually" mean since the polygon is a discrete function of the number of sides? Also, what does "inscribed" mean if the curve is not convex? And how does this "limit" of curves illustrate the definition?]

The only whiff of suggestion that there might be a problem with existence of limits that I have found is in the discussion of tangents and concavity, where it is acknowledged that vertical tangents can occur, where

$\frac{dy}{dx}$ is "infinite" (the author then says that all higher derivatives are also "infinite" there). But not to worry: just interchange x and y, and the derivative then becomes zero.

Two arguments are given purporting to prove equality of mixed partials, both based on a vague assertion that "processes of differentiation take no cognizance of the *particular values* of any constants involved," neither of which hold water.

There is eventually a little more care, if not rigor. In the discussion of Taylor series, there is an (oversimplified) discussion of continuity (only jump discontinuities are considered), and an acknowledgment that the Taylor series of a function does not converge if the function or one of its derivatives becomes infinite or has a discontinuity, or even if the error terms do not approach zero – i.e. the Taylor series converges to the function except when it doesn't (evidently sequences can fail to have limits or series can fail to converge, but functions always have limits).

The Intermediate Value Theorem is considered "obvious," and it is asserted that a quantity can change from positive to negative or vice versa only by "passing through 0 or infinity." And in discussing maxima and minima, it is implicitly assumed that a function can change sign only at a discrete set of points (i.e. the converse of the First Derivative Test is essentially argued).

On the other hand, extensive and detailed applications are given to the geometric structure of curves, with intricate constructions rarely discussed in modern texts. For example, I learned what the "roulette" of a curve rolling on another curve is, and how to describe and calculate it.

The author argues that the term "indeterminate form" is not the best terminology since the limit is not really "indeterminate" (he of course does not worry about whether it *exists*, but merely whether and how it can be calculated; he calls the limit as $x \to a$ the *true value*, or sometimes the *true limiting value*, of the expression at $x = a$), and suggests the terms "undetermined" and "illusory" instead; but I don't find these terms any better, and he does acknowledge that even then the term "indeterminate form" was well established. A similar argument is given in Courant.

Euler, *Foundations of Differential Calculus* [Eul00]

English translation of the first part of Euler's 1755 treatise on calculus. Primarily of (considerable) historical interest today, although almost all of the mechanical parts of calculus covered today in first-semester courses are here and treated in a nicely understandable way, if not always too rigorously (this was the first text about which this can be reasonably said), and as in most writings of great mathematicians there is much wisdom and insight in the exposition. And the man certainly could calculate! But, on the other hand, it is remarkable and revealing that even one of the greatest mathematicians had such a degree of misunderstanding about things we today regard as fairly basic, both in mathematics and physics (e.g. Euler seemed to consider it self-evident that matter is infinitely divisible). Throughout this book there is much beating of gums about the nature of infinity and the "infinitely small," which he repeatedly states is actually zero or nothing while at the same time can be used as a divisor of another infinitely small quantity. This book graphically illustrates how far our understanding of the foundations of analysis has progressed since then.

Griffin, *An Introduction to Mathematical Analysis* [Gri21] and *Mathematical Analysis: Higher Course* [Gri27]

A widely-used two volume textbook sequence for the first two years of university mathematics, from the 1920s (would that more of our entering students today were able to take a course at this level! Granted, a much higher percentage of the population attends universities now than 100 years ago.) The books cover roughly what we do today in a two-year calculus sequence, along with trigonometry and some statistics. There is not a high level of rigor, and many things are oversimplified and/or not justified adequately, but

the books are fairly "honest" mathematically, with numerous cautionary comments of the type we might give today, such as:

> "**We cannot divide by zero.** Erroneous statements are often made in this connection about 'infinity' – whatever that may be. The correct statement is that we cannot divide by zero."

> "It is highly useful to give a meaning to a power x^n even when the exponent is negative, fractional, or zero. Obviously x^n for such n has no meaning [in the usual sense] and is quite absurd ... [The definition is then introduced.] ... These definitions are the only ones possible if the usual laws of exponents are to apply in all cases. ... It is a great advantage [to have these definitions of power functions]. This is possible because we have *generalized the meaning* of a power, to allow n to be negative or fractional. Such generalizations are very typical of modern mathematics."

> Concerning complex numbers, "The names 'real' and 'imaginary' are very misleading, as they suggest that one kind of number actually exists and that the other does not. The fact is, these are merely *different kinds* of numbers. Imaginary numbers can be given a perfectly concrete interpretation, which makes them exceedingly useful."

The definition of limit is actually correctly given, although just described in words (somewhat bizarrely using the symbol n to denote what we usually call ϵ). But some limits, including the crucial one defining e (V.5.3.5.), are simply computed (estimated) numerically with no proof that they actually exist, or even suggestion that such a proof is necessary.

There are a few misstatements and misleading statements due to oversimplification, such as an assertion that the converse of the First Derivative Test holds (cf. V.3.10.10.). The derivation and proof of the Product, Quotient, and Chain Rules are mathematically inadequate (V.3.10.5., V.5.5.8.). The following "proof" that two functions with the same derivative differ by a constant is typical of the level of rigor in the books:

> "For if the derivatives are equal, the functions must be changing at the same rate; and therefore their difference is not changing but remains constant."

(I have not found any mention of the Completeness Property of $\mathbb{R}$ in either volume; cf. V.12.3.5.(iv).) And there is apparently nothing to prove with the Implicit Function Theorem:

> "Any equation involving x and y requires y to vary with x in some definite way. If the value of y is not given in terms of x explicitly, i.e. in the form $y = f(x)$, we say that the equation defines y 'implicitly' as a function of x; or that y is an *implicit function*."

He does sometimes acknowledge things like

> "Certain critical questions arise in some of these steps which will not be discussed here."

The style is more cookbook than we like today, with many facts and formulas simply stated (one can usually find some justification by looking hard enough, although often not a mathematically adequate one), with the admonition that the student should "memorize" them.

These books give a picture of the state of lower-division undergraduate mathematics instruction in the early-to-mid twentieth century.

Osgood, *Advanced Calculus* [**?**]

This is the second volume of a two-volume sequence from the 1920s. The first volume is titled *Introduction to the Calculus*, but evidently (I have not yet had an opportunity to review the first volume) it would be

more accurate to describe the first volume as one-variable calculus and the second as multivariable calculus, although it also contains chapters on differential equations and the calculus of variations. The topics covered are similar to the ones in [CJ99]–[CJ89], to which it is compared in the review of that work by R. Boas in the *Mathematical Reviews*. That review complained that Courant's book relied excessively on motivations from physics, but Osgood's book is tied far more closely to physics. In fact, Osgood, although he was one of the best American mathematicians of the early twentieth century, seems to have had a decidedly old-fashioned view that mathematics and physics are not really distinct subjects. Indeed, on p. v he calls geometry "the noblest branch of physics" and again on p. 65 when he attempts to justify the integral formula for the area of a surface () he states:

> "It is a fundamental principle of elementary geometry to refer all geometrical truth back directly to the definitions and axioms. What are the axioms on which this formula depends? The answer is: *The formula itself is an axiom.* The justification for this axiom is the same as for any other physical law, namely, that the physical science, here geometry, built on it is in accord with experience."

The level of rigor in this book is higher than in the contemporary [Gri27], but still not entirely adequate, although the book is pretty honest mathematically and the author acknowledges many technicalities which must be proved but aren't, and references are often given. He does sometimes say things like "It is here not a question of a rigorous proof of a theorem, the truth of which no one doubts." (p. 193) And in discussing a problem in the calculus of variations, he states on p. 408:

> "[W]e are not concerned here with the difficult question of proving mathematically that our problem (after still further restrictions) has one and only one solution, and that this solution satisfies the further demand of possessing a second derivative. On the contrary, we take for granted, in a given case, as more or less plausible from the physical evidence, that the problem will admit a unique solution, continuous together with its derivatives of the first two orders, for we know from experience that problems with a physical pedigree do usually admit the kind of solution expected – and we then turn our efforts to finding this solution."

He does, though, go on to say: "Incidentally, we get some interesting surprises as to what was to be expected from the physical evidence."

In fact, a considerable portion of the book is devoted to applications of various aspects of calculus to physical problems.

I could not classify this book as a classic to the same degree as [CJ99]–[CJ89] (which was somewhat later), but it was about the best English-language text available at the time.

Real Analysis, Measure Theory, and Integration

These textbooks cover the basic theory of measure and integration, and most also treat the basics of functional analysis.

Many general measure theory texts follow what I call the "reinventing the wheel" plan: first develop Lebesgue measure and integration, then redo everything in general, in many cases just referring back to the Lebesgue case for proofs of results. This approach is certainly defensible pedagogically, and is preferred by many people, but it is not to my personal taste; I much prefer to start out with general measure theory, spending considerable time discussing and emphasizing the critical special case of Lebesgue measure as I go, as well as discussing other special cases such as probability spaces.

> "One of the many fundamental questions which any author on the subject must decide, is whether to begin with 'general' measure theory or with 'Lebesgue' measure and integration. The point is that Lebesgue measure is rather more than just the most important example of a measure space. It is so close to the heart of the subject that the great majority of the ideas of the elementary theory can be fully realised in theorems about Lebesgue measure. ... If you take the view, as I certainly do when it suits my argument, that the business of pure mathematics is to express and extend the logical capacity of the human mind, and that the actual theorems we work through are merely vehicles for these ideas, then you can correctly point out that all the really important things in [elementary measure theory] can be done without going to the trouble of formulating a general theory of abstract measure spaces; and that by studying the relatively concrete example of Lebesgue measure on r-dimensional Euclidean space you can avoid a variety of irrelevant distractions.
>
> If you are quite sure, as a teacher, that none of your pupils will wish to go beyond the elementary theory, there is something to be said for this view. I believe, however, that it becomes untenable if you wish to prepare any of your students for more advanced ideas. The difficulty is that, with the best will in the world, anyone who has worked through the full theory of Lebesgue measure, and then comes to the abstract theory of general measure spaces, is likely to go through it too fast, and at the end finds himself uncertain about just which ninety percent of the facts he knows are generally applicable. I believe it is safer to keep the special properties of Lebesgue measure clearly labelled as such from the beginning."
>
> *D. Fremlin*[9]

The "reinventing the wheel" approach seems to me like presenting Linear Algebra by first stating and proving all the theorems of the subject for $\mathbb{R}^n$, and then restating and partially reproving them for general vector spaces.

Bogachev, *Measure Theory* [Bog07]

A very thorough two-volume treatment of measure theory. In order to cover the huge volume of material, many topics, some of them important, are relegated to "exercises" where arguments are only sketched or outlined.

The English translation was apparently done by the author himself. Although this was a commendable effort, and in most places the translation is pretty good, the results are not entirely satisfactory, and in a

[9] [**?**], V. 1, p. 9.

few places the reader must puzzle out what the author is trying to say (of course, in many books written by native English speakers, probably including my own, the reader must do the same thing!)

Dieudonné, *Foundations of Modern Analysis* [Die60]

This well-regarded book, the first of a comprehensive nine-volume series on analysis, develops basic analysis from a "modern" point of view, in the BOURBAKI tradition (the author was a leading member of BOURBAKI). To illustrate the approach, note that Banach and Hilbert spaces are treated before reaching differential calculus, which is first done on Banach spaces and then specialized to the real numbers. The exposition is elegant and carefully done. The author has strong opinions on many subjects and is not hesitant to express them, not all of which I agree with; although he was a formidable mathematician with a strong command of the big picture of mathematics, he had his own point of view (among other things, very geometric) which I do not always share. This is a valuable and widely-used book, but I am not convinced the approach is necessarily the best one for someone learning the subject. Concerning the series as a whole, it has been suggested to me that my own book is unnecessary since Dieudonné already covered the entire subject (I of course disagree!)

My opinion of Dieudonné is colored by my personal experience with him, which is admittedly quite limited and may have given an inaccurate picture. My only encounter with him came when he gave a colloquium at Berkeley while I was a graduate student. The talk was about a technical result on Lie groups which was probably not appropriate for a general mathematical audience, made worse by the fact that he repeatedly refused to define terms or discuss general considerations, sneering that he didn't have to explain such things at *Berkeley*. The student sitting next to me seriously considered asking "What's a Lie group?" at the end (this student knew the answer very well). It was the most arrogant mathematics talk I have ever heard, and I remember thinking I had never met anyone whose name so exactly described his opinion of himself.

Doob, *Measure Theory* [Doo94]

This book, written by an eminent probabilist, is somewhat idiosyncratic, but valuable. It has nice, efficient treatments of a number of topics not readily found in standard texts. The approach is not especially wedded to probability theory, but the chapter on martingales (invented by the author) is especially valuable for probabilists. Many arguments represented as proofs in this book are really only outlines, and need many details to be filled in by the reader (not necessarily a bad thing so long as the necessity is recognized!)

Dunford and Schwartz, *Linear Operators I* [DS88]

This first of a three-volume set is usually regarded as an encyclopedic treatise on Operator Theory. But chapters 3 and 4 give an extensive treatment of measure theory and related topics. One of the best places to read about finitely additive measures and integration (from the beginning, a "measure" is a finitely additive complex measure), and vector measures.

Folland, *Real Analysis* [Fol99]

This is one of my personal favorite Real Analysis texts. It contains elegant and comprehensive treatments of many important topics. However, the basic measure theory part is much too concise to be a good choice as a text for a beginning course, especially for masters students, as I discovered when I tried it. The approach of this book is the exact opposite of the "reinventing the wheel" plan, in my opinion too much so: it treats Lebesgue measure almost as an afterthought, as a special case of Stieltjes (Radon) measures.

Fremlin, *Measure Theory* [Fre04]–[Fre06]

A monumental five-volume treatise on almost all aspects of the subject. Not suitable as a text, although the first two volumes are designed to be accessible to students, but a detailed and comprehensive reference work.

Halmos, *Measure Theory* [Hal50]

The first (1950) "modern" treatment of measure theory, and still one of the best. It is an excellent treatment of the basics of measure theory (done from the standpoint of what we call "premeasures on rings"), and includes some more advanced topics such as Haar measure. In fact, one can hardly improve on the exposition of many of the topics covered.

Hewitt and Stromberg, *Real and Abstract Analysis* [HS75]

A classic book. It is dense in places and not always easy going, but well written and contains a wealth of material.

Lieb and Loss, *Analysis* [LL97]

This is a valuable book with a little different emphasis than most. The authors' philosophy is to seamlessly merge classical and functional analysis, which they have done well. The treatments of measure theory and basic functional analysis are rather sketchy, but the coverage of L^p-spaces and Sobolev spaces is excellent, among the most complete in any general analysis book. The general approach is a low-tech and computational one, emphasizing inequalities and "hard analysis" arguments rather than abstract general theorems (quite the opposite of Pedersen), an approach which has much merit and should particularly appeal to applied analysts; the primary intended audience for the book is evidently applied mathematicians and mathematical physicists, and there are many physics-related applications woven throughout. The book is written in somewhat of a stream-of-consciousness style, and would have benefited from some more organizational attention and effort on the part of the authors; but on the other hand, it is full of insightful comments and observations. A few things are a little sloppy: for example, the proof of Hölder's Inequality 2.3 uses Jensen's Inequality and thus is only valid as written on finite measure spaces (the inequality is true in general, which can be shown by an additional MCT argument, or see XX.15.2.6. for a simpler proof). This is not a book I would be likely to use as a text, but it is a great supplementary text.

McDonald and Weiss, *A Course in Real Analysis* [MW99]

This is overall one of the better measure theory texts, and one that I have used several times. The treatments of most standard subjects range from fairly good to excellent. Some of the writing is somewhat turgid, and some of the definitions (e.g. Borel sets, absolutely continuous functions) are nonstandard but equivalent to the usual ones. A few proofs are more laborious or roundabout than necessary. A few important basic results, e.g. the Lebesgue Density Theorem, are not covered. This text follows the "reinventing the wheel" plan. This book has a lot of exercises, some of which are fairly challenging, but generally with hints and guidance.

Natanson, *Theory of Functions of a Real Variable*

Another standard Soviet text. Well done, and covers material not found in other books (e.g. the Banach-Tarski paradox). Volume I is pretty routine, but Volume II is more interesting. The English translation leaves something to be desired, and is confusing in places: for example, in several instances the phrase "replace x by y" should be "replace y by x" (this is simply a translation error; the original Russian text is correct).

Pedersen, *Analysis NOW* [Ped89]

The author, a specialist in Operator Algebras, takes an uncompromising approach to measure and integration using functional analysis methods (he says the title means "Analysis via norms, operators, and weak topologies"). This is a valid approach with much to recommend it (in my opinion as a fellow Operator Algebraist), but should probably be complemented by a more traditional treatment. The text is clearly written, and witty in places (different readers will have different opinions about the wit).

Royden, *Real Analysis* [Roy88]

This is one of the "classic" texts (and the one I learned measure theory from as a student). Despite its age, it is still a viable and worthwhile option. The third edition has some additions and improvements, but also more errors than the second edition. This book follows the "reinventing the wheel" plan, to an even greater extent than most. An expanded and updated fourth edition, prepared by M. Fitzpatrick, has recently appeared.

Rudin, *Real and Complex Analysis* [Rud87]

"Big Rudin." Also a classic book, widely used. The author rather uniquely covers and integrates Real Analysis (measure and integration) with Complex Analysis. These subjects are usually treated separately, but there is much merit in the author's approach which highlights the overlap and interplay of the two theories (although I have chosen not to follow this path in this book). Rudin does not follow the "reinventing the wheel" plan; general measure and integration is developed in Chapter 1, and Lebesgue measure is obtained by applying the Riesz Representation Theorem to the Riemann integral.

Saks, *Theory of the Integral* [Sak64]

This was the first comprehensive book on measure and integration (1933), and is very impressive in its breadth and depth, as well as giving a nice overall view of the logic and philosophy of the subject. While the abstract Lebesgue integral on general measure spaces is treated, it is somewhat in passing and the discussion is practically limited to subsets of Euclidean space. At least half the book is devoted to a study of the interrelated topics of indefinite integration, bounded variation, and absolute continuity. The book is still well worth careful study. Treatments of some topics have been improved since this book, but many have not, and indeed this book covers a number of important topics not treated adequately in most newer references. It is overall quite well written and readable, although dense and detailed. Some of the notation and terminology is outdated and requires care on the reader's part to decipher. Definitely a book for adults.

Swartz, *Measure, Integration and Function Spaces* [Swa94]

This book has a rather unconventional approach to the subject with much to recommend it. It treats finitely additive, signed, and complex measures from the beginning rather than as an afterthought. It has a good treatment of limits of measures, the Nikodým Convergence and Boundedness Theorems, and the Vitali-Hahn-Saks Theorem. The treatment of fundamentals of functional analysis includes ordered vector spaces.

Taylor, *Measure Theory and Integration* [Tay06]

A good general treatment of the subject, with topics leading in the direction of geometric measure theory not well covered in other texts in this group. However, the author is not as careful as he should be in specifying finite or σ-finite restrictions in some results and problems (I found several places where such restrictions are necessary and not mentioned).

Complex Analysis

There are dozens, perhaps hundreds, of complex analysis texts. We will only describe a few of the best. Many others are at least decent.

Ahlfors, *Complex Analysis* [Ahl78]

Generally regarded as *the* classic complex analysis text, and still worth studying.

Brown and Churchill

One of the classic complex analysis texts (I used an early edition as a textbook as a student, when it was just Churchill). Oriented primarily for nonmathematics students, e.g. physicists and engineers, but unlike some books of this type it is mathematically solid.

Gamelin, *Complex Analysis* [Gam01]

One of the very best complex analysis texts.

Greene and Krantz, *Function Theory of One Complex Variable* [GK06]

I had a somewhat better opinion of this book before I used it (third edition) as a text for a class. It is overall quite good, especially the later chapters (e.g. the treatment of the Riemann Mapping Theorem is generally excellent). But the first three chapters with the basic theory are somewhat of an organizational mess, and some of the exercises are not well thought out (in fact, I had some trouble finding enough good and suitable exercises for my class). And the authors beat the $\frac{\partial}{\partial z}$ and $\frac{\partial}{\partial \bar{z}}$ notation to death, and insist on writing $\frac{\partial f}{\partial \bar{z}}$ instead of f' for the derivative of a holomorphic function f throughout most of the book, and even $\left(\frac{\partial}{\partial z}\right)^n f$ for the n'th derivative, making for confusing formulas which are hard to read.

Krantz, *Complex Variables, A Physical Approach with Applications and MATLAB*

This book seems designed as a modern replacement for Brown and Churchill. It is a rather slow-paced treatment of basic complex analysis aimed at a nonmathematical audience. Far less detailed and sophisticated than [GK06]. Generally well-done, with good insights and explanations of standard results. The MATLAB part (also covering some other similar packages) is not at all deep, barely going beyond how to do complex arithmetic. A good book to read as an introduction to the subject.

Remmert, *Theory of Complex Functions* [Rem91] and *Classical Topics in Complex Function Theory* [Rem98]

This is more or less a two-volume sequence, but there is a vast difference in depth and sophistication between the two volumes. The first is a treatment of the basic theory, suitable for undergraduates, and the second is a dense, detailed study of more advanced topics. Both are entertainingly written with extensive historical notes, including anecdotes only loosely related to the subject at hand (such as noting that Pringsheim was at first denied Habilitation "on account of the great ignorance of the candidate"; he apparently refused to explain how to solve quadratic equations.) The books are probably not suitable as primary texts, but are excellent supplemental texts.

Stein and Sakarchi, *Complex Analysis* [SS03]

Part of the authors' series of books on analysis. This book stands on its own as well as a volume in the series. Covers basic complex analysis, not as thoroughly as in [Gam01] or [GK06] but well, and then has

detailed treatments of number-theoretic functions and the Prime Number Theorem, and of elliptic functions. A very worthwhile addition to the literature, but perhaps not the best choice as a text for an introductory complex analysis course.

Functional Analysis

Kreyszig, *Introductory Functional Analysis with Applications* [Kre89]

A classic text, and one of the few options for a class without a measure theory prerequisite. It has a vaguely applied orientation, but mainly just covers the basic theory, and covers it well. Still a good choice as a text for an introductory course at the advanced undergraduate level.

Lax, *Functional Analysis* [Lax02]

This is overall one of the best texts on functional analysis, by one of the leading applied analysts of the 20th century. It includes many topics not covered in other standard texts, and the treatment of most topics is excellent. It has thorough coverage of theoretical functional analysis as well as some applied areas. (There are, however, a rather large number of typographical errors, most of which are inconsequential.) The author's applied orientation shows through not only in his choice of material, but also in the style and tone of some of the presentation and his notation and terminology; however, this does not make the book less useful or valuable for "pure" mathematicians. Some of the exposition is more reflective of the 1950s than 2002, when the book was published: for one example close to my heart, the antiquated term "B*-algebra" is used, a term I have not otherwise seen since about 1980; specialists in this field have almost universally used the name "C*-algebra" since the 1960s.

Specialized Topics in Real Analysis

Diestel and Uhl, *Vector Measures* [DU77]

Despite the name, only the first two chapters cover the theory of vector measures and integration *per se*; the rest of the book consists of applications to the structure theory of Banach spaces. The first two chapters give a nice introduction to vector measures and Bochner integration. The authors' personalities and strong opinions make for interesting reading.

König, *Measure and Integration* [Kön97]

This is a highly original book and contains much material not available elsewhere about construction of measures from primitive data. But the author's opaque writing style and frequently nonstandard notation and terminology make for tough reading.

Krantz and Parks, *The Implicit Function Theorem* [KP02a]

The most thorough discussion available of the various versions of the Implicit Function Theorem and Inverse Function Theorem and their histories.

Krantz and Parks, *A Primer of Real Analytic Functions* [KP02b]

This is a quite thorough treatment of the theory of real analytic functions, covering not only the basic one-variable and multivariable theories, but also more advanced topics such as the Cauchy–Kovalevskaya Theorem, quasi-analytic and Gevrey classes, and the embedding theorem for real analytic manifolds. Complete (or at least semi-complete) proofs are given for most elementary results, but as the topics become more advanced the book becomes more of a survey. Generally well written and readable, although, as with many of Krantz's books, there is an annoying lack of quality control and attention to detail in places. (I do not mean to slight the contributions of Parks, who has long been an acquaintance of mine.)

Oxtoby, *Measure and Category* [Oxt71]

This book is rather unique and does not directly overlap with any other books I know of (except for [BJ95], which is a sequel of sorts), although most topics can be found scattered elsewhere. The theme is the duality between the theories of Lebesgue measure and Baire category. Many results in one of these theories have an analog in the other, which is often more than merely an analog. Along the way, topics like the Banach-Mazur game are discussed and woven into the story in an illuminating way.

Schechter, *Handbook of Analysis and its Foundations* [Sch97]

I think the review of this book on MathSciNet is far too negative. This book, if properly viewed and used, is a valuable resource. Despite the name, not too much classical analysis is covered; instead, the author concentrates on an efficient and logical development of much of the mathematics a modern analyst needs to know, beginning with sets, relations, and orderings, proceeding through algebra and logic, topology and uniformity, and concluding with an extensive treatment of topological vector spaces. This skeleton outline does not do justice to the range of mathematics covered. Dependence of results on set-theoretic axioms is carefully discussed.

Globally, this book is very much in the style of Bourbaki in its organization, attempting to present the most efficient logical path through the development of the material. But the local style is very different: Schechter allows himself much room for motivation, explanation, entertaining writing, and even some humor, things famously and deliberately excluded from Bourbaki (although there is some subtle humor in

Bourbaki reflecting the personalities of the members, beginning with the name itself; indeed, according to [RS09], Grothendieck was not a successful member of Bourbaki in part because he "lacked humor.")

I discovered this book when my manuscript was already well along, and I was struck by how many passages were similar to things I had written, tried to write, or planned to write, often done better than I could have myself. Needless to say, I approve of the author's views about many things in mathematics. (This is not to say that this book is overall very similar to my own; there are very significant differences.)

That said, I do not think this book would be an appropriate place to learn about most of the material covered; it would be most suitable for someone who had already been introduced to most of the major topics. The author did not assume any prerequisites beyond calculus except for a degree of mathematical maturity, but the book would be best used to fill in holes in one's background and to see how everything fits together logically.

This book would have been much more effective as an online book with hyperlinks, organized more in the manner I envision for my own manuscript (i.e. less linearly). The author's logical progression of development means that many results and ideas are introduced before they are actually used, and there are numerous cross-references to other parts of the book. If one simply wants to understand a certain topic or result, a maze of tedious searches through the book via cross-references is often necessary to get the full picture, which would be much more efficient and less annoying if done via hyperlinks. For example, there is a very nice and complete proof of the Brouwer Fixed-Point Theorem contained in the book, but pieces of the proof are scattered through several parts of the book. I hope eventually an online edition of this book can be produced with hyperlinks (there is currently a CD-ROM version, but it appears to be identical to the printed version).

van Rooij and Schikhof, *A Second Course on Real Functions* [vRS82]

This nice book concentrates on a detailed study of differentiation and integration of functions on the real numbers. The material is restricted in scope, but the treatment is quite complete, containing facts, results, and examples otherwise scattered through the literature.

Measure and Probability

These texts are generally very good, and could serve nicely for a first-semester course in measure theory, as well as for a probability course.

Billingsley, *Probability and Measure* [Bil95]

A current standard for this genre. Contains a lot of material on probabilistic number theory as well as general measure theory and, of course, standard probability theory.

Dudley, *Real Analysis and Probability* [Dud02]

An excellent treatment of both measure theory and probability. The two subjects are separated more strictly than in other books of this type, and emphasis is placed on Borel measures on metric spaces, which are especially important in probability. The advanced topics discussed are primarily, but not exclusively, directed towards probability. The writing is clear and readable, and in places rather informal and even chatty, with some important definitions and secondary results buried in passages of text. The writing style somewhat obscures the very careful treatment of the mathematics; I particularly admire the efforts the author made to present the most elegant proofs available of the results. He has a good sense of how to present material in a mathematically honest but non-overwhelming way. There are also extensive and valuable historical notes.

Kingman and Taylor, *Introduction to Measure and Probability* [KT66]

Parthasarathy, *Introduction to Probability and Measure* [Par05]

Geometric Measure Theory

Geometric Measure Theory fairly rapidly splits into two rather distinct parts: "smooth" geometric measure theory and the study of fractals. Some of these references (e.g. Morgan) concentrate on the first, and others (e.g. Edgar) the second. But the basic tools like Hausdorff measures are common to both, and all these references have at least some relevance to both parts.

Edgar, *Measure, Topology, and Fractal Geometry* [Edg08]

An undergraduate-level introduction to fractals and the tools needed to study them. Contains nice but somewhat "light" (focused on the needs of the book) introductions to metric spaces, topological dimension, and measure theory (Lebesgue and Hausdorff measures). A good place to begin reading about the subjects covered.

Edgar, *Integral, Probability, and Fractal Measures* [Edg98]

A "sequel" to *Measure, Topology, and Fractal Geometry*. The title gives a good summary of the topics covered.

Falconer

Federer, *Geometric Measure Theory* [Fed69]

The "bible" of Geometric Measure Theory, with encyclopedic coverage of the subject. The more I study this book, the more I appreciate it; it is a treasure trove of material. At first I found it tough going since the exposition is quite dense (but very efficient). Some of the notation and terminology is nonstandard, and some of the English is rather stilted, but one gets used to these things (and we have an excellent guide in Morgan [Mor09]). And there is more explanation and motivation than I gave credit for at first, along with occasional statements like "I do not know whether this [more general statement] is true" and "I would not know how to prove this [more general statement]." I have found *no* errors or questionable proofs (this doesn't mean there aren't any!), and the only thing I could quibble with is occasional language. I have used this book as primary source for my expositions of various topics. The author is (or was) clearly a master of this subject, and it is pretty much all here.

Mattila, *Geometry of Sets and Measures in Euclidean Spaces* [Mat95]

Morgan, *Geometric Measure Theory* [Mor09]

This is generally a delightful and stimulating book to read, and gives perhaps the most readable introduction to ("smooth") geometric measure theory. It is liberally supplied with computer-generated pictures of objects studied, along with snapshots and anecdotes about the author and other figures in geometric measure theory. Newer editions also have an extensive discussion of soap bubbles and other packing problems. Read this book first (and keep it on hand) before tackling Federer. Note, however, that some arguments represented as proofs in this book are too informal to be considered rigorous.

Rogers, *Hausdorff Measures* [Rog70]

This is an excellent treatment of Hausdorff measures and other basic topics of Geometric Measure Theory. Also discusses Suslin and analytic sets and some Descriptive Set Theory.

Descriptive Set Theory

Bartoszyński and Judah, *Set Theory: On the Structure of the Real Line* [BJ95]

This book more or less picks up where [Oxt71] ends. But the treatment is quite technical, and mostly of interest only to specialists in Set Theory.

Kechris, *Classical Descriptive Set Theory* [Kec95]

The best introductory treatment of Descriptive Set Theory. Covers the basics of the subject thoroughly from many points of view, and makes a convincing case for the importance of the subject and its place in the world of mathematics. More readable than Moschovakis for nonexperts in set theory, although it is more of a complement than a substitute.

Moschovakis, *Descriptive Set Theory* [Mos80]

The standard treatment of Descriptive Set Theory before Kechris. Requires more detailed familiarity with set theory and logic. Well written and readable, however, and a valuable supplement to Kechris.

Foundations, Set Theory, Logic, Large Cardinals

Kanamori, *The Higher Infinite* [Kan03]

This is, and will likely remain (additional volumes are planned), the definitive treatment of the theory of large cardinals. Much of the text is tough going for readers not expert in set theory and mathematical logic, but there are also extensive explanations and motivation for important topics and results, so the book is valuable even for nonspecialists. The order of topics follows the historical development, so the origins of the ideas can be traced. There are also numerous and quite candid descriptions and evaluations of the work of individual mathematicians involved in this subject, which make for interesting reading.

Landau, *Foundations of Analysis* [Lan51]

The first complete treatment of the construction of the real and complex numbers from the Peano axioms (although the various constructions had long since been known and carefully described). The text is rather dry and formal with little explanatory material (although the introduction is engagingly written), but not too much is really necessary and the arguments are easy to follow.

Lawvere and Rosebrugh, *Sets for Mathematics*

This is an appealing and readable treatment of set theory and the foundations of mathematics using categories, and makes a good case that this is the most natural and fruitful approach to foundations for most working mathematicians. However, I find somewhat unconvincing the authors' cavalier assertions that category theory and "naive set theory" (which is what is obtained when the category framework is stripped away) as presented in this way are free from logical contradiction and paradox. For one example, does every category have a skeleton? This is the Axiom of Choice on steroids: the AC is equivalent to the statement that every small category has a skeleton.

Lucas, *The Conceptual Roots of Mathematics* [Luc00]

This is an entertaining and enlightening book, as well as a good source of pithy quotes. The theme is an argument for a "chastened logicism" as the correct philosophy of mathematics; but there is much more of interest, such as one of the best discussions I have read of second-order logic. I know nothing of the author,

but he comes across as something of a character. The quality of both the mathematics and the writing is somewhat uneven, ranging from silly (arguing that Euclidean geometry is "better" than noneuclidean, or invoking stereotypes of Scotsmen and Welshmen to explain some terms) to trenchant.

Mendelson, *Introduction to Mathematical Logic* [Men15]

One of the standard texts on mathematical logic, now in its sixth edition. Gives a very good development of mathematical logic, first-order theories, Gödel's theorems, and various versions of set theory, and has extensive discussion of computability.

However, I got an uneasy feeling of circularity while reading this book (my uneasiness may not be specific to this book but may be with the subject in general, although it came into particular focus with the author's treatment of it). The book begins with a fairly informal introduction in which terms like *set*, *function*, and *countable set* are described. Then the more formal presentation beginning in Chapter 1 gives the axiomatization of a formal theory, starting with a countable set of symbols. Whoa! what exactly is our starting point? Is it set theory (perhaps a naive version)? Is it $\mathbb{N}$? If so, what is the point of eventually axiomatizing $\mathbb{N}$ (or set theory), as is done for $\mathbb{N}$ (in both first- and second-order ways) in Chapter 3 and for set theory (also in multiple ways) in Chapter 4? And if we axiomatize $\mathbb{N}$, don't we have to prove that the $\mathbb{N}$ we have satisfies the axioms (which seems like an impossible task if $\mathbb{N}$ exists without a real definition)? There is also a clear circularity, or at least logical *non sequitur*, when in Chapter 2 alephs and even the Axiom of Choice are used and ultraproducts constructed, before set theory, ordinals, cardinals, and the AC are discussed in Chapter 4. The author's philosophy, as expressed in the introduction, is that "Impeccable definitions have little value at the beginning of the study of a subject. The best way to find out what mathematical logic is about is to start doing it." He does say in a footnote in the introduction that "Care will be exercised to avoid using any ideas or procedures that may lead to the paradoxes ..." but I am not convinced. (My copy of the book is the third edition, but judging from descriptions of the changes in later editions these points do not seem to be treated differently there.)

Nagel and Newman, *Gödel's Proof* [NN01]

A nice introduction to Gödel's Incompleteness Theorem, readable for nonexperts, although it contains a serious misstatement (see H. Putnam's review in *Philos. Sci.* 27(1960), p. 205-207).

Potter, *Set Theory and its Philosophy* [Pot04]

This is an impressive and valuable treatment of set theory which bridges philosophy and mathematics. The mathematical part of the book contains good treatments of the construction of number systems, cardinals, ordinals, and the axiom of choice, as well as set theory itself. Set theory is axiomatized by the theory of levels, but there is also extensive discussion of other systems such as ZF which give theories essentially equivalent for the purposes of providing a foundation for mathematics. Includes extensive historical discussions of the development of the various topics and the philosophy underlying them. Very nicely written, and the philosophical part is varied and balanced. Nonetheless, mathematicians may not agree with all of the philosophical arguments given, and there are a few minor mathematical errors (perhaps no more than in any other book on the list!), such as the formula on p. 232 and accompanying comments.

Sierpiński, *Cardinal and Ordinal Numbers* [Sie65]

An encyclopedic treatment of ordered sets, ordinals, and cardinals. Contains much material about the order structure of the real line. Interesting reading, with many clever arguments and observations. Some of the language used is a little unusual mathematically, and logically ambiguous in places; for example, a

number of results are phrased as "We can prove with the aid of the Axiom of Choice that ...." Apparently this language was used because at the time of the publication of this book, it was not definitely established that these results were independent of ZF set theory (most, if not all, have been subsequently shown to be independent).

Smullyan, *Gödel's Incompleteness Theorems* [Smu92]

A very readable and mathematically complete treatment of various versions of Gödel's Incompleteness Theorem. Perhaps the best source to learn about the details of the theorem.

Nonstandard Analysis

Davis, *Applied Nonstandard Analysis* [Dav77]

Goldblatt, *Lectures on the Hyperreals* [Gol98]

Hurd and Loeb, *An Introduction to Nonstandard Real Analysis* [HL85]

Robinson, *Non-Standard Analysis* [Rob66]

The original treatment of Nonstandard Analysis, and as such of historical and conceptual value. But it is outdated as an exposition of the subject: for example, the hyperreals are constructed using the Compactness Theorem of model theory rather than via ultrafilters (the distinction is more apparent than real, since ultrafilters are used to prove the Compactness Theorem). No longer the best reference for learning about the subject. The last chapter is a discussion of Robinson's interpretation of the history and philosophy of calculus; while interesting and valuable as motivation for his development of Nonstandard Analysis, some of his opinions are questionable at best.

Constructive Analysis

Bishop and Bridges, *Constructive Analysis* [BB85]

This is a "second edition" of Bishop's treatise, largely prepared by Bridges after Bishop's death.

Bridges and Vîţă, *Techniques of Constructive Analysis* [BV06]

A sequel of sorts to [BB85], although it does not presuppose any knowledge about constructive analysis. It includes much material that postdates the publication of [BB85], and is more reflective of the state of the art in this subject. Nicely written, and a good place to learn about some constructive analysis without becoming embroiled in the philosophical controversies underlying it: the authors present the subject in the right way, in my opinion, as an interesting and useful part of mathematics rather than the only true way to do all of mathematics.

The authors' presentation does, however, have some foundational obscurities which must be sorted out. In their approach, a "set" is a collection of objects constructed according to some property, from objects which have "already" been constructed [if a construction is done once, is it permanent, or can it be "lost", e.g. with the death of its constructor?], along with an equivalence relation of "equality" on the objects. In practice the mathematician cannot distinguish between equivalent objects. This approach implicitly presupposes two levels of existence and knowledge: an omniscient level where the objects really exist and where true equality of objects makes sense, and where the equivalence relation can be truly defined, and a lower level

of perception on which human mathematicians operate where objects can only be described or distinguished by constructive arguments. Viewed this way, constructive analysis nicely models the way analysis is actually carried out by mathematicians and especially by computers, and the algorithmic arguments forced by this approach are of obvious importance.

From the point of view I take in the last paragraph, what the authors call the "equality" relation should be called "indistinguishability," and what the authors call "inequality" (constructive proof that two objects are not "equal") should be called "distinguishability." But there are still some logical difficulties: for example, in practice it is not obvious that "indistinguishability" is an equivalence relation, specifically that it is transitive. For (perhaps oversimplified) example, suppose we say we can only distinguish real numbers if they differ by at least 10^{-6} (this is reasonable from a computer architecture point of view, perhaps with a different constant). We can then have three numbers a, b, c with a and b indistinguishable, b and c indistinguishable, but a and c distinguishable. In fact, any type of archimedean property of the real numbers would seem to be incompatible with indistinguishability being transitive; real numbers indistinguishable from 0 would be analogous to infinitesimals in nonstandard analysis (ironically, constructivists, notably including BISHOP [Bis77], are some of the most vocal critics of nonstandard analysis).

Another somewhat jarring feature of the presentation of the book is the rather brief and unmotivated adoption of the set $\mathbb{N}$ as the foundation for all of analysis. The only justification given for this is two quotes from KRONECKER and BISHOP in effect saying that they regarded $\mathbb{N}$ as a gift from God; the authors seem to suggest that if it was good enough for them, it is good enough for us too. This approach is somewhat of an injustice: in fact, there are reasonable arguments in favor of viewing $\mathbb{N}$ as a starting point, such as ones described in [Hey66]. But I still think the worship of $\mathbb{N}$ as the starting point of mathematics is a philosophical weak point of intuitionism/constructivism.

Heyting, *Intuitionism: an Introduction* [Hey66]

Number Theory

Baker, *Transcendental Number Theory* [Bak90]

Perhaps the definitive work on the subject, covering the author's revolutionary contributions. The techniques cover a broad range, from algebra and number theory to geometry and hard analysis. Worth looking at, if only to get a feeling for the character of the subject.

Beiler, *Recreations in the Theory of Numbers* [**?**]

This book is a collection of an astounding supply of facts and problems about the positive integers, of varying depth ranging from high-school level puzzles to ones still unsolved by top mathematicians. It is mind-boggling how many person-hours must have been needed over the centuries to find and prove all these results and formulas. The properties of the natural numbers probably have such a combination of intricate patterns and baffling irregularities as to pique the curiosity and challenge the intellect of mankind for all time. And, as the author says,

> "Contemplating such efforts within the frame of a world troubled by war and ceaseless unrest, it seems that the men [sic] who spent so large a part of their lives in this kind of work must have found the secret to an inner peace and happiness that all men crave but that no exigent clamorer for worldly goods can attain."

This book dates from the mid-1960s and is somewhat out of date, especially with regard to what is known about primes and factoring; the best computers at that time were no more powerful, perhaps less, than cell phones of today (although maybe the comparison is not quite straightforward). However, this does not diminish the book's interest and appeal. I learned at least one new word (*aliquot*), and a couple of easily-remembered largish primes, 11111111111111111111111 (23 1's) and $9999999900000001 = 10^{16} - 10^8 + 1$. (The number consisting of n 1's is denoted R_n and called a *repunit* by the author. It is now known [Underwood 2022, using many hours of computer time] that R_{49081} is prime.)

The author is apparently an engineer, but he evidently has a deep understanding and appreciation of the nature of mathematics, and the book is almost entirely amply solid and rigorous, especially for one intended as a popular treatment. I noticed a few small misstatements, such as the assertion on p. 258 that the continued fraction expansion of an irrational number is periodic (he seems only to have quadratic numbers in mind in this discussion); he also calls 1 a prime on p. 211. His instincts as an engineer show through occasionally: on page 280 he states that "The Arabs knew that it was impossible [i.e. knew that FLT was true for $n = 3$] 700 years before Fermat's day and even possessed a proof, albeit a faulty one." (How did they "know" this if they did not have a valid proof? For that matter, how does one "know" something without possessing a proof at all, even a faulty one? Actually, can a faulty "proof" even be legitimately called a proof?) And on p. 124, after a long discussion of how close the acute angles are to 45° in what he calls the 100th Pythagorean triangle with consecutive legs, whose legs are 77-digit numbers differing by one, he concludes, "We had, therefore, best call the one-hundredth, a 45° isosceles right triangle and let it go at that." Spoken like a true engineer. (If we have a sequence converging to zero, would he say we had best call the terms zero from some point on and let it go at that?)

I had fun reading this nicely written book, although I am not inclined to spend hours working on numerical puzzles (maybe why I'm not a number theorist), and recommend it for recreational reading.

Hardy and Wright, *Introduction to the Theory of Numbers* [HW08]

Khinchin, *Continued Fractions* [Khi97]

This little gem is probably the best introduction to the theory of continued fractions. The first 2/3 of the book is an elementary and very readable introduction to the algebraic theory. The last part is on a considerably more sophisticated level, giving a nice and still very readable treatment of the probabilistic aspects of the theory and full proofs of the author's important results about the probabilistic distribution of terms in continued fractions.

Niven, *Irrational Numbers* [Niv56]

Geometry

Berger, *Geometry I, II* [Ber09]

Originally a single volume in French. A wide-ranging and informative treatment of geometry, including such topics as map projections for the earth.

Blumenthal, *Theory and Applications of Distance Geometry* [Blu70]

This is an interesting, valuable, and rather unique treatment of geometry from the metric space point of view. The subject matter is covered thoroughly, not only metric characterizations of Euclidean space, but also curvature and elliptic and hyperbolic spaces. Some of the notation and terminology is outdated, and some sections and arguments seem a bit laborious and unfocused. Nonetheless, it is the best reference on these topics, and virtually the only one on some of them.

Greenberg, *Euclidean and Non-Euclidean Geometry* [Gre93]

An elementary college-level text on axiomatic geometry, with emphasis on the history and nature of the Parallel Postulate. Very readable, and mostly carefully written, although EUCLID'S postulates are rather mangled (they are eventually written properly, but seemingly only as an afterthought).

Hartshorne, *Geometry: Euclid and Beyond* [Har00]

A careful and comprehensive axiomatic treatment of elementary geometry, both Euclidean and Noneuclidean. Written on a more sophisticated level than Greenberg, but very readable. The connections between geometry and modern algebra are extensively discussed and explained. The author has included nice touches such as hand-drawing most of the diagrams to maintain the spirit of EUCLID.

Moise, *Elementary Geometry from an Advanced Standpoint* [Moi90]

Topology

Topology, as covered in the literature, seems to be essentially two separate subjects: point-set topology and algebraic topology. Few books treat both parts of topology in any detail. Even within each part, standard references vary greatly in which topics are covered in detail, since the subjects are too large to be covered comprehensively in any one reference. Different standard references are better or more complete for different topics, to a much greater extent than in basic real analysis, and it is not feasible in brief reviews to give a full summary of which references are good for which topics (and this is a somewhat subjective matter anyway). Readers are encouraged to explore broadly.

Alexandroff (Aleksandrov) and Hopf, *Topologie I* [AH74]

In German. The first comprehensive book on topology, both point-set and algebraic (1935). I was rather surprised how much of the basics of topology had reached its modern form (some of it for the first time) in this book, since the whole subject had barely begun by the beginning of the twentieth century, although HAUSDORFF'S book [Hau49] must be regarded as more revolutionary. In particular, much of the topology we cover in this book appears here in essentially modern form. Although this is denoted "Volume I", there was never a Volume II or, unfortunately, an English translation.

Borsuk, *Theory of Retracts* [Bor67]

Dugundji, *Topology* [Dug78]

Engelking, *General Topology* [Eng89]

The third iteration of this book. This is the most comprehensive text on general topology, and seems to have become the standard reference, although there is essentially no mention of algebraic topology concepts, even the fundamental group (in fact, homotopy is barely mentioned at all). It is also nicely written, with considerable explanation, motivation, and historical notes. However, it is probably too detailed (and esoteric in many places) to be a good text for an introductory course. Copies are also rather difficult and expensive to come by.

Engelking, *Theory of Dimensions Finite and Infinite* [Eng95]

Revised version of *Dimension Theory* (available online). A modern version of Hurewicz and Wallman, covering the theory of dimension beyond the separable metric case, and the theory of infinite-dimensional spaces. Nicely written, and a good place to learn about dimension theory even for separable metric spaces, the case of most general interest.

Hatcher, *Algebraic Topology* [Hat02]

One of the best modern treatments of classical algebraic topology, and freely available online to boot. Very clearly written, with a highly readable exposition of most important basic aspects of algebraic topology. Perhaps the best source for learning the fundamentals of the subject, and a model for what I hope will someday be available online in all parts of mathematics, a goal to which I hope my own book will make a contribution. The author promises additional volumes, which I sincerely hope to see.

Hausdorff, *Grundzüge der Mengenlehre [Fundamentals of Set Theory]* [Hau49]

This was a path-breaking book (1914), and still merits study. The first half of the book is an exposition of set theory, ordered sets, ordinals, and cardinals; while it is undoubtedly a better treatment of these subjects than was available previously, this part does not really break new ground. But the second half of the book can be fairly regarded as the beginning of the modern subject of point-set topology, although the notion of an abstract topological space was first described by FRÉCHET in 1906. It contains the first modern definition of a metric space (modeled on previous work of FRÉCHET as well as Hausdorff himself), and essentially the modern definition of a topological space (there had also been previous work on this, notably by HILBERT). A quite remarkable percentage of the topology we cover in this book appears here. While the topology of the real numbers and Euclidean space was already well understood (primarily via convergence of sequences and series and continuity of functions), the transition to abstract spaces as we view them today sprang forth rather suddenly in amazingly detailed and polished form in this book.

Later editions of this book (e.g. [Hau44]), and the English translations (e.g. [Hau57]), are considerably abbreviated, leaving out much of the text of the first edition including most of the topology (but with a more refined exposition of what remains), with added material on Descriptive Set Theory. So the later editions are really a different book whose title more accurately describes the contents. This book is mainly of historical interest today, especially the first edition.

Hurewicz and Wallman, *Dimension Theory* [HW41]

Kelley, *General Topology* [Kel75]

Kunen and Vaughan, ed., *Handbook of Set-Theoretic Topology* [KV84]

The title of this collection of articles should be taken seriously (at least the "set-theoretic topology" part; the "handbook" part is pretty misleading – this is a massive 1273-page work which would qualify as a handbook only for the likes of Paul Bunyan, although I note that several "handbooks" of similar size on other subjects are also published by Oxford University Press, so maybe the British have bigger hands). Set-theoretic topology is not quite the same thing as point-set topology; the topology treated in this book is based on the intricacies of set theory, including set-theoretic axioms and large cardinal hypotheses. Almost all the results discussed are either trivial or inapplicable for metrizable spaces, and many of them even for spaces which are separable and/or first countable. The articles seem to be carefully written and comprehensive, but will be of interest primarily to specialists in this field, who would be characterized more as set theorists than mainstream topologists. Even the articles on Borel measures, Banach spaces, and topological groups, which should be of interest to a functional analyst such as myself, have a heavy dose of set-theoretic content, although I found parts of each to be of interest. Besides these, the only article which evinced more than a passing interest from me is P. NYIKOS's discussion of nonmetrizable manifolds, and these still strike me as being mostly a curiosity. Several other articles do have preliminary sections which are well-written overviews of parts of the subject for nonexperts.

Kuratowski, *Topology I, II* [Kur66], [Kur68]

Massey, *Algebraic Topology* [Mas77], [Mas91]

Munkres, *Algebraic Topology* [Mun84]

Munkres, *Differential Topology* [Mun66]

Munkres, *Topology* [Mun75]

Seifert and Threlfall, *A Textbook of Topology* [ST80]

An early (1934) treatment of algebraic topology. Nicely written and still very readable, and of the early topology texts it is the only one which is still worthwhile studying for content instead of just historical value. Parts are rather outdated today; the English translation, published in 1980, describes some but not all of these (and even an exposition current in 1980 is now a bit dated). The small amount of point-set topology covered is particularly rudimentary and outdated. A few things in the exposition are at least misleading if not inaccurate:

> In §52, an argument is given that purports to show that every finitely generated group occurs as the fundamental group of a closed 4-manifold (homogeneous 4-complex). But the argument () clearly only works as written for finitely *presented* groups, a much smaller class; and in fact the fundamental group of a closed manifold (and, more generally, a compact metrizable ANR, cf. ()), is always finitely presented. Actually the statement in the book is somewhat ambiguous since their complexes are not required to be finite; the question of whether every finitely generated group, or more generally any countable group, occurs as the fundamental group of a (possibly noncompact) 4-manifold appears to be up in the air, but the answer is positive for open subsets of $\mathbb{R}^5$; cf. http://math.stackexchange.com/questions/1362655/fundamental-group-of-open-subsets-of-mathbbrn.
>
> The definition of "manifold" in the book is what is now called a *combinatorial homology manifold*, which is significantly different from a topological manifold: a combinatorial homology manifold is compact, with no boundary, triangulable, but not necessarily homogeneous. Thus only closed manifolds are combinatorial homology manifolds, and only those which can be triangulated, but on the other hand there are combinatorial homology manifolds which are not manifolds, such as the suspension of the Poincaré homology 3-sphere ().

Spanier, *Algebraic Topology* [Spa81]

My immediate impression of this book when I first picked it up as a student was that it is quite heavy for its size. When I started reading it, this impression persisted in a metaphorical sense: it is pretty dense and detailed. But it is also pretty readable, and it was a standard reference and text at that time for good reason. It may not still be state of the art, but it is still a worthwhile reference.

Willard, *General Topology* [Wil04]

Numerical Analysis

Acton, *Numerical Methods That (Usually) Work* [Act90]

A lively and informative book. The contents could be uncharitably described as a large bag of tricks, but the book is really much more than that: it is a detailed discussion of the art of computation (with emphasis on the *art*), meaning finding, adapting, and using appropriate numerical procedures for computational problems. The author was an engineer, with a fairly typical engineering perspective on mathematics, but informed by extensive experience in computation, and there is much wisdom and common (even uncommon) sense here. His writing style is refreshing and somewhat self-effacing; although he was an acknowledged expert in computation, he never talks down to the reader. For example, he states, "I have tried to make my explanations clear, but sad experience has shown that you will not really understand what I am talking

about until you have made some of the same mistakes that I have made." There is not much attention to mathematical rigor, and even the mathematical details of standard procedures are not discussed in depth, but there are beautiful heuristic descriptions of many procedures. This book is a must-read for anyone preparing to work in computation.

Hamming, *Numerical Methods for Scientists and Engineers* [Ham86]

As a nonexpert in numerical analysis, I found this to be one of the most valuable books for understanding the philosophy of the subject. While many of the nuts and bolts of numerical analysis are treated in detail, and in a way I mostly found mathematically satisfactory, the book's main strength is extensive explanations of *why* things are done as they are, from the point of view of the user. Pure (and even applied) mathematicians may be occasionally offended by sarcastic comments of the author (his colleague Brian Kernighan has fondly called Hamming a "classic curmudgeon"[10]). For example, he describes the difference between Numerical Analysis and Numerical Methods thus:

> "Numerical analysis seems to be the study in depth of a few, somewhat arbitrarily selected, topics and is carried out in a formal mathematical way devoid of relevance to the real world. Numerical methods, on the other hand, try to meet the need for methods to cope with the potentially infinite variety of problems that can arise in practice. The methods given are generally chosen for their wide applicability in creating formulas and algorithms as well as for the particular result being found at that point."

(In fact, I disagree with much of the author's philosophy of mathematics as described in [Ham98], although I appreciate his point of view.) But from this book one really gets a feeling for how engineers and other consumers of numerical analysis view the subject and how their needs are accommodated. The author's stated motto is "The purpose of computing is insight, not numbers," and he delivers on this goal.

Whittaker and Robinson, *The Calculus of Observations* [WR67]

A book reflecting the state of the art in numerical analysis about a hundred years ago. Interesting to peruse, if for no other reason than to marvel at the extensive computations mathematicians of yesteryear carried out by hand (and the elaborate techniques developed to accomplish them) which few today would attempt. Nor would they usually need to with the availability of computers; but the ability (and willingness) to do such hand computations may be somewhat of a lost art, and may have even given these mathematicians insights not easily obtainable today (although, to be sure, we get other insights from computers which they could not).

Miscellaneous Topics

Wagon, *The Banach-Tarski Paradox* [Wag93]

This is one of the best mathematics books I have read. It covers an enormous range of fascinating topics in a very engaging way, and has succeeded in giving a quite readable treatment of some deep topics without pulling punches. Not all of the topics covered are directly related to measure theory, but many are, sometimes in surprising ways. This book really gives the reader a feeling for the way many diverse parts of mathematics fit together and interact.

[10] *Princeton Alumni Weekly*, 1/11/17.

History of Mathematics

Ball, *A Short Account of the History of Mathematics* [Bal60]

This book, first published in 1891, is badly out of date and inconsistent with modern historical scholarship in many parts; thus its modern-day value as a serious historical work is limited. It is still interesting reading, however, with many personal tidbits which are generally more reliable than those in BELL'S books.

Bell, *The Development of Mathematics* [Bel92]

This is a more solid and serious book than *Men of Mathematics*, and has been generally praised by historians of mathematics, although the early chapters have faced some criticism. It is written in an engaging style, with many interesting insights and strong opinions. But it should not be the only book one reads about the history of mathematics.

Boyer, *A History of Mathematics* [Boy89]

Eves, *An Introduction to the History of Mathematics* [Eve90]

Graham and Kantor, *Naming Infinity* [GK09]

The thesis of this book is that "Name Worshipping," a sect of the Russian Orthodox Church, played a significant role in the mathematical careers of DMITRI EGOROV and NIKOLAI LUZIN, and in the development of Descriptive Set Theory by the Moscow school. Part of the book is a detailed account of the lives of these mathematicians and their colleagues and students, which is one of the most detailed studies yet of the important group in Moscow commonly called "Lusitania." Other parts describe the practice of Name Worshipping and attempt to make the connection with the mathematical philosophies of EGOROV and, especially, LUZIN.

I found the book somewhat disappointing until I reached the last chapter, where the authors put things together nicely; this chapter is a glimpse of what the earlier part of the book could have been. Unfortunately, it seems that the authors, who are themselves certainly well versed in mathematics, shied away from depth and technicalities in most of the book in a misguided effort to make it accessible and interesting to non-mathematicians, and thus made their arguments less convincing to mathematicians. I found many of their arguments in the earlier part of the book strained and unconvincing. For example, they often seem to argue that due to the Name Worshipping philosophy, LUZIN *et al.* conflated naming mathematical objects with establishing their existence. No one with a significant background in mathematics could accept this, and it is clear from a quote of LUZIN in the Appendix that this was not what he thought (the authors also explain in the last chapter that they did not believe this argument or intend to make it). Several analogies proposed such as the comparison with GROTHENDIECK are quite strained: GROTHENDIECK'S talent described in the book (only one of his many talents) was his ability to come up with *good* names for mathematical objects which would inform and inspire himself and others; in the philosophy described in the book, where simply naming or explicitly describing mathematical objects was the crucial matter, the quality of the names chosen would seem to be irrelevant. And although the book draws extensive comparisons between the mathematics in Moscow and developments in France during the same period and immediately preceding, I was surprised that the religious influences on the work and philosophy of CANTOR, who got the whole subject of set theory going, were almost totally ignored (it seems clear from other sources that CANTOR was indeed significantly influenced by Christian theology and/or Jewish traditions).

This is overall an interesting book to read, but it could have been a lot better.

Grattan-Guinness, *From the Calculus to Set Theory, 1630-1910* [Gra00]

Gray, *Plato's Ghost: The Modernist Transformation of Mathematics* [Gra08]

This is a massive and thorough study of the revolution in the nature and conception of mathematics which took place primarily in the late nineteenth and early twentieth centuries, and represents a formidable piece of scholarship which is also beautifully and entertainingly written.

The book ends rather with a whimper than a bang, however (maybe I was just tired out when I got this far). The last chapter consists of speculations about the philosophy of mathematics, and ends with an assessment of whether the modernist approach "won." The author tries to claim that there are areas of mathematics where anti-modernism still reigns, such as the subject of differential equations (ODE and PDE). But I find this argument unconvincing. Before the modernist revolution, mathematics consisted almost exclusively of the study of objects mathematicians regard as "concrete", at that time numbers, Euclidean space, and well-behaved functions on them. Mathematics today consists of an entire spectrum of study of objects ranging from concrete to extremely abstract ones defined axiomatically, with no clear dividing line. Some parts of today's mathematics continue to study the traditional concrete objects, e.g. number theory and complex analysis; differential equations and differential geometry can be regarded as the principal remaining parts of traditional real analysis with an active research component (one might add probability). In addition, we now have an ever-increasing collection of things regarded by at least some mathematicians as specific concrete objects, e.g. manifolds, well-ordered sets, Hilbert space. But even the subjects superficially unchanged from 200 years ago, such as number theory and differential equations, use a range of techniques far expanded from the traditional ones which could be developed or even contemplated only as a result of the modernist revolution. For example, the modern study of differential equations heavily uses methods from functional analysis, listed by the author as one of the shining examples of modernist mathematics. (Of course, traditional techniques also continue to be used; after all, even calculus as known and understood in the eighteenth century is still extremely important and useful.) The subject of differential equations also heavily uses sophisticated modeling and simulation techniques made possible by the existence of powerful computers, a development not anticipated even at the end of the modernist revolution; the present use of computation in mathematics may mark another revolution which will turn out to be as profound as the modernist one.

Although parts of this book are dense and a challenge to slog through, it is highly recommended.

Jahnke (Ed.), *A History of Analysis* [Jah03]

Kline, *Mathematical Thought from Ancient to Modern Times* [Kli72]

This is Kline's *magnum opus* and principal scholarly work on the history of mathematics; he has a number of other books written more at the popular level (which are somewhat repetitious and not always as well researched). In this, as well as his other books, there is an undercurrent of Kline's philosophy of mathematics which I personally take some issue with, although this does not seriously undercut the value of these books as historical works. There are also small doses of sexism here and there in Kline's books.

Moore, *Zermelo's Axiom of Choice* [Moo82]

A careful and thorough study of the Axiom of Choice – its prehistory, history, controversy, variations, and intimate connections with the very nature of modern mathematics.

Pier, *Mathematical Analysis During the 20th Century* [Pie01]

The author has set himself a monumental and perhaps impossible task in trying to summarize and organize the development of analysis during the twentieth century, and he has partially succeeded, perhaps as well as any one individual could in a book of this length published at the immediate end of the period studied (he is well aware of the inherent problems of such a project, and indeed divides his description of each part of analysis into "evolution" covering the period 1900-1950 and "flashes" covering 1950-2000). But due to the sheer volume and scope of the material, the coverage is necessarily rather shallow, and the author's judgment about what the most important developments were is highly subjective. For example, in the area I personally know best, operator algebras, he hits some of the high points of the subject (with occasionally muddled description) but misses others, instead discussing some developments at length which are generally regarded as peripheral, such as the axioms for a C*-algebra; I doubt that any leading operator algebra specialist would regard the author's treatment as an adequate or satisfactory history of the subject of operator algebras during the 20th century. (To be fair, he seems to have done a more accurate and thorough job on some other parts of analysis he may be more personally familiar with.)

One of the most interesting parts of the book is the introduction, where the definition, meaning, and scope of the term "analysis" is discussed. The author then takes a fairly broad interpretation of its meaning, including subjects from topology to algebraic geometry under the umbrella of 20th century "analysis."

Much of the material is presented via extensive quotations from the writing of leading mathematicians. These make for interesting and thought-provoking reading. The book is very helpful in giving the reader a feeling for the overall flavor of twentieth-century analysis, although it is far from a definitive study.

Roberts, *King of Infinite Space* [Rob06]

A biography of Donald Coxeter (the title is an apparent allusion to *Hamlet*). This was a more engaging book than I expected. It is nicely written, and the author has really done her homework, with extensive and trenchant quotes from many leading mathematicians she apparently interviewed. The mathematics is pretty much right, both in detail and philosophical form, especially impressive from an author who does not claim any particular expertise in mathematics; it is conveyed more accurately than by some highly-trained authors who should know better. The general theme of the book, which I think is slightly (but only slightly) overblown, is that classical geometry was a dying subject kept alive primarily by Coxeter before finding a renaissance in such modern areas as computer graphics. In fact, classical geometry had been declining in importance since the eighteenth century, when analytic methods began to supplant geometric ones.

A subplot which is greatly oversimplified and rather inaccurate, in my opinion, is that there was a war between formal and visual mathematics in the twentieth century, which is roughly characterized as Bourbaki vs. Coxeter. Bourbaki was indeed a champion of a formal, logical approach to mathematics, but it would be totally wrong to claim that the members of Bourbaki (even Dieudonné, portrayed as the principal villain) did not approve of use of pictures, diagrams, and other visual aids in understanding and creating mathematics; they just felt that the ultimate logical justification for results which may have been found by such methods must be done formally, a sentiment with which even Coxeter would have largely agreed. And it is quite another matter that Bourbaki chose to leave pictures and diagrams almost completely out of their exposition of mathematics, which was never intended to supplant all other approaches. (And Bourbaki was hardly the first to do this: in his *Méchanique analytique*, Lagrange bragged:

> "No diagrams will be found in this work. The methods which I expound in it demand neither constructions nor geometrical or mechanical reasonings, but solely algebraic [analytic] operations subjected to a uniform and regular procedure."

(cf. [Kli72, p. 615]).) Blaming BOURBAKI for the failures of the "new math" is about as misguided as blaming JESUS for present-day fundamentalist Christianity. A new exposition of part of mathematics can *never* be a negative, except *possibly* if it is mathematically incorrect (but see the review of [**?**]); it can only help a competent reader, in combination with other expositions, gain a better understanding of the subject. The problem comes when some people who lack the vision and expertise of the authors decide that a particular work is gospel and should be followed (often with a twisted interpretation due to a lack of understanding) to the exclusion of all others, and a fashion (which mathematics is certainly not immune from) is established. This was the principal problem with the "new math," and even (as the author points out) with the mathematics curriculum at some leading universities, although it is questionable whether all these problems were even directly a result of the influence of BOURBAKI. In any event, it is nice to see that visual mathematics is making a comeback, and while COXETER was probably not as single-handedly responsible as this book portrays, he was certainly a major influence.

Struik, *A Concise History of Mathematics* [Str87]

Philosophy of Mathematics

Some of the books in other sections above are partially concerned with the philosophy of mathematics.

Chaitin, *Information-Theoretic Incompleteness* [Cha92] and *The Unknowable* [Cha99]

Entertainingly written accounts of Chaitin's philosophy of mathematics and knowledge, focusing on complexity. The presentation is unusual, intermixing autobiographical passages with the mathematics. It is easy to dismiss the autobiographical parts as egotistical self-promotion (e.g. "Gödel discovered incompleteness, Turing discovered uncomputability, and I discovered randomness"), but, although I have no personal knowledge about the author, I see a certain naiveté about the stories that suggests such a conclusion may be unfair. The mathematical part is heavy stuff, and I do not feel comfortable with my own understanding of it, but it is certainly a lot of food for thought.

Corfield, *Towards a Philosophy of Real Mathematics* [Cor03]

This is not the first criticism of the preoccupation of philosophers of mathematics with foundations at the expense of studying its actual practice, but it is the most detailed and articulate one to date. The author's knowledge and understanding of mathematics is clearly much broader and deeper than that of most other philosophers of mathematics, and is quite impressive. He makes a compelling case against traditional philosophy of mathematics, but I have a hard time figuring out exactly what he proposes as a replacement: he says he is arguing for a Bayesian theory, but spends a good bit of time pointing out the absurdities of a strict Bayesian interpretation of mathematics and science (e.g. that existing evidence cannot help confirm the plausibility of a new theory that explains it). In any event, as evidenced by his choice of title, he clearly thinks that answering the questions he raises is a work in progress. Although this cannot be considered the definitive work on the subject, it is a significant and thought-provoking contribution to the process.

Lakatos, *Proofs and Refutations* [Lak76]

A thought-provoking discussion of Lakatos' ideas on the nature of mathematical proof, and by transfer his whole philosophy of mathematics. Written primarily as a dialog based around EULER'S formula for the relation of the number of vertices, edges, and faces of a polyhedron ($V-E+F=2$). Interesting reading even for those, like me, who are somewhat skeptical of his conclusions. See [] and [Cor03] for detailed analyses of Lakatos' work.

Introduction to Proofs

I rarely spend time looking at books at this level, so I do not really know what is available. I imagine there are quite a few others, since there is a market for them as textbooks.

Hammack, *Book of Proof* [**?**]

Online. When my department created an Introduction to Proofs class, this was adopted as the text. I taught the course once, and I found this to be a reasonable text, although with some serious objections. I would use it again, especially since it has been improved since, and is freely available online. (I object to making students buy textbooks at inflated prices, especially ones which are regularly revised not to improve the text but to deliberately make used copies of previous editions worthless. These comments will undoubtedly not ingratiate me to textbook authors and publishers, but I don't care.) As I do with most texts I use in class, I assembled an extensive list of errata and comments, which I gave to my students and sent to the author. I was pleased to see that the next revision of the text incorporated most of my comments; I like to think this was a result of my list, but this may or may not be the case (anyway, the author never acknowledged my communication).

Houston, *How to Think Like a Mathematician* [**?**]

I recently found this at a local used book store for $1, and I am quite impressed by it. It is not exactly an Introduction to Proofs text, but advice on how to read, write, and do mathematics (it is called "A Companion to Undergraduate Mathematics"). It is very nicely written and contains a lot of solid advice for an aspiring young mathematician. Much of this advice echoes my own advice scattered through this manuscript, but more extensive and better organized together. (The similarity is neither coincidence nor plagiarism: most mathematicians would probably give similar advice if they took the time and effort to write it.) I found only a few small statements to quibble with. In a footnote, it is stated that a theorem is "a mathematical statement which is true." No: a theorem is a mathematical statement which has a proof, which is not quite the same thing (see GÖDEL). The description of the Diagonalization Argument on p. 226-7 is not quite correct: it must be assumed that the decimals in the list are written without an infinite string of 9s (otherwise the second number in the list could be $.24999\cdots$ and the generated number could be $.25000\cdots$, for example). In Theorem 27.15, EUCLID never said there were infinitely many primes (to be fair, the author never explicitly said he did); the proof, and EUCLID'S statement, just says there is no largest prime. And on p. 262, under "How to prove that something is unique," it is an overuse of contradiction to assume one has two distinct objects and then prove they are the same; it is simpler and more direct to just assume we have elements x and y and show directly they are the same (to be sure, this would often be proved by contradiction). Also, Example 8.7(iii) is wrong.

Nontechnical and Popular Books

Barrow, *Pi in the Sky* [**?**]

Another book about mathematics written by an astrophysicist (cf. LIVIO). This book can best be described as "very uneven": parts are very cogent and thought-provoking, with some deep insights, while others are way off base. The author's primary focus is understandably the use of mathematics as the basis for science, and how the philosophy of mathematics relates to the structure of the universe (in a broad sense). After an introduction and chapters on the origins of the number system and numerology, which are interesting but seemingly off the main subject (although the idea is that they represent the origins of the philosophies of mathematics), the bulk of the book is a discussion of intuitionism, formalism, and Platonism,

largely from the point of view of how these philosophies impact the relationship of mathematics to science, especially physics. The discussion of each of the philosophies is interesting but flawed, in some respects fatally flawed.

The discussion of intuitionism is quite flawed. The author conflates at least four distinct notions:

1. Intuitionism.
2. Constructivism.
3. n-Valued logic, $n \geq 3$.
4. Rejection of *reductio ad absurdum.*

Various intuitionists/constructivists believe some, but not all, of these, sometimes in more nuanced versions. See () and () for more comments on this. In his discussion of multivalued logics, he (correctly) credits Lukasiewicz, Post, and Tarski, leaving the impression that they were intuitionists. I have seen no evidence that this was the case; at least Tarski, with whom I was acquainted as a student, was not an intuitionist – if anything, he was a logicist, although, like most mathematicians, he was hard to categorize.

In the discussion of formalism, the author seems to conflate formalism and logicism, which may seem superficially similar but are quite distinct. Russell was a logicist, and certainly not a formalist, and definitely did not believe mathematics is meaningless. Hilbert also did not believe mathematics is meaningless and did not espouse the crude ("neutered") version of formalism that mathematics is just manipulation of symbols according to postulated rules, and neither, I think it is safe to say, did any member of Bourbaki, even Dieudonné. In fact, it is a stretch to even call Dieudonné a formalist; consider the quote in () and his very geometric approach to Analysis. I will not repeat here my comments about Bourbaki in the review of *King of Infinite Space*, but a formal approach to mathematics is not the same as formalism: saying that formal mathematical structures are "meaningless" is not at all the same thing as saying mathematics is meaningless. In fact, the quote from Bourbaki on p. 134 belies that they believed mathematics is meaningless.

Some of the misstatements in the book are just silly. For example, the author repeatedly states that Euclid's fifth postulate is that "parallel lines do not intersect." No: this is Euclid's *definition* of parallel lines; his fifth postulate has to do with existence and uniqueness of parallel lines (and not really even that: see I.6.1.48. and I.6.1.54.).

The author states on p. - that "Cantor was not Jewish." This is actually a somewhat complex matter which largely depends on one's definition of "Jewish," and a flat statement is not warranted or even helpful. Cantor was not a practicing Jew, at least openly. His religious beliefs appear to be Christian (whether Protestant or Catholic is unclear). His father was a devout Lutheran, of Jewish ancestry; his mother was from a Catholic family, likely also of Jewish ancestry, at least partially. There are indications that some Jewish traditions persisted in the family, and of course there is Cantor's use of Hebrew letters as mathematical symbols. One assessment I read somewhere concluded that Cantor was "barely a Jew." But Hitler would have considered him a Jew.

The heart of the problems of the book is the inadequate and misleading discussion of logic in its various forms. The only distinction the author makes in any detail is between "classical" (two-valued) and "intuitionist" (multivalued) logic. But mathematical logic, even "ordinary" two-valued mathematical logic, is not the same as Aristotelian logic, which is not precise enough to form a basis for mathematics. (There is also a "metalogic," essentially Aristotelian logic, which is used in the metalanguage, as the author does point out.) But both two-valued logic and multivalued logic come in different forms. Multivalued logic is off the mathematical mainstream (and it is inaccurate to equate it to intuitionist logic, which also comes in various forms). The biggest distinction which must be made is among two-valued logics: first-order logic,

used almost exclusively by logicians, and second-order logic, which seems to be used by most mainstream mathematicians. There is a whiff of discussion in the book of second-order logic, in the footnote on p. 121, but it is only a whiff and GENTZEN'S result is not really second-order logic.

GÖDEL'S Incompleteness Theorem is a theorem in first-order logic and there are grave difficulties in extending or interpreting it in second-order logic. Second-order Peano arithmetic is (presumably) consistent, finitely axiomatized, contains first-order Peano arithmetic, and is categorical and thus complete in a sense (). This is a crucial point about the Incompleteness Theorem which is overlooked by the author (and practically all other authors writing about the Theorem). This greatly limits the applicability of GÖDEL'S theorem, and many of the sweeping philosophical consequences claimed, in this book and elsewhere, are unwarranted.

While on the subject of the Incompleteness Theorem, there is also the often-overlooked hypothesis that the theory be "finitely describable" (recursively axiomatizable). This author, to his credit, does know that there is a hypothesis of this sort, but only mentions it inadequately in passing. He also gets it somewhat wrong: recursively axiomatizable theories need not be finitely axiomatizable, and even first-order Peano arithmetic is not finitely axiomatizable; the discussion of axiom schemes on p. 126 is incorrect.

So Gödel Incompleteness does not (directly) apply to second-order logic. But second-order logic has a different kind of incompleteness since there is no second-order mechanical proof theory (). Thus logicians regard second-order logic as an unsatisfactory foundation for mathematics, although mathematicians seem to freely use it (even the Completeness Axiom of the real numbers is second-order). I think a very interesting book along the lines of the one being reviewed could explore the implications of using second-order logic in mathematics and its relation to the physical world. But this book is not it.

Bellos, *The Grapes of Math* [**?**]

Although I am not impressed by the cutesy title (the title may not be the author's doing – the British version has the title *Alex Through the Looking Glass*), this is an interesting and fun book to read. It is not quite nontechnical, but does not presuppose any knowledge for the reader beyond some high-school algebra. Many interesting topics are discussed, in a clear, entertaining, and mostly mathematically satisfactory way. The examples in the next-to-last chapter on proofs are strikingly similar to ones in my book (to be clear, almost all of my versions were written and online before this book was published; it's almost as if the author read my manuscript, although probably not since it is not mentioned and he omits some difficulties I discuss).

Bellos, *Here's Looking at Euclid* [**?**]

An earlier book by the same author with a cutesy title (the British version is titled *Alex's Adventures in Numberland*). It is on a slightly more elementary level than *The Grapes of Math*. The author is an entertaining writer and does his homework, and there is enough here to make this book an interesting read even for professionals.

Part of this book is a discussion of how the human mind understands mathematics. Although it is a popular book and not intended as a scholarly publication, and although the author has no advanced training in either mathematics or psychology, I find the discussion (few if any of the ideas are original with the author) more intelligent and insightful than [LN00].

Most of the mathematics discussed is elementary. There is almost nothing to quibble with, although occasionally something is "dumbed down" to the point of making the discussion at least somewhat misleading. For example, on p. 107 he attempts to define transcendental numbers without mentioning the word *polynomial*, saying "a transcendental number is an irrational that cannot be described by an equation with a finite amount of terms." He adds that they are "so mysterious and evasive that finite mathematics could not capture them." But equations like $2^x = 3$ or $\sin x = 1$ have a finite amount of terms and transcendental

solutions (even $\sum_{k=0}^{\infty} \frac{1}{k!}$ is arguably a finite expression). His discussion of infinite series also only applies to nonnegative series, an excusable simplification in a book of this sort, although it may merit a comment.

He lionizes EUCLID (slightly) too much: on p. 56 he writes, "*The Elements* is a magnum opus of pedantry and rigor. Nothing is ever assumed. Every line follows rigorously from the line before." This is largely, but not entirely true; see the discussion in (). Indeed, compare the "proof" on that page with (); how do you know the circles intersect? (On the back cover, EUCLID is called "the greatest mathematician of all time," but this was undoubtedly written by a publicist and not the author.) And the discussion of the "Parallel Postulate" on p. 268-9 is flawed (cf. ()): WALLIS'S version is equivalent to Playfair's Axiom, the usual statement of the Parallel Postulate, and not to the version at the bottom of p. 268, and arguably not equivalent to EUCLID'S fifth postulate.

On p. 248, he refers to GALILEO measuring the "distances to stars." I doubt GALILEO ever tried to do this with any star but the sun, or got any reasonable answers if he did. People at that time, even geniuses like GALILEO, had no conception of how far away stars are or any way of estimating it. He was likely trying to measure the distance to planets, for which he had at least a hope of reasonably succeeding. (Even reasonably estimating the distance to the moon would have been a challenge at that time.)

Enough nitpicking. This is a good book.

Cheng, *How to Bake* π [Che15]

This is a generally very nicely-done book, intending to give a comfortable introduction to real mathematics (particularly category theory, the author's specialty) for a lay audience. A subtheme is a discussion of baking, one of the author's interests, and perceptive analogies are made between baking and mathematics. The author is very good at putting mathematical concepts in a form which should be understandable to someone without much mathematical background. However, the desire to be as nontechnical as possible does lead to some misstatements, or at least misleading statements, for a mathematically literate reader. The worst problem I found was in the discussion of topology on p. 90-95, where the concepts of *homeomorphism* and *homotopy equivalence* are confused (of course I mean the discussion confuses them, not that the author does): the equivalence described somewhat informally appears to be homeomorphism, but the examples, e.g. the capital letters of the alphabet, are of homotopy equivalence ("number of holes"). The discussion is continued on p. 174-175, somewhat more carefully but not quite adequately. A more minor misstatement is on p. 120, where some properties of addition on $\mathbb{Z}$ are listed, and it is asserted that these listed properties are the axioms for a group. But one needs also that $a+0=a$ for any a, which does not follow from the listed properties (one could define $+$ on $\mathbb{Z}$ by setting $a+b=b$ for any a and b). This is carefully and correctly done later on p. 133-136. The book concludes with very insightful discussions of knowledge vs. understanding and the place of mathematics in general society.

Elwes, *Mathematics 1001* []

This is an interesting compilation of "1001 bite-sized explanations" of "absolutely everything that matters in mathematics" (or maybe at least everything the author knows). The explanations, roughly a paragraph long, are of uneven quality, but mostly quite good, and are entertainingly written. The weakest chapter seems to be Analysis (maybe I'm just more attuned to noticing objectionable things here): difficulties range from outright misstatements, for example claiming (under "analytic continuation") that any analytic function on a piece of the plane extends to an analytic function on the whole plane, an amateurish mistake, to explanations not technically wrong but misleading or excessively shallow. For example, under "implicit differentiation" (one of my pet topics: I often use the treatment of the Implicit Function Theorem as an early benchmark in judging the quality of any basic analysis or general-purpose book) he first points out that $\frac{dy}{dx}$ can be found from an equation $y = f(x)$ by the techniques of calculus already discussed, and then states, "However, there is no law that says we have to begin with y alone on one side of the equation." Well, OK, there is no law that says we have to restrict ourselves to the hypotheses of any theorem; but if you want to assert the conclusion of the theorem in a situation where the hypotheses are not satisfied, you'd better be prepared to justify it, and quoting the theorem is of no help in this. In this case, there is no further hint that this situation is not a straightforward extension of ordinary calculus. On the same page is a fallacious "proof" of the Chain Rule (cf. V.3.10.5.). Another misleading explanation from the Algebra chapter is an incorrect statement of the cubic formula (the signs must be chosen opposite in the two terms). And he uses some nonstandard terminology (at least for the US; maybe it's different in the UK), such as "gradient" instead of "slope" of a line. The chapters on logic and metamathematics are the most interesting (to me) and appear to be the best, not surprising since these are the areas of the author's primary expertise. This book is interesting and entertaining to read, but with a bit of caution.

Klein, *Elementary Mathematics from an Advanced Standpoint: Arithmetic, Algebra, Analysis* [Kle04]

This is not intended to be a popular book, and is not exactly nontechnical. Its goal was twofold: to describe Klein's ideas on how mathematics should be presented and taught in schools, and to give schoolteachers knowledge and insight on the true nature of the mathematics they teach (he says on p. 162: "... the teacher's knowledge should be far greater than that which he presents to his pupils. He must be familiar with the cliffs and the whirlpools in order to guide his pupils safely past them.") Although it was written more than a hundred years ago, most of it is as relevant today as when it first appeared, and it is full of valuable insights even today from a first-rate mathematical mind which paid unusually great attention to how mathematics should be taught. I myself learned some things about mathematics and its development, such as a beautiful proof of Taylor's theorem with Lagrange remainder via difference calculus, regarded by Klein as the best and most natural proof of the theorem (and the way TAYLOR himself found the theorem, although he did not come close to giving what would be today recognized as a proof), which I have not found in modern texts, probably because difference calculus is rarely discussed except in numerical analysis texts. See () for the argument. And HILBERT in his *Grundlagen der Geometrie* described something close to the nonstandard real numbers (they may have even been found earlier by someone else). This book is well worth reading and studying by any potential mathematician or mathematics teacher. The reader is cautioned that the English translation in this edition contains an unusually large number of misprints, most of which are harmless.

Livio, *Is God a Mathematician?* [Liv09]

This is generally a well-written, informative, and entertaining study of the "unreasonable effectiveness of mathematics" as a basis for science and the underlying principles of the universe. There is little actual religious content, mainly just a discussion of the religious beliefs of major figures such as GALILEO, DESCARTES,

and NEWTON. The author is an astrophysicist and not a mathematician, and his primary interest (at least in this book) is the relation between mathematics and science; but he has deep knowledge and understanding of mathematics and gets it almost all right. There are a few misstatements, such as that "a fully consistent theory of analysis was only formulated in the 1960s" (an apparent reference to nonstandard analysis), and the crucial words *finitely describable* (technically, *recursively axiomatizable*) are omitted from the statements of Gödel's theorems (he is hardly the first author to make this mistake, and in his case it may well be an inadvertent error rather than a misunderstanding). His discussion of the philosophy of mathematics is extensive and intelligently argued, although he repeatedly makes rather unconvincing arguments against Platonism, and he rather too glowingly refers to [LN00], which in my opinion is hardly more than a work of fiction. He buys a little too much into the "crisis" characterization of the modernist revolution precipitated by the discovery of noneuclidean geometry, but his discussion of this subject is more accurate and intelligent than many others. He does not pretend to have the answers to the main questions he discusses (no one else does either), but makes worthwhile contributions to the discussion.

Maor, *e: the Story of a Number* [**?**]

A wide-ranging account of some parts of the history of mathematics, loosely based around the number e. The author is a historian of mathematics and knows his stuff. I found the book almost entirely mathematically accurate and honest, as well as being engagingly written, although I am skeptical of the author's claim that it is accessible to a general audience; I think at least a one-semester calculus course is pretty much necessary to understand and appreciate the book. I found very little to quibble with in the text. The only mathematical error I noticed was on p. 179, where the author asserts that if a complex function is differentiable at a point, then it is conformal there (see XII.1.1.14. for the correct statement). On p. 181 he says the three great achievements of nineteenth-century mathematics are complex analysis, abstract algebra, and noneuclidean geometry. Crucial advances all, especially if noneuclidean geometry includes Riemannian geometry; but I would say that set theory *á la* Cantor and the foundations of (real) analysis are just as important. I also question his assertion that most texts on calculus and analysis treat the calculus of the logarithm function before the exponential, suggesting (but never really claiming) that his approach of doing the exponential function first is somehow novel. I have always done the exponential function first in my calculus classes, in the same way he does it, and the textbooks I have used have almost always done the same. If one draws the graphs of 2^x and 3^x, it is geometrically very plausible that there is a b between 2 and 3 for which the tangent line to the graph of b^x at $(0, 1)$ has slope 1. Anyway, these are minor points and do not detract from the book.

Pickover, *The Math Book* []

Each summer I spend time at our family cottage on a lake in Ontario. I often check out popular math books from the local library (they have a better collection than one might expect) to read on the island while not doing more serious things like swimming and boating. This is one of those books.

This lavishly illustrated book, organized chronologically, gives one-page treatments of topics the author considers interesting and important, ranging from prehistoric to 2007. On the page facing each piece of text is a color illustration, many of which are visually stunning and some of which are even directly connected with the topic discussed. The topics range from deep and important subjects in mathematics to simple puzzles (although some of the puzzles turn out to be not so simple and do require some nontrivial mathematics to be developed to solve them). The author is not a mathematician (his degree is in biochemistry), but he is fairly knowledgable about mathematics and gets it mostly, but not entirely, right. The history is also not entirely right, and some of it seems to be straight out of BELL'S *Men of Mathematics*. He does correct some of the

problems and misconceptions in the notes at the end of the book: for example, after a text attributing sole responsibility for "Stirling's formula" to Stirling, in the notes he acknowledges it is largely (in fact almost entirely) due to de Moivre. Many, perhaps too many, of the photos are computer graphics of fractals and other mathematical art, one of the author's interests – he has been significantly involved in producing such images. This book is worth checking out if you spend a week on an island in Canada.

Rucker, *Infinity and the Mind* [Ruc82]

In my opinion, the most interesting of Rucker's books. Contains accurate and readable, although non-rigorous, treatments of the theory of ordinals and of Gödel's Incompleteness Theorem.

Smullyan, *The Lady or the Tiger?* [**?**]

The first half of this book is a collection of Smullyan's famous logic puzzles. The second half is a delightful treatment of Gödel's Incompleteness Theorem in the form of a "mathematical novel."

Smullyan, *Forever Undecided* [**?**]

This book falls somewhere between *The Lady or the Tiger?* and *Gödel's Incompleteness Theorems*. It is not a technical book, although it covers some symbolic logic, but it is also not nearly as lighthearted as Smullyan's other popular books. The topics discussed and explained are somewhere between philosophy and mathematics, based on various versions of Gödel's Incompleteness Theorem.

The two books *The Lady or the Tiger?* and *Forever Undecided* are available together in a single paperback volume.

Stillwell, *The Story of Proof* [Sti22]

This nice book does not cleanly fit into either the History of Mathematics or Popular Books category. It is an account of how the concepts of mathematics and their proofs developed, but is more about the substance of mathematics than its history. Although some parts, especially the earlier ones, can be understood at the popular level, the depth and sophistication of the mathematics increases significantly through the book (not a criticism at all!), and most of it is only understandable to someone with a significant background in mathematics (it would be accessible to graduate students or advanced undergraduates in mathematics).

Many of the arguments presented are clever and insightful, and some were even new to me; and they are accompanied by very nice illustrations and graphics. My only criticism is that it is not pointed out that many of the "proofs," especially earlier ones, are not rigorous or complete by modern standards. For example, some arguments concern tangent lines, which are not defined; the author also talks frequently about arc length and surface area without definitions, suggesting they have an intrinsic self-evident meaning. But even Archimedes understood that surface area is not self-evident and had to be defined, which is not easy to do carefully. Even carefully defining the area of a plane figure or volume of a solid takes some work. A few added comments here and there would remedy these shortcomings.

Books to be Wary of

Alexander, *Infinitesimal: How a Dangerous Mathematical Theory Shaped the Modern World* [Ale15]

The theme of this book is that the mathematical "theory of indivisibles" precipitated an intense battle in the sixteenth and seventeenth centuries, not just among mathematicians but even among political and social theorists. Throughout the book the author tries to characterize the opposition to the theory of indivisibles as a manifestation of a rigid, authoritarian social philosophy, and the belief in the theory of

indivisibles as representing ideals of free inquiry and, ultimately, democracy. While this may be rather a stretch, oversimplification, or extrapolation, it certainly has an element of truth, and the historical events are engagingly described in considerable detail in support of the thesis.

I saw and browsed this book in a bookstore and it looked interesting, so I got a copy and began reading it systematically. I could hardly get past the introduction, since it was so full of mathematical misconceptions. The most egregious, affecting the entire thesis of the book, is confusion of the mathematical terms *indivisible* and *infinitesimal* (which mean entirely different things – whatever an infinitesimal is, it is not indivisible! If it were, what could an equation like $dy = \frac{1}{2}\,dx$ possibly mean? Granted, before the late seventeenth century the difference was not well understood even by mathematicians. See [She18] for a detailed study.) In fact, the book probably should have been titled *Indivisible*, since this term is the one causing most of the controversy discussed in the book. (Debates about the existence of indivisibles and the nature of the continuum go back to ancient Greece, and were revived in Europe at least as early as the fourteenth century; cf. e.g. [Boy59].) The only mention of infinitesimals as mathematicians since Archimedes have almost always thought of them is a few oblique references to the development of calculus, which really has nothing to do with the theory of indivisibles; the only relevance to the book's thesis is that (early) calculus was based on controversial (although not actually new) ideas which were not rigorously established.

But there are other serious mathematical inadequacies, such as the attempt to describe what the theory of indivisibles means (it essentially means that a line is a set of points, or, what is not exactly the same thing, that a line is "made up" of a collection of points, and similarly a plane figure is "made up" of a collection of parallel line segments, etc., not that a line segment is actually indivisible). The author is a historian, not a mathematician, and it is perhaps unrealistic to expect him to get the mathematics right, but many authors of other books about mathematics have done better with no more formal mathematical training. Perhaps he was simply presenting the idea as it was (mis)understood by people like Renaissance Jesuits as well as many mathematicians of the time, but if so he should have said so and pointed out the misconception. And he never mentioned the resolution of the paradoxes discussed, not found until the twentieth century using measure theory and the difference between countably and uncountably infinite sets (see I.3.1.2. ff. for a discussion). The author seems to subscribe to the common picture of a line as a string of beads, each point having a "next" point, a picture every modern mathematician has had to try to correct in the classroom. This picture was common at one time even among mathematicians: Galileo believed that a line segment was made up of infinitely many infinitely small points separated by infinitely many infinitely small gaps.

Pushing on, I found the rest of the book more interesting, with much detailed history I didn't know which I assume is accurately portrayed since this should be the author's forte. The book traces the formation of the Jesuit order of the Catholic Church, largely as a response to the Reformation and resulting loss of influence and control of the Church, and the incorporation of mathematics into the curriculum of the Jesuit colleges and, more generally, as a subject as worthy of study as philosophy and theology. Mathematics as taught by the Jesuits was in a form compatible with overall Jesuit philosophy, as a body of eternal truths and not as an intellectual pursuit in which new ideas were sought or even tolerated. As the author puts it on p. 78:

> "But why study mathematics? Only because it provided a model of perfect rational order and certainty, and an example of how universal truths governed the world. If mathematics was to become a field of far-reaching innovation, in which new truths were proposed and then subjected to challenge and debate, then it would be worse than useless. It would be dangerous, as it would compromise the very foundations of the truth it was meant to buttress."

The Jesuits had enormous influence on the intellectual life of continental Europe in the sixteenth and seventeenth centuries, partially through their colleges, which were in some respects the best system of higher education at the time.

A good deal of the book is devoted to Thomas Hobbes and John Wallis. The mathematical part of their longstanding intellectual battle (there was a considerable nonmathematical component too, see e.g. [**?**]) is described as being mainly a disagreement about mathematical philosophy. While there is probably some truth to this, I think there is a more important reason. The author does point out that Hobbes's (later) mathematical work was not taken seriously by the English mathematical establishment, but attributes this to political prejudice. In fact, Hobbes was a hopeless mathematical charlatan, and his mathematical work was no more worthy of being taken seriously than that of modern-day cranks (cf. Figure I.13). Wallis was an excellent mathematician, one of the world's leading mathematicians of the seventeenth century (although overshadowed in Britain by Newton and Gregory), and very likely had the attitude widespread among mathematicians of not suffering fools gladly (see e.g. the quote in II.1.). The author may perhaps again be excused for not having the ability to make a judgment about the quality of Hobbes's mathematics. The statement on p. 213 that Hobbes was "one of the most respected English mathematicians of the day" [1640] is doubtful (although English mathematics was in a rather primitive state at the time); I am not aware of a single (correct) original contribution he made to mathematics.

The author's assertion that Wallis created a new kind of experimental mathematics, and especially that he did so to conform to the philosophy of the Royal Society, is nonsense. All mathematics is experimental when first being developed, and mathematical results have been found by induction since ancient Greece and even earlier. Wallis used techniques going back to Archimedes and probably even Euclid's predecessors (ancient Greeks apparently did not record how they obtained their results, and in any event knowledge about their methods has been lost), and in fact paid considerable homage to Greek mathematics and could be considered the leading historian of mathematics of his era. Wallis may not have paid sufficient attention to carefully proving his results, but he undoubtedly could have (and probably did mentally) in many cases using the elaborate double *reductio ad absurdum* technique of Archimedes, but evidently was more interested in using his time and effort to explore new things. Also note that it was not uncommon at that time for mathematicians to announce results without giving a hint of a proof, often in the form of an implicit challenge to other mathematicians (in effect saying, "I can prove this, can you?") Wallis was surely aware, as contemporaries like Newton were, that some of the techniques did not (at that time) have a sound logical footing, but these mathematicians had strong intuition and their results invariably turned out to be correct and logically justifiable (unlike Hobbes). It seems farfetched that Wallis declined to try to prove his results simply in order to adhere to the Royal Society philosophy. The author seems to be obsessively insistent on attributing the primary motivation for every mathematical or scientific study to adherence to one side or the other of the philosophical dichotomy which is the theme of the book (Maslow's hammer at work!)

Incidentally, Wallis did not initiate the idea that mathematical objects exist and only need to be explored, and that mathematical truths are discovered; this philosophy of mathematics is called *Platonism* since it goes back (at least) to Plato. (Hobbes was also a Platonist and, contrary to what is suggested in the book, Euclid probably was too.)

There is at least one historical inaccuracy I noticed. On p. 290, it is stated:

> "Following the war over the infinitely small, advanced mathematics in Italy came to a standstill, whereas English mathematics quickly became one of the dominant national traditions in Europe, rivaled only by the French."

This may be an accurate assessment about Italy. But although England in many respects, including technology, industrialization, and certain intellectual and social areas, became a world leader, it notably remained a mathematical backwater for more than 100 years after Newton, particularly in analysis (other than mathematical physics). In fact, there was a nineteenth century analysis renaissance in Italy before there was one

in England. It was French, Swiss, and German mathematicians, not English, who were at the forefront of analysis in the eighteenth and nineteenth centuries (I count LAGRANGE as French although he was actually born in Italy). The only British analyst of note during this period was MACLAURIN (who was Scottish).

There have been criticisms of the reasons the author attributes for the Jesuit opposition to the theory of indivisibles, that it is inconsistent with ARISTOTLE and EUCLID and the notion that mathematics represents eternal truth. In [She18] (which I find an articulate criticism of much of Alexander's book, although I cannot evaluate the historical aspects), it is argued that the principal reason for the opposition was that it contradicted the Catholic doctrine of transubstantiation of Christ's body and blood in the Eucharist (this idea was first presented in [Red87]). Note that I only discovered and read these other references after I had written most of this review.

It is remarkable that so many people, both mathematicians and nonmathematicians, spent so much time and effort and generated such acrimony over a dispute of such little ultimate (mathematical) consequence, with both sides advancing arguments which today would be regarded as foolishly misguided. I wish the author had been able to convey the obsolescence better. The book is worth reading for its historical content, as long as the author's attributions and conclusions are taken with a degree of skepticism.

Bell, *Men of Mathematics* [**?**]

This book is entertaining reading, but unreliable as history; some parts are downright fanciful. The chapters on GALOIS and CANTOR are the worst from the standpoint of historical accuracy. Although Bell was a respectable mathematician, some critics have asserted that as a biographer he seems to have been the type of writer who never lets the facts get in the way of a good story. Overall, however, the book is not nearly as bad as these critics have made it out to be: the vast majority of the book is factually accurate, and the errors appear to be entirely the result of careless or inadequate research and unwarranted extrapolation rather than deliberate falsification; the accounts are much closer to the actual events than in many modern "historical" movies. In fact, the author has gone to some pains to qualify historical assertions he felt were not adequately established. Despite the title, the book does not have a sexist tone, and in fact was fairly progressive for its time (1937), giving sympathetic treatment to several women mathematicians and the obstacles they faced; for example, Bell gave the first popular (if dated and flawed) account of SOFIA KOVALEVSKAYA. There is some gratuitous antisemitism in the chapter on CANTOR (there is controversy over to what degree, if any, CANTOR himself was Jewish, although he clearly had at least some Jewish ancestry). This book may be inspiring to a young reader who is aware of its historical inaccuracies, and it also serves a useful purpose in countering the stereotype of mathematicians as nerds of borderline sanity and little personality or variety. The description of Bell in [New88] seems fairly apt: "lively, stimulating, inoffensively crotchety and opinionated, with a good sense of historical circumstance, a fine impatience with humbug, a sound grasp of the entire mathematical scene, and a gift for clear and orderly explanation."

Griffiths and Hilton, *A Comprehensive Textbook of Classical Mathematics* [GH78]

This book is hard to categorize. The aim is to give a unified exposition of "classical mathematics" (or what may be loosely described as "undergraduate mathematics") from a modern perspective: logic and set theory, abstract algebra, elementary geometry and topology, and undergraduate analysis. Some things (generally simpler ones) are proved in detail, but most of the deeper or more difficult results are just stated with some informal or intuitive justification given (which is clearly represented as such; the book is pretty honest mathematically). I would not describe the treatment as inspired, and it is generally pretty conventional, but there is much wisdom and insight in the descriptions. And, importantly, the authors explicitly describe, and follow, what they call the *spiral principle*, which underlies both the historical development of mathematics

and the way it is, or should be, taught: fundamental topics and principles are first introduced and developed in a natural and intuitive way, and after they have become somewhat familiar are redeveloped in a more careful way. This is nicely summarized in this statement from p. 249, in the chapter "The Logic of Geometry", after Euclid's geometry (with modern improvements) is described:

> "A person might object to all this procedure on the grounds that the definitions given are artificial without a prior knowledge of Euclidean geometry; to which we can only reply that *once* they are given, a logically watertight system of geometry can be developed, and it matters not a whit how the definitions happened to be suggested. This is typical of the way in which a great deal of mathematics has grown. First a terminology and body of results is obtained by logically unsatisfactory (but often aesthetically pleasing) methods. Then the theory is seen to have flaws which prevent further progress. Finally the theory is reset in a possibly different language, and derived in a correct logical manner, usually gaining greater power."

On the other hand, the book contains surprisingly many (especially for an apparently revised second edition) missteps, most of which are fairly minor and many of little real consequence; I think at least nearly all of them could be described as "careless." For example, both the statement of the Implicit Function Theorem (37.4.2) and Theorem 15.5.2 are false as stated; additional hypotheses are needed (continuity of the relevant derivatives suffices). In contrast, the statement of the Mean Value Theorem, or more accurately of "Rolle's conditions" on p. 483 (ROLLE never stated, and probably never considered, such conditions), contains a superfluous continuity assumption on the derivative, which is not used in the proof or mentioned when the theorem is quoted in several other places. It is stated that the definition of "finite set" in 7.2.4 (what is usually called *Dedekind-finite*) is equivalent to the one in 7.1.4 (apparently meaning 2.11.1), but some form of the Axiom of Choice, which is only mentioned later, is needed. (Theorem 7.1.4 also has an obvious superfluous hypothesis.) And the authors seem to like to invent their own terminology, which is sometimes at odds with convention. Some of this invented terminology is picturesque, such as their name "mostest theorem" for what is normally called the Extreme Value Theorem, but some is not an improvement, such as "flat" for "affine subspace" or "laws of indices" for "rules of exponents" (maybe some of their terms are in more general use in Britain than in the U.S.). And what they call the "Axiom of Choice" (that an element can be chosen out of one nonempty set) is either trivial or a nonmathematical assertion depending on one's point of view (cf. II.6.1.6.); what is usually called the Axiom of Choice is called the "Strong Axiom of Choice".

This book could be quite helpful to an advanced undergraduate or knowledgeable nonmathematician in helping to "put the pieces together", somewhat in the manner of [Sch97] (although the books are very different), as long as it is recognized that the small inaccuracies mean that the details of the exposition must be regarded with a degree of skepticism. But, although this book is described and clearly intended as a textbook, I find it hard to think it could be suitable as a textbook for a course it would be worthwhile for a mathematics student to take; it might be suitable for a "capstone" course for students preparing to be math teachers at the high-school or community college level.

Guillen, *Bridges to Infinity* [Gui83]

This book is not entirely about mathematics, and I will not pass judgment on the nonmathematical part. But it is so riddled with mathematical errors and misconceptions that the credibility of the entire work is questionable. There are also sloppy historical errors, such as calling VON NEUMANN a "German mathematician."

Jackson (ed.), *Mathematics: an Illustrated History of Numbers* [**?**]

This book is representative of a genre of lavishly produced and marketed "popular" books about mathematics for mass audiences. It has received glowing reviews in the mainstream press (e.g. *The Booklist, Washington Post*) written by reviewers who were not competent to evaluate it.

The book is divided into 100 units ("ponderables"), many of which are actually physics. At first glance it is appealing, but one soon finds it full of misstatements, misconceptions, and outright nonsense (both mathematical and historical) in places. This is now apparently a third revised edition (I can only imagine what was revised and corrected since so many problems remain). For example, in the unit on e, the number is first described:

> "e arises from the construction of the so-called natural logarithm, which defines a new number – a fixed-value 'base' the natural logarithm of which is 1. The natural logarithm of any given number x is the power to which this base must be raised to equal x. In other words, $\ln(e^x) = x$ and $\ln(e) = 1$, since $e^1 = e$. This logarithmic function is natural in mathematical terms since it is defined around oneness. As you can see, e is the public face of the exponential function, e^x, where $x = 1$."

In fact, nearly every page seems to have a mathematical howler. A few other samples:

> $\sqrt{2}$ is made up of an infinite string of numbers and therefore has no exact value.
>
> In at least two places, it is stated that Euclid's axioms were "inconsistent."
>
> The three-body problem shows that the universe is indeterminate.
>
> Russell's Paradox was a death blow to Cantor's set theory.
>
> The real numbers are uncountable because between any two integers there are infinitely many of them.
>
> Unlike straight lines, geodesics are closed, and geodesics were used to build the geodesic dome.

Not everything in this book is nonsense, far from it; but there is so much that it takes a mathematically literate reader to separate the wheat from the chaff. To be perhaps overly charitable to the authors (none of whom are mathematicians, although this is really neither here nor there), some of the nonsense may be just an overzealous attempt to "dumb down" sophisticated mathematical ideas for a lay audience. But the book is not suitable even for one of its stated target audiences, young people interested in mathematics, since too many misconceptions could be ingrained from it.

Lakoff and Nuñez, *Where Mathematics Comes From* [LN00]

This book purports to be a study of how the human mind perceives and understands mathematics. But the authors, a linguist and a cognitive psychologist, repeatedly demonstrate an embarrassing lack of understanding of basic undergraduate mathematics (e.g. limits); they would be well advised to study a book like *Closer and Closer* to see what they are missing. They regard the equation $e^{i\pi} = -1$ as "one of the deepest equations in all of mathematics," and suggest that even most modern mathematicians do not understand it. There is also much discussion of the philosophy of mathematics, which is shallow and logically flawed, sometimes with a vicious undertone. While I cannot evaluate the authors' cognitive psychology arguments, this is not a book that mathematicians should take very seriously (how can they expound intelligently about how humans understand mathematics if they don't understand it themselves?) See http://www.bruceblackadar.com for a more extensive review.

Polya, *Induction and Analogy in Mathematics* [Pól90a] and *Patterns of Plausible Inference* [Pól90b]

This two-volume work is in many ways an admirable book. It is a thorough discussion of mathematical reasoning and problem solving by one of the masters, and it is filled with solid and valuable advice on how to approach and solve mathematical problems. The techniques are illustrated by numerous examples, containing clever and insightful arguments (originally dating back as far as ancient Greece) from which much can be learned. And he does draw and explain an important distinction between *plausible reasoning* and *demonstrative reasoning*.

But the treatment of demonstrative reasoning is seriously flawed: the book is unfortunately riddled with what I call "mathematical dishonesty." Many of the arguments represented as proofs are logically incomplete or nonrigorous (see I.6.1.7. for a discussion of some of these; earlier, he seems to represent the principle of induction as so self-evident that it merits no comment.) This is a much more serious problem in a book of this sort than in books like [Doo94] or [Mor09] (where there are gaps in arguments but no attempt is made to hide them, and the gaps are generally routine to fill in), since the main subject of this book is mathematical reasoning. Polya himself certainly knew better. Some rather minor changes in the language used, and a few added comments, could have maintained mathematical honesty while still making the main points the author wanted to convey. But he seemed to deliberately go out of his way to obscure or gloss over some of the mathematical difficulties in his arguments, and as a result many readers could be misled about the nature of full demonstrations in mathematics.

Seife, *Zero: The Biography of a Dangerous Idea* [Sei00]

Who knew mathematics was so dangerous (cf. [Ale15])? This book is a popular account of the history of zero in a broad sense, including limits and singularities; zero and infinity are treated together as the limiting cases of the positive real numbers, and "reciprocals" of each other. The "danger" described is mostly really discomfort; the only danger (other than true physical danger to a few people for expressing heretical ideas) is difficulties and absurdities caused by trying to divide by zero or carelessly add infinite series. The last third of the book is mostly physics, describing singularities like black holes and the big bang and the attempts to reconcile relativity and quantum mechanics. The author is a clear and entertaining writer, with enough background in mathematics that the actual mathematics is mostly described accurately, with occasional language imprecision (e.g. he repeatedly refers to sums like

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots$$

as having "infinite terms" rather than an infinite *number* of terms), but his arguments and conclusions are often sensationalized: the introduction and acceptance of zero as a number was a slow and difficult process which certainly caused discomfort, but it is quite a stretch to call zero "dangerous." There is a lot of interesting history, which is for the most part well researched but not entirely accurate. Comparing with [Ale15], both books suffer from overly broad and simplistic interpretations and characterizations (sometimes in conflict between the two books, although the overlap in subject matter is not great), this book somewhat more so, although the mathematics is described more accurately. This is an interesting book to read so long as not everything is taken at face value.

Nonmathematical Books

Okrent, *Highly Irregular* [Okr21]

This is a delightfully written and engagingly scholarly account of how English came to be such an irregular language. It explains the origin of the spelling and grammatical irregularities, as well as some

common expressions which seem to defy logic. Well worth reading for both native and non-native English speakers.

A Stalin-era joke from the Soviet Union:[11]

> "Two militiamen are invited to Comrade Kozlov's birthday party. What should we give him? asks one. The best gift is a book, says the other. No, says the first militiaman. Comrade Kozlov already has a book."

[11] Elena Gorokhova, *A Mountain of Crumbs*, p. 7.

Bibliography

[Aar20] Scott Aaronson. The busy beaver frontier. *ACM SIGACT News*, 51(3):32–54, 2020.

[AB68] R. D. Anderson and R. H. Bing. A complete elementary proof that Hilbert space is homeomorphic to the countable infinite product of lines. *Bull. Amer. Math. Soc.*, 74:771–792, 1968.

[AC59] R. D. Anderson and Gustave Choquet. A plane continuum no two of whose nondegenerate subcontinua are homeomorphic: An application of inverse limits. *Proc. Amer. Math. Soc.*, 10:347–353, 1959.

[Act90] Forman S. Acton. *Numerical methods that work*. Mathematical Association of America, Washington, DC, 1990. Corrected reprint of the 1970 edition.

[Add59] J. W. Addison. Separation principles in the hierarchies of classical and effective descriptive set theory. *Fund. Math.*, 46:123–135, 1959.

[AF03] Robert A. Adams and John J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[AH74] P. Alexandroff and H. Hopf. *Topologie. I.* Springer-Verlag, Berlin, 1974. Berichtigter Reprint, Die Grundlehren der mathematischen Wissenschaften, Band 45.

[Ahl78] Lars V. Ahlfors. *Complex analysis*. McGraw-Hill Book Co., New York, third edition, 1978. An introduction to the theory of analytic functions of one complex variable, International Series in Pure and Applied Mathematics.

[AK88] William Aspray and Philip Kitcher, editors. *History and philosophy of modern mathematics*, Minnesota Studies in the Philosophy of Science, XI, Minneapolis, MN, 1988. University of Minnesota Press.

[Akh88] N. I. Akhiezer. *The calculus of variations*. Harwood Academic Publishers, Chur, 1988. Translated from the Russian by Michael E. Alferieff.

[Ale47] P. Alexandroff. On the dimension of normal spaces. *Proc. Roy. Soc. London*, 189:11–39, 1947.

[Ale15] Amir Alexander. *Infinitesimal—how a dangerous mathematical theory shaped the modern world*. Scientific American/Farrar, Straus and Giroux, New York, 2015.

[Alf71] Erik M. Alfsen. *Compact convex sets and boundary integrals*. Springer-Verlag, New York-Heidelberg, 1971. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 57.

[AMR88] R. Abraham, J. E. Marsden, and T. Ratiu. *Manifolds, tensor analysis, and applications*, volume 75 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1988.

[AMS73] Piotr Antosik, Jan Mikusiński, and Roman Sikorski. *Theory of distributions. The sequential approach*. Elsevier Scientific Publishing Co., Amsterdam; PWN—Polish Scientific Publishers, Warsaw, 1973.

[And58] R. D. Anderson. A characterization of the universal curve and a proof of its homogeneity. *Ann. of Math. (2)*, 67:313–324, 1958.

[And66] R. D. Anderson. Hilbert space is homeomorphic to the countable infinite product of lines. *Bull. Amer. Math. Soc.*, 72:515–519, 1966.

[AP73] P. S. Aleksandrov and B. A. Pasynkov. *Vvedenie v teoriyu razmernosti: Vvedenie v teoriyu topologicheskikh prostranstv i obshchuyu teoriyu razmernosti*. Izdat. "Nauka", Moscow, 1973.

[Are46a] Richard Arens. Topologies for homeomorphism groups. *Amer. J. Math.*, 68:593–610, 1946.

[Are46b] Richard F. Arens. A topology for spaces of transformations. *Ann. of Math. (2)*, 47:480–495, 1946.

[Art64] Emil Artin. *The gamma function*. Translated by Michael Butler. Athena Series: Selected Topics in Mathematics. Holt, Rinehart and Winston, New York, 1964.

[AS60] Lars V. Ahlfors and Leo Sario. *Riemann surfaces*. Princeton Mathematical Series, No. 26. Princeton University Press, Princeton, N.J., 1960.

[Ass79] Patrice Assouad. étude d'une dimension métrique liée à la possibilité de plongements dans $\mathbf{R}^n$. *C. R. Acad. Sci. Paris Sér. A-B*, 288(15):A731–A734, 1979.

[Atk53] F. V. Atkinson. On relatively regular operators. *Acta Sci. Math. Szeged*, 15:38–56, 1953.

[AU23] P. Alexandroff and P. Urysohn. Une condition necessaire et suffisante pour qu'une classe $(\mathscr{L})$ soit une classe $(\mathscr{D})$. *C. R. Acad. Sci. Paris*, 177:1274–1277, 1923.

[AW98] Herbert Alexander and John Wermer. *Several complex variables and Banach algebras*, volume 35 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 1998.

[AW07] Kathryn B. Andrist and David G. Wright. The Whitehead manifold has no orientation reversing homeomorphism. *Topology Proc.*, 31(1):1–5, 2007.

[Axl95] Sheldon Axler. Down with determinants! *Amer. Math. Monthly*, 102(2):139–154, 1995.

[Axl97] Sheldon Axler. *Linear algebra done right*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, second edition, 1997.

[AZ10] Martin Aigner and Günter M. Ziegler. *Proofs from The Book*. Springer-Verlag, Berlin, fourth edition, 2010.

[Bak90] Alan Baker. *Transcendental number theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, second edition, 1990.

[Bal60] W. W. Rouse Ball. *A short account of the history of mathematics*. Dover Publications Inc., New York, 1960.

[Bar00] Edward J. Barbeau. *Mathematical fallacies, flaws, and flimflam*. MAA Spectrum. Mathematical Association of America, Washington, DC, 2000.

[Bat93] S. M. Bates. Toward a precise smoothness hypothesis in Sard's theorem. *Proc. Amer. Math. Soc.*, 117(1):279–283, 1993.

[Bau61] Heinz Bauer. Šilovscher Rand und Dirichletsches Problem. *Ann. Inst. Fourier Grenoble*, 11:89–136, XIV, 1961.

[BB85] Errett Bishop and Douglas Bridges. *Constructive analysis*, volume 279 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1985.

[BB92] J. M. Borwein and P. B. Borwein. Strange series and high precision fraud. *Amer. Math. Monthly*, 99(7):622–640, 1992.

[BB08] Jonathan Borwein and David Bailey. *Mathematics by experiment*. A K Peters Ltd., Wellesley, MA, second edition, 2008. Plausible reasoning in the 21st Century.

[BBG04] Jonathan Borwein, David Bailey, and Roland Girgensohn. *Experimentation in mathematics*. A K Peters Ltd., Natick, MA, 2004. Computational paths to discovery.

[BBI01] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001.

[BD09] Jonathan Borwein and Keith Devlin. *The computer as crucible*. A K Peters Ltd., Wellesley, MA, 2009. An introduction to experimental mathematics.

[BdL59] Errett Bishop and Karel de Leeuw. The representations of linear functionals by measures on sets of extreme points. *Ann. Inst. Fourier. Grenoble*, 9:305–331, 1959.

[Bec71] Petr Beckmann. *A history of π (pi)*. The Golem Press, Boulder, Colo., second edition, 1971.

[Bel77] J. L. Bell. *Boolean-valued models and independence proofs in set theory*. Oxford Logic Guides. Clarendon Press, Oxford, 1977. With a foreword by Dana Scott.

[Bel87] E. T. Bell. *Mathematics, Queen and servant of science*. MAA Spectrum. Mathematical Association of America, Washington, DC, 1987. Reprint of the 1951 original, With a foreword by Martin Gardner.

[Bel92] E. T. Bell. *The development of mathematics*. Dover Publications Inc., New York, second edition, 1992.

[Ber07] Felix Bernstein. Zur Theorie der trigonometrischen Reihe. *J. Reine Angew. Math.*, 132:270–278, 1907.

[Ber72] J. Bertrand. *Calcul des probabilités*. Chelsea Publishing Co., Bronx, N.Y., 1972. Réimpression de la deuxième édition de 1907.

[Ber03] Marcel Berger. *A panoramic view of Riemannian geometry*. Springer-Verlag, Berlin, 2003.

[Ber09] Marcel Berger. *Geometry I–II*. Universitext. Springer-Verlag, Berlin, 2009. Translated from the 1977 French original by M. Cole and S. Levy, Fourth printing of the 1987 English translation [MR0882541].

[Bes24] A. S. Besicovitch. über analytische funktionen mit vorgeschriebenen werten ihrer ableitungen. *Mathematische Zeitschrift*, 21:111–118, 1924.

[Bes45] A. S. Besicovitch. A general form of the covering principle and relative differentiation of additive functions. *Proc. Cambridge Philos. Soc.*, 41:103–110, 1945.

[Bes52] A. S. Besicovitch. On existence of subsets of finite measure of sets of infinite measure. *Indag. Math.*, 14:339–344, 1952. Nederl. Akad. Wetensch. Proc. Ser. A **55**.

[BG80] Richard L. Bishop and Samuel I. Goldberg. *Tensor analysis on manifolds*. Dover Publications, Inc., New York, 1980. Corrected reprint of the 1968 original.

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[Bin48] R. H. Bing. A homogeneous indecomposable plane continuum. *Duke Math. J.*, 15:729–742, 1948.

[Bin49] R. H. Bing. A convex metric for a locally connected continuum. *Bull. Amer. Math. Soc.*, 55:812–819, 1949.

[Bin51a] R. H. Bing. Concerning hereditarily indecomposable continua. *Pacific J. Math.*, 1:43–51, 1951.

[Bin51b] R. H. Bing. Higher-dimensional hereditarily indecomposable continua. *Trans. Amer. Math. Soc.*, 71:267–273, 1951.

[Bin64] R. H. Bing. Some aspects of the topology of 3-manifolds related to the Poincaré conjecture. In *Lectures on modern mathematics, Vol. II*, pages 93–128. Wiley, New York, 1964.

[Bir37] Garrett Birkhoff. Moore-Smith convergence in general topology. *Ann. of Math. (2)*, 38(1):39–56, 1937.

[Bis77] Errett Bishop. Book Review: Elementary calculus. *Bull. Amer. Math. Soc.*, 83(2):205–208, 1977.

[BJ35] R. P. Boas Jr. A theorem on analytic functions of a real variable. *Bull. Amer. Math. Soc.*, 41(4):233–236, 1935.

[BJ59] R. H. Bing and F. B. Jones. Another homogeneous plane continuum. *Trans. Amer. Math. Soc.*, 90:171–192, 1959.

[BJ95] Tomek Bartoszyński and Haim Judah. *Set theory*. A K Peters Ltd., Wellesley, MA, 1995. On the structure of the real line.

[BK87] Steven R. Bell and Steven G. Krantz. Smoothness to the boundary of conformal maps. *Rocky Mountain J. Math.*, 17(1):23–40, 1987.

[Bla] Bruce Blackadar. A family of mutually nonhomeomorphic separable contractible 2-manifolds of cardinality $2^{2^{\aleph_0}}$. http://arxiv.org/abs/1505.01441.

[Bla77] Charles E. Blair. The Baire category theorem implies the principle of dependent choices. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.*, 25(10):933–934, 1977.

[Bla83] Bruce E. Blackadar. The regular representation of local affine motion groups. *Pacific J. Math.*, 108(2):265–274, 1983.

[Bla97] David Blatner. *The joy of π*. Walker and Company, New York, 1997.

[Bla06] B. Blackadar. *Operator algebras*, volume 122 of *Encyclopaedia of Mathematical Sciences*. Springer-Verlag, Berlin, 2006. Theory of C^*-algebras and von Neumann algebras, Operator Algebras and Non-commutative Geometry, III. Revised and corrected manuscript available at http://wolfweb.unr.edu/homepage/bruceb/.

[Bli46] Gilbert A. Bliss. *Lectures on the Calculus of Variations*. University of Chicago Press, Chicago, Ill., 1946.

[Blo97] Ethan D. Bloch. *A first course in geometric topology and differential geometry*. Birkhäuser Boston, Inc., Boston, MA, 1997.

[Blu70] Leonard M. Blumenthal. *Theory and applications of distance geometry*. Second edition. Chelsea Publishing Co., New York, 1970.

[Blu75] Leonard M. Blumenthal. Four-point properties and norm postulates. In *The geometry of metric and linear spaces (Proc. Conf., Michigan State Univ., East Lansing, Mich., 1974)*, pages 1–13. Lecture Notes in Math., Vol. 490. Springer, Berlin, 1975.

[BM48] Salomon Bochner and William Ted Martin. *Several Complex Variables*. Princeton Mathematical Series, vol. 10. Princeton University Press, Princeton, N. J., 1948.

[BM88] Beverly L. Brechner and John C. Mayer. Antoine's necklace or How to keep a necklace from falling apart. *College Math. J.*, 19(4):306–320, 1988.

[Boa64] R. P. Boas, Jr. Mathematical Notes: Yet Another Proof of the Fundamental Theorem of Algebra. *Amer. Math. Monthly*, 71(2):180, 1964.

[Boa86] R. P. Boas. The Teaching of Mathematics: Counterexamples to L'Hopital's Rule. *Amer. Math. Monthly*, 93(8):644–645, 1986.

[Boa95] Ralph P. Boas, Jr. *Lion hunting & other mathematical pursuits*, volume 15 of *The Dolciani Mathematical Expositions*. Mathematical Association of America, Washington, DC, 1995. A collection of mathematics, verse and stories, Edited and with an introduction by Gerald L. Alexanderson and Dale H. Mugler.

[Boa96] Ralph P. Boas. *A primer of real functions*, volume 13 of *Carus Mathematical Monographs*. Mathematical Association of America, Washington, DC, fourth edition, 1996. Revised and with a preface by Harold P. Boas.

[Bog07] V. I. Bogachev. *Measure theory. Vol. I, II.* Springer-Verlag, Berlin, 2007.

[Bol61] Oskar Bolza. *Lectures on the calculus of variations.* 2nd ed. Chelsea Publishing Co., New York, 1961.

[Bon63] F. F. Bonsall. On the representation of points of a convex set. *J. London Math. Soc.*, 38:332–334, 1963.

[Bor36] Karol Borsuk. Ensembles dont les dimensions modulaires de Alexandroff coincident avec la dimension de Menger-Urysohn. *Fund. Math.*, 27:77–93, 1936.

[Bor47] Karol Borsuk. An example of a simple arc in space whose projection in every plane has interior points. *Fund. Math.*, 34:272–277, 1947.

[Bor49] Émile Borel. *Éléments de la Théorie des Ensembles.* Éditions Albin Michel, Paris, 1949.

[Bor67] Karol Borsuk. *Theory of retracts.* Monografie Matematyczne, Tom 44. Państwowe Wydawnictwo Naukowe, Warsaw, 1967.

[Bor69] Carlos J. R. Borges. A study of absolute extensor spaces. *Pacific J. Math.*, 31:609–617, 1969.

[Bot63] H. G. Bothe. Eine Einbettung m-dimensionaler Mengen in einen $(m+1)$-dimensionalen absoluten Retrakt. *Fund. Math.*, 52:209–224, 1963.

[Bou40] N. Bourbaki. *Éléments de mathématique. Part I. Les structures fondamentales de l'analyse. Livre III. Topologie générale. Chapitres I et II.* Actualités Scientifiques et Industrielles [Current Scientific and Industrial Topics], No. 858. Hermann & Cie, Paris, 1940.

[Bou98] Nicolas Bourbaki. *General topology. Chapters 1–4.* Elements of Mathematics (Berlin). Springer-Verlag, Berlin, 1998. Translated from the French, Reprint of the 1989 English translation.

[Boy03] Werner Boy. Über die Curvatura integra und die Topologie geschlossener Flächen. *Math. Ann.*, 57(2):151–184, 1903.

[Boy59] Carl B. Boyer. *The history of the calculus and its conceptual development.* Dover Publications, Inc., New York, 1959.

[Boy89] Carl B. Boyer. *A history of mathematics.* John Wiley & Sons Inc., New York, second edition, 1989. Edited and with a preface by Uta C. Merzbach.

[BP75] Czesław Bessaga and Aleksander Pełczyński. *Selected topics in infinite-dimensional topology*, volume Tom 58 of *Monografie Matematyczne [Mathematical Monographs].* PWN—Polish Scientific Publishers, Warsaw, 1975.

[BPR06] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in real algebraic geometry*, volume 10 of *Algorithms and Computation in Mathematics.* Springer-Verlag, Berlin, second edition, 2006. Available online at www.math.purdue.edu/~sbasu/.

[BR87] Ola Bratteli and Derek W. Robinson. *Operator algebras and quantum statistical mechanics. 1.* Texts and Monographs in Physics. Springer-Verlag, New York, second edition, 1987. C^*- and W^*-algebras, symmetry groups, decomposition of states.

[Bre07] David M. Bressoud. *A radical approach to real analysis*. Classroom Resource Materials Series. Mathematical Association of America, Washington, DC, second edition, 2007.

[Bri98] Douglas S. Bridges. *Foundations of real and abstract analysis*, volume 174 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

[Bro35] Arthur B. Brown. Functional dependence. *Trans. Amer. Math. Soc.*, 38(2):379–394, 1935.

[Bro69] Andrew Browder. *Introduction to function algebras*. W. A. Benjamin, Inc., New York-Amsterdam, 1969.

[Bro96] Andrew Browder. *Mathematical analysis*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1996. An introduction.

[Bru83] Norbert Brunner. Geordnete Läuchli Kontinuen. *Fund. Math.*, 117(1):67–73, 1983.

[BS61] A. S. Besicovitch and I. J. Schoenberg. On Jordan arcs and Lipschitz classes of functions defined on them. *Acta Math.*, 106:113–136, 1961.

[BT54] A. S. Besicovitch and S. J. Taylor. On the complementary intervals of a linear closed set of zero Lebesgue measure. *J. London Math. Soc.*, 29:449–459, 1954.

[BT82] Raoul Bott and Loring W. Tu. *Differential forms in algebraic topology*, volume 82 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Berlin, 1982.

[Bun59] V. Bunyakovskii. Sur quelques inégalités. *Mem. Acad. Imp. des Sciences St. Petersburg, Series VII*, I, No. 9:1–18, 1859.

[Bus52] K. A. Bush. Continuous functions without derivatives. *Amer. Math. Monthly*, 59:222–225, 1952.

[BV06] Douglas S. Bridges and Luminiţa Simona Vîţă. *Techniques of constructive analysis*. Universitext. Springer, New York, 2006.

[BZ88] Yu. D. Burago and V. A. Zalgaller. *Geometric inequalities*, volume 285 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1988. Translated from the Russian by A. B. Sosinskiĭ, Springer Series in Soviet Mathematics.

[Caj93] Florian Cajori. *A history of mathematical notations*. Dover Publications, Inc., New York, 1993. Including Vol. I. Notations in elementary mathematics; Vol. II. Notations mainly in higher mathematics, Reprint of the 1928 and 1929 originals.

[Car40] Henri Cartan. *Sur les classes de fonctions définies par des inégalités portant sur leurs dérivées successives*. Actual. Sci. Ind., no. 867. Hermann et Cie., Paris, 1940.

[Car65] C. Carathéodory. *Calculus of variations and partial differential equations of the first order. Part I: Partial differential equations of the first order*. Translated by Robert B. Dean and Julius J. Brandstatter. Holden-Day, Inc., San Francisco-London-Amsterdam, 1965.

[Car67] C. Carathéodory. *Calculus of variations and partial differential equations of the first order. Part II: Calculus of variations*. Translated from the German by Robert B. Dean, Julius J. Brandstatter, translating editor. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1967.

[Cat06] F. S. Cater. A simple proof that a linearly ordered space is hereditarily and completely collectionwise normal. *Rocky Mountain J. Math.*, 36(4):1149–1151, 2006.

[Cau21] Augustin-Louis Cauchy. *Cours d'Analyse*. 1821. Reprint Editions Jacques Gabay 1989.

[Čer64] A. V. Černavskiĭ. Finite-to-one open mappings of manifolds. *Mat. Sb. (N.S.)*, 65 (107):357–369, 1964.

[CG96] John H. Conway and Richard K. Guy. *The book of numbers*. Copernicus, New York, 1996.

[Cha92] G. J. Chaitin. *Information-theoretic incompleteness*, volume 35 of *World Scientific Series in Computer Science*. World Scientific Publishing Co. Inc., River Edge, NJ, 1992.

[Cha99] Gregory J. Chaitin. *The unknowable*. Springer Series in Discrete Mathematics and Theoretical Computer Science. Springer-Verlag Singapore, Singapore, 1999.

[Cha05] Gregory Chaitin. *Meta math!* Pantheon Books, New York, 2005. The quest for Omega.

[Che41] Claude Chevalley. Two theorems on solvable topological groups. In *Lectures in Topology*, pages 291–292. University of Michigan Press, Ann Arbor, Mich., 1941.

[Che92] Paul R. Chernoff. A simple proof of Tychonoff's theorem via nets. *Amer. Math. Monthly*, 99(10):932–934, 1992.

[Che15] Eugenia Cheng. *How to bake π*. Basic Books, New York, 2015. An edible exploration of the mathematics of mathematics.

[CHM10] Carlos A. Cabrelli, Kathryn E. Hare, and Ursula M. Molter. Classifying Cantor sets by their fractal dimensions. *Proc. Amer. Math. Soc.*, 138(11):3965–3974, 2010.

[Cho56] Gustave Choquet. Existence des représentations intégrales au moyen des points extrémaux dans les cônes convexes. *C. R. Acad. Sci. Paris*, 243:699–702, 1956.

[CJ89] Richard Courant and Fritz John. *Introduction to calculus and analysis. Vol. II*. Springer-Verlag, New York, 1989. With the assistance of Albert A. Blank and Alan Solomon, Reprint of the 1974 edition.

[CJ99] Richard Courant and Fritz John. *Introduction to calculus and analysis. Vol. I*. Classics in Mathematics. Springer-Verlag, Berlin, 1999. Reprint of the 1989 edition.

[CK70] J. Cheeger and J. M. Kister. Counting topological manifolds. *Topology*, 9:149–151, 1970.

[Cla36] James A. Clarkson. Uniformly convex spaces. *Trans. Amer. Math. Soc.*, 40(3):396–414, 1936.

[CMMS04] Carlos Cabrelli, Franklin Mendivil, Ursula M. Molter, and Ronald Shonkwiler. On the Hausdorff h-measure of Cantor sets. *Pacific J. Math.*, 217(1):45–59, 2004.

[CMPS05] C. Cabrelli, U. Molter, V. Paulauskas, and R. Shonkwiler. Hausdorff measure of p-Cantor sets. *Real Anal. Exchange*, 30(2):413–433, 2004/05.

[CMR95] Laura E. Clemens, Kathy D. Merrill, and David W. Roeder. Continued fractions and series. *J. Number Theory*, 54(2):309–317, 1995.

[Coh66] Paul J. Cohen. *Set theory and the continuum hypothesis*. W. A. Benjamin, Inc., New York-Amsterdam, 1966.

[Con74] Randolph Constantine, Jr. Variation norm convergence of function sequences. *Proc. Amer. Math. Soc.*, 45:339–345, 1974.

[Con94] Alain Connes. *Noncommutative geometry*. Academic Press Inc., San Diego, CA, 1994.

[Con95] John B. Conway. *Functions of one complex variable. II*, volume 159 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[Cor03] David Corfield. *Towards a philosophy of real mathematics*. Cambridge University Press, Cambridge, 2003.

[CR79] Richard Courant and Herbert Robbins. *What is mathematics?* Oxford University Press, New York, 1979. An elementary approach to ideas and methods.

[Cri92] Mihai Cristea. Local inversion theorems and implicit functions theorems without assuming differentiability. *Bull. Math. Soc. Sci. Math. Roumanie (N.S.)*, 36(84)(3-4):227–236, 1992.

[CS78] D. W. Curtis and R. M. Schori. Hyperspaces of Peano continua are Hilbert cubes. *Fund. Math.*, 101(1):19–38, 1978.

[CS99] J. H. Conway and N. J. A. Sloane. *Sphere packings, lattices and groups*, volume 290 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, third edition, 1999. With additional contributions by E. Bannai, R. E. Borcherds, J. Leech, S. P. Norton, A. M. Odlyzko, R. A. Parker, L. Queen and B. B. Venkov.

[CSB54] E. Corominas and F. Sunyer Balaguer. Conditions for an infinitely differentiable function to be a polynomial. *Revista Mat. Hisp.-Amer. (4)*, 14:26–43, 1954.

[CUN02] D. Cruz-Uribe and C. J. Neugebauer. Sharp error bounds for the trapezoidal rule and Simpson's rule. *JIPAM. J. Inequal. Pure Appl. Math.*, 3(4):Article 49, 22, 2002.

[CV98] Charles O. Christenson and William L. Voxman. *Aspects of topology*. BCS Associates, Moscow, ID, second edition, 1998.

[CWY10] Stanley Chang, Shmuel Weinberger, and Guoliang Yu. Taming 3-manifolds using scalar curvature. *Geom. Dedicata*, 148:3–14, 2010.

[Dau88] Joseph W. Dauben. Abraham Robinson and nonstandard analysis: history, philosophy, and foundations of mathematics. In *History and philosophy of modern mathematics (Minneapolis, MN, 1985)*, Minnesota Stud. Philos. Sci., XI, pages 177–200. Univ. Minnesota Press, Minneapolis, MN, 1988.

[Dav52] R. O. Davies. Subsets of finite measure in analytic sets. *Indag. Math.*, 14:488–489, 1952. Nederl. Akad. Wetensch. Proc. Ser. A **55**.

[Dav59] Philip J. Davis. Leonhard Euler's integral: A historical profile of the gamma function. *Amer. Math. Monthly*, 66:849–869, 1959.

[Dav77] Martin Davis. *Applied nonstandard analysis*. Wiley-Interscience [John Wiley & Sons], New York, 1977. Pure and Applied Mathematics.

[DB86] Robert S. Doran and Victor A. Belfi. *Characterizations of C^*-algebras*, volume 101 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker Inc., New York, 1986.

[Dev73] Keith J. Devlin. *Aspects of constructibility*. Lecture Notes in Mathematics, Vol. 354. Springer-Verlag, Berlin, 1973.

[Dev77] Keith J. Devlin. *The axiom of constructibility*. Lecture Notes in Mathematics, Vol. 617. Springer-Verlag, Berlin, 1977. A guide for the mathematician.

[Dev84] Keith J. Devlin. *Constructibility*. Perspectives in Mathematical Logic. Springer-Verlag, Berlin, 1984.

[DF04] David S. Dummit and Richard M. Foote. *Abstract algebra*. John Wiley & Sons Inc., Hoboken, NJ, third edition, 2004.

[Die60] J. Dieudonné. *Foundations of modern analysis*. Pure and Applied Mathematics, Vol. X. Academic Press, New York-London, 1960.

[Dix53] J. Dixmier. Sur les bases orthonormales dans les espaces préhilbertiens. *Acta Sci. Math. Szeged*, 15:29–30, 1953.

[Dix69] Jacques Dixmier. *Les C^*-algèbres et leurs représentations*. Number 29 in Cahiers Scientifiques. Gauthier-Villars, Paris, 2nd edition, 1969. Reprinted by Éditions Jacques Gabay, Paris, 1996. Translated as *C^*-algebras*, North-Holland, Amsterdam, 1977. First Edition 1964.

[dlVP46] Ch. J. de la Vallée Poussin. *Cours d'Analyse Infinitésimale*. Dover Publications, New York, N. Y., 1946.

[DM68] P. H. Doyle and D. A. Moran. A short proof that compact 2-manifolds can be triangulated. *Invent. Math.*, 5:160–162, 1968.

[Doo94] J. L. Doob. *Measure theory*, volume 143 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994.

[Dos80] Raouf Doss. The Hahn decomposition theorem. *Proc. Amer. Math. Soc.*, 80(2):377, 1980.

[Dow55] C. H. Dowker. Local dimension of normal spaces. *Quart. J. Math. Oxford Ser. (2)*, 6:101–120, 1955.

[DR69] Roy O. Davies and C. A. Rogers. The problem of subsets of finite positive measure. *Bull. London Math. Soc.*, 1:47–54, 1969.

[DS88] Nelson Dunford and Jacob T. Schwartz. *Linear operators. Part I.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.

[DU77] J. Diestel and J. J. Uhl, Jr. *Vector measures.* American Mathematical Society, Providence, R.I., 1977. With a foreword by B. J. Pettis, Mathematical Surveys, No. 15.

[Dud92] Underwood Dudley. *Mathematical cranks.* MAA Spectrum. Mathematical Association of America, Washington, DC, 1992.

[Dud02] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[Dug78] James Dugundji. *Topology.* Allyn and Bacon Inc., Boston, Mass., 1978. Reprinting of the 1966 original, Allyn and Bacon Series in Advanced Mathematics.

[Duj01] Andrej Dujella. Newton's formula and the continued fraction expansion of $\sqrt{d}$. *Experiment. Math.*, 10(1):125–131, 2001.

[Dum00] Michael Dummett. *Elements of intuitionism*, volume 39 of *Oxford Logic Guides.* The Clarendon Press Oxford University Press, New York, second edition, 2000.

[Dun90] William Dunham. *Journey through genius.* Wiley Science Editions. John Wiley & Sons, Inc., New York, 1990. The great theorems of mathematics.

[Dun18] William Dunham. *The calculus gallery.* Princeton Science Library. Princeton University Press, Princeton, NJ, 2018. Masterpieces from Newton to Lebesgue, Reprint of the 2005 edition [MR2112402], With a new preface by the author.

[DW96] H. Garth Dales and W. Hugh Woodin. *Super-real fields*, volume 14 of *London Mathematical Society Monographs. New Series.* The Clarendon Press Oxford University Press, New York, 1996. Totally ordered fields with additional structure, Oxford Science Publications.

[Eag25] Albert Eagle. *A practical treatise on Fourier's theorem and harmonic analysis for physicists and engineers.* Longmans, Green and Co., London, 1925.

[Ebb07] Heinz-Dieter Ebbinghaus. *Ernst Zermelo.* Springer, Berlin, 2007. An approach to his life and work, In cooperation with Volker Peckhaus.

[Edg75] G. A. Edgar. A noncompact Choquet theorem. *Proc. Amer. Math. Soc.*, 49:354–358, 1975.

[Edg98] Gerald A. Edgar. *Integral, probability, and fractal measures.* Springer-Verlag, New York, 1998.

[Edg08] Gerald Edgar. *Measure, topology, and fractal geometry.* Undergraduate Texts in Mathematics. Springer, New York, second edition, 2008.

[Edw79] C. H. Edwards, Jr. *The historical development of the calculus.* Springer-Verlag, New York-Heidelberg, 1979.

[Eff67] Edward G. Effros. Structure in simplexes. *Acta Math.*, 117:103–121, 1967.

[Ele97] Neven Elezović. A note on continued fractions of quadratic irrationals. *Math. Commun.*, 2(1):27–33, 1997.

[EM42a] Samuel Eilenberg and Saunders MacLane. Group extensions and homology. *Ann. of Math. (2)*, 43:757–831, 1942.

[EM42b] Samuel Eilenberg and Saunders MacLane. Natural isomorphisms in group theory. *Proc. Nat. Acad. Sci. U.S.A.*, 28:537–543, 1942.

[EM45] Samuel Eilenberg and Saunders MacLane. General theory of natural equivalences. *Trans. Amer. Math. Soc.*, 58:231–294, 1945.

[Enf73] Per Enflo. A counterexample to the approximation problem in Banach spaces. *Acta Math.*, 130:309–317, 1973.

[Eng89] Ryszard Engelking. *General topology*, volume 6 of *Sigma Series in Pure Mathematics*. Heldermann Verlag, Berlin, second edition, 1989. Translated from the Polish by the author.

[Eng95] Ryszard Engelking. *Theory of dimensions finite and infinite*, volume 10 of *Sigma Series in Pure Mathematics*. Heldermann Verlag, Lemgo, 1995.

[EoGKP88] P. Erd˝os, C. D. Godsil, S. G. Krantz, and T. D. Parsons. Intersection graphs for families of balls in $\mathbf{R}^n$. *European J. Combin.*, 9(5):501–505, 1988.

[Esc62] L. È. Èl′ sgol′c. *Calculus of variations*. Pergamon Press Ltd., London-Paris-Frankfurt, Addison-Wesley Publishing Co., Inc., Reading, Mass., 1962.

[Euc02] Euclid. *Euclid's Elements*. Green Lion Press, Santa Fe, NM, 2002. All thirteen books complete in one volume, The Thomas L. Heath translation, Edited by Dana Densmore.

[Eul00] L. Euler. *Foundations of differential calculus*. Springer-Verlag, New York, 2000. Translated from the Latin by John D. Blanton.

[Eve90] Howard Eves. *An introduction to the history of mathematics*. Saunders Series. Saunders College Publishing, Philadelphia, PA, sixth edition, 1990. With cultural connections by Jamie H. Eves.

[Ewi85] George M. Ewing. *Calculus of variations with applications*. Dover Publications, Inc., New York, 1985. Corrected reprint of the 1969 original.

[FA48] Ralph H. Fox and Emil Artin. Some wild cells and spheres in three-dimensional space. *Ann. of Math. (2)*, 49:979–990, 1948.

[Fab66] J. Fabius. A probabilistic example of a nowhere analytic C^∞-function. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 5:173–174 (1966), 1966.

[Fal14] Kenneth Falconer. *Fractal geometry*. John Wiley & Sons, Ltd., Chichester, third edition, 2014. Mathematical foundations and applications.

[Fc57] G. M. Fihtengol′ c. *Foundations of mathematical analysis. Vol. I–II*. Gosudarstv. Izdat. Tehn.-Teor. Lit., Moscow, 1957.

[Fed69] Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[Fed95] V. V. Fedorchuk. A differentiable manifold with noncoinciding dimensions for the continuum hypothesis. *Mat. Sb.*, 186(1):149–160, 1995.

[Fed98] V. V. Fedorchuk. The Uryson identity and the dimension of manifolds. *Uspekhi Mat. Nauk*, 53(5(323)):73–114, 1998.

[FH10] J. F. Feinstein and M. J. Heath. Swiss cheeses, rational approximation and universal plane curves. *Studia Math.*, 196(3):289–306, 2010.

[Fic89] G. M. Fichtenholz. *Differential- und Integralrechnung. I–III*, volume 61 of *Hochschulbücher für Mathematik [University Books for Mathematics]*. VEB Deutscher Verlag der Wissenschaften, Berlin, thirteenth edition, 1989. Translated from the Russian by Ludwig Boll.

[Fil69] V. V. Filippov. A bicompactum with noncoinciding inductive dimensions. *Dokl. Akad. Nauk SSSR*, 184:1050–1053, 1969.

[Fil72] V. V. Filippov. The inductive dimension of a product of bicompacta. *Dokl. Akad. Nauk SSSR*, 202:1016–1019, 1972.

[FK78] W. G. Fleissner and K. Kunen. Barely Baire spaces. *Fund. Math.*, 101(3):229–240, 1978.

[FL94] Zoltán Füredi and Peter A. Loeb. On the best constant for the Besicovitch covering theorem. *Proc. Amer. Math. Soc.*, 121(4):1063–1073, 1994.

[Fol99] Gerald B. Folland. *Real analysis*. Pure and Applied Mathematics (New York). John Wiley & Sons Inc., New York, second edition, 1999. Modern techniques and their applications, A Wiley-Interscience Publication.

[For60] A. R. Forsyth. *Calculus of variations*. Dover Publications, Inc., New York, 1960.

[Fow99] David Fowler. *The mathematics of Plato's academy*. The Clarendon Press, Oxford University Press, New York, second edition, 1999. A new reconstruction.

[Fox43] Ralph H. Fox. On fibre spaces. II. *Bull. Amer. Math. Soc.*, 49:733–735, 1943.

[Fra22] Adolf Fraenkel. Zu den Grundlagen der Cantor-Zermeloschen Mengenlehre. *Math. Ann.*, 86(3-4):230–237, 1922.

[Fra65] S. P. Franklin. Spaces in which sequences suffice. *Fund. Math.*, 57:107–115, 1965.

[Fra66] Abraham A. Fraenkel. *Set theory and logic*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1966.

[Fre37] Hans Freudenthal. Entwicklungen von Räumen und ihren Gruppen. *Compositio Math.*, 4:145–234, 1937.

[Fre64] Peter Freyd. *Abelian categories. An introduction to the theory of functors*. Harper's Series in Modern Mathematics. Harper & Row, Publishers, New York, 1964.

[Fre84] D. H. Fremlin. *Consequences of Martin's axiom*, volume 84 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1984.

[Fre04] D. H. Fremlin. *Measure theory. Vol. 1.* Torres Fremlin, Colchester, 2004. The irreducible minimum, Corrected third printing of the 2000 original.

[Fre06] D. H. Fremlin. *Measure theory. Vol. 4.* Torres Fremlin, Colchester, 2006. Topological measure spaces. Part I, II, Corrected second printing of the 2003 original.

[Fre13] Edward Frenkel. *Love and math.* Basic Books, New York, 2013. The heart of hidden reality.

[Fri37] A. H. Frink. Distance functions and the metrization problem. *Bull. Amer. Math. Soc.*, 43:133–142, 1937.

[Fuc59] L. Fuchs. On character groups of discrete abelian groups. *Acta Math. Acad. Sci. Hungar.*, 10:133–140 (unbound insert), 1959.

[Fuk52] Masanori Fukamiya. On a theorem of Gelfand and Neumark and the B^*-algebra. *Kumamoto J. Sci. Ser. A.*, 1(1):17–22, 1952.

[FW99] George K. Francis and Jeffrey R. Weeks. Conway's ZIP proof. *Amer. Math. Monthly*, 106(5):393–399, 1999.

[Gab11] David Gabai. The Whitehead manifold is a union of two Euclidean spaces. *J. Topol.*, 4(3):529–534, 2011.

[Gag58] Emilio Gagliardo. Proprietà di alcune classi di funzioni in più variabili. *Ricerche Mat.*, 7:102–137, 1958.

[Gam01] Theodore W. Gamelin. *Complex analysis.* Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2001.

[Gar01] Martin Gardner. *The colossal book of mathematics.* W. W. Norton & Co. Inc., New York, 2001. Classic puzzles, paradoxes, and problems, Number theory, algebra, geometry, probability, topology, game theory, infinity and other topics of recreational mathematics.

[Gel41] I. Gelfand. Normierte Ringe. *Rec. Math. [Mat. Sbornik] N. S.*, 9 (51):3–24, 1941.

[GF63] I. M. Gelfand and S. V. Fomin. *Calculus of variations.* Revised English edition translated and edited by Richard A. Silverman. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963.

[GH78] H. B. Griffiths and P. J. Hilton. *A comprehensive textbook of classical mathematics.* Springer-Verlag, New York-Heidelberg, 1978. A contemporary interpretation, Reprinted from the 1970 edition.

[GH96a] Mariano Giaquinta and Stefan Hildebrandt. *Calculus of variations. I*, volume 310 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. The Lagrangian formalism.

[GH96b] Mariano Giaquinta and Stefan Hildebrandt. *Calculus of variations. II*, volume 311 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. The Hamiltonian formalism.

[GH19] Victor Guillemin and Peter Haine. *Differential forms*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2019.

[GJ76] Leonard Gillman and Meyer Jerison. *Rings of continuous functions*. Springer-Verlag, New York, 1976. Reprint of the 1960 edition, Graduate Texts in Mathematics, No. 43.

[GK06] Robert E. Greene and Steven G. Krantz. *Function theory of one complex variable*, volume 40 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, third edition, 2006.

[GK09] Loren Graham and Jean-Michel Kantor. *Naming infinity*. The Belknap Press of Harvard University Press, Cambridge, MA, 2009. A true story of religious mysticism and mathematical creativity.

[Gle52] Andrew M. Gleason. Groups without small subgroups. *Ann. of Math. (2)*, 56:193–212, 1952.

[GLS94] Daniel Gorenstein, Richard Lyons, and Ronald Solomon. *The classification of the finite simple groups*, volume 40 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1994.

[GM78] J. D. Gray and S. A. Morris. When is a function that satisfies the Cauchy-Riemann equations analytic? *Amer. Math. Monthly*, 85(4):246–256, 1978.

[GN43] I. Gelfand and M. Neumark. On the imbedding of normed rings into the ring of operators in Hilbert space. *Rec. Math. [Mat. Sbornik] N.S.*, 12(54):197–213, 1943.

[God04] Roger Godement. *Analysis. I*. Universitext. Springer-Verlag, Berlin, 2004. Convergence, elementary functions, Translated from the 1998 French edition by Philip Spain.

[Gol54] S. W. Golomb. Checker boards and polyominoes. *Amer. Math. Monthly*, 61:675–682, 1954.

[Gol80] Herman H. Goldstine. *A history of the calculus of variations from the 17th through the 19th century*, volume 5 of *Studies in the History of Mathematics and Physical Sciences*. Springer-Verlag, New York-Berlin, 1980.

[Gol88] Warren Goldfarb. Poincaré against the Logicists. In *History and philosophy of modern mathematics (Minneapolis, MN, 1985)*, Minnesota Stud. Philos. Sci., XI, pages 61–81. Univ. Minnesota Press, Minneapolis, MN, 1988.

[Gol98] Robert Goldblatt. *Lectures on the hyperreals*, volume 188 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998. An introduction to nonstandard analysis.

[Goo86a] K. R. Goodearl. *Partially ordered abelian groups with interpolation*, volume 20 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1986.

[Goo86b] Nicolas D. Goodman. Mathematics as an objective science [Amer. Math. Monthly **86** (1979), no. 7, 540–551; MR0542765 (80m:00008)]. In *New directions in the philosophy of mathematics*, pages 79–94. Birkhäuser Boston, Boston, MA, 1986.

[Gou03] Ed. Goursat. Sur la théorie des fonctions implicites. *Bull. Soc. Math. France*, 31:184–192, 1903.

[Gou59] Édouard Goursat. *A course in mathematical analysis: Vol 1: Derivatives and differentials, definite integrals, expansion in series, applications to geometry. Vol. 2, Part 1: Functions of a complex variable. Vol. 2, Part 2: Differential equations*. Translated by E. R. Hedrick (Vol. 1), and E. R. Hedrick and O. Dunkel (Vol. 2). Dover Publications, Inc., New York, 1959.

[Gow95] W. T. Gowers. Recent results in the theory of infinite-dimensional Banach spaces. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pages 933–942, Basel, 1995. Birkhäuser.

[Gow02] Timothy Gowers. *Mathematics*, volume 66 of *Very Short Introductions*. Oxford University Press, Oxford, 2002. A very short introduction.

[GP10] Victor Guillemin and Alan Pollack. *Differential topology*. AMS Chelsea Publishing, Providence, RI, 2010. Reprint of the 1974 original.

[GR09] Robert C. Gunning and Hugo Rossi. *Analytic functions of several complex variables*. AMS Chelsea Publishing, Providence, RI, 2009. Reprint of the 1965 original.

[Gra00] *From the calculus to set theory, 1630–1910*. Princeton Paperbacks. Princeton University Press, Princeton, NJ, 2000. An introductory history, Edited by I. Grattan-Guinness, Reprint of the 1980 original.

[Gra08] Jeremy Gray. *Plato's ghost*. Princeton University Press, Princeton, NJ, 2008. The modernist transformation of mathematics.

[Gra10] Judith V. Grabiner. *A historian looks back*. MAA Spectrum. Mathematical Association of America, Washington, DC, 2010. ıt The calculus as algebra and selected writings.

[Gre93] Marvin Jay Greenberg. *Euclidean and non-Euclidean geometries*. W. H. Freeman and Company, New York, third edition, 1993. Development and history.

[Gri21] Frank L. Griffin. *Introduction to mathematical analysis*. Houghton-Mifflin, 1921.

[Gri27] Frank L. Griffin. *Mathematical analysis: higher course*. Houghton-Mifflin, 1927.

[Gri46] G. F. C. Griss. Negationless intuitionistic mathematics. *Nederl. Akad. Wetensch., Proc.*, 49:1127–1133 = Indagationes Math. 8, 675–681 (1946), 1946.

[Gri51] G. F. C. Griss. Negationless intuitionistic mathematics. IVa. *Nederl. Akad. Wetensch. Proc. Ser. A.* **54** = *Indagationes Math.*, 13:452–462, 1951.

[Gri85] Eric L. Grinberg. On the smoothness hypothesis in Sard's theorem. *Amer. Math. Monthly*, 92(10):733–734, 1985.

[Gro57] Alexander Grothendieck. Sur quelques points d'algèbre homologique. *Tohoku Math. J. (2)*, 9:119–221, 1957.

[Gro99] Misha Gromov. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152 of *Progress in Mathematics*. Birkhäuser Boston, Inc., Boston, MA, 1999. Based on the 1981 French original [MR0682063 (85e:53051)], With appendices by M. Katz, P. Pansu and S. Semmes, Translated from the French by Sean Michael Bates.

[Gui83] Michael Guillen. *Bridges to infinity*. Jeremy P. Tarcher Inc., Los Angeles, CA, 1983. The human side of mathematics.

[Gun90a] Robert C. Gunning. *Introduction to holomorphic functions of several variables. Vol. I.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990. Function theory.

[Gun90b] Robert C. Gunning. *Introduction to holomorphic functions of several variables. Vol. II.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1990. Local theory.

[Gun90c] Robert C. Gunning. *Introduction to holomorphic functions of several variables. Vol. III.* The Wadsworth & Brooks/Cole Mathematics Series. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA, 1990. Homological theory.

[H90] Lars Hörmander. *An introduction to complex analysis in several variables*, volume 7 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, third edition, 1990.

[Had47] *The Royal Society. Newton Tercentenary Celebrations, 15–19 July 1946*. Published for The Royal Society at the University Press, Cambridge, 1947.

[Hag77] Charles L. Hagopian. A characterization of solenoids. *Pacific J. Math.*, 68(2):425–435, 1977.

[Hal50] Paul R. Halmos. *Measure Theory*. D. Van Nostrand Company, Inc., New York, N. Y., 1950.

[Hal74] Hubert Halkin. Implicit functions and optimization problems without continuous differentiability of the data. *SIAM J. Control*, 12:229–236, 1974. Collection of articles dedicated to the memory of Lucien W. Neustadt.

[Hal82] Paul Richard Halmos. *A Hilbert space problem book*. Springer-Verlag, New York, second edition, 1982. Encyclopedia of Mathematics and its Applications, 17.

[Hal07] Thomas C. Hales. Jordan's proof of the Jordan Curve Theorem. *Studies in Logic, Grammar and Rhetoric*, 10 (23):45–60, 2007.

[Ham86] R. W. Hamming. *Numerical methods for scientists and engineers*. Dover Publications, Inc., New York, second edition, 1986.

[Ham98] R. W. Hamming. Mathematics on a distant planet. *Amer. Math. Monthly*, 105(7):640–650, 1998.

[Har00] Robin Hartshorne. *Geometry: Euclid and beyond*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, 2000.

[Hat02] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002. Online copy available at http://www.math.cornell.edu/~hatcher/AT/ATpage.html.

[Hau14] F. Hausdorff. Bemerkung über den Inhalt von Punktmengen. *Math. Ann.*, 75(3):428–433, 1914.

[Hau32] Felix Hausdorff. Zur Theorie der linearen metrischen Räume. *J. Reine Angew. Math.*, 167:294–311, 1932.

[Hau44] F. Hausdorff. *Mengenlehre*. Dover Publications, New York, N. Y., 1944.

[Hau49] Felix Hausdorff. *Grundzüge der Mengenlehre*. Chelsea Publishing Company, New York, N. Y., 1949.

[Hau57] Felix Hausdorff. *Set theory*. Chelsea Publishing Company, New York, 1957. Translated by John R. Aumann, et al.

[HD64] Hugo Hadwiger and Hans Debrunner. *Combinatorial geometry in the plane*. Translated by Victor Klee. With a new chapter and other additional material supplied by the translator. Holt, Rinehart and Winston, New York, 1964.

[Hem46] Erik Hemmingsen. Some theorems in dimension theory for normal Hausdorff spaces. *Duke Math. J.*, 13:495–504, 1946.

[Hen80] Julien Hennefeld. Classroom Notes: A Nontopological Proof of the Uniform Boundedness Theorem. *Amer. Math. Monthly*, 87(3):217, 1980.

[Her06] Horst Herrlich. *Axiom of choice*, volume 1876 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.

[Hey66] A. Heyting. *Intuitionism: An introduction*. Second revised edition. North-Holland Publishing Co., Amsterdam, 1966.

[HF11] Thomas Hales and Samuel Ferguson. *The Kepler conjecture*. Springer, New York, 2011. The Hales-Ferguson proof, Including papers reprinted from Discrete Comput. Geom. **3**6 (2006), no. 1, Edited by Jeffrey C. Lagarias.

[HH99] John Hamal Hubbard and Barbara Burke Hubbard. *Vector calculus, linear algebra, and differential forms*. Prentice Hall, Inc., Upper Saddle River, NJ, 1999. A unified approach.

[Hil43] Einar Hille. Review of "On the embedding of normed rings into the ring of operators in Hilbert space," by I. Gelfand and M. Neumark. *Math. Reviews*, MR0009426 (5,147d), 1943.

[Hil87] F. B. Hildebrand. *Introduction to numerical analysis*. Dover Publications, Inc., New York, second edition, 1987.

[Hir94] Morris W. Hirsch. *Differential topology*, volume 33 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994. Corrected reprint of the 1976 original.

[HL85] Albert E. Hurd and Peter A. Loeb. *An introduction to nonstandard real analysis*, volume 118 of *Pure and Applied Mathematics*. Academic Press Inc., Orlando, FL, 1985.

[Hof71] Joseph E. Hofmann. über die ersten infinitesimalmathematischen Studien von Johann Bernoulli. In *Actes XIIe Congrès Internat. d'Histoire des Sciences (Paris, 1968), Tome IV: Histoire des Mathématiques et de la Mécanique*, pages 89–95. Librairie Sci. et Techn. A. Blanchard, Paris, 1971.

[How95] J. D. Howroyd. On dimension and on the existence of sets of finite positive Hausdorff measure. *Proc. London Math. Soc. (3)*, 70(3):581–604, 1995.

[HR98] Paul Howard and Jean E. Rubin. *Consequences of the axiom of choice*, volume 59 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1998. With 1 IBM-PC floppy disk (3.5 inch; WD).

[HS75] Edwin Hewitt and Karl Stromberg. *Real and abstract analysis*. Springer-Verlag, New York, 1975. A modern treatment of the theory of functions of a real variable, Third printing, Graduate Texts in Mathematics, No. 25.

[HT96] Stefan Hildebrandt and Anthony Tromba. *The parsimonious universe*. Copernicus, New York, 1996. Shape and form in the natural world.

[HT06] Horst Herrlich and Eleftherios Tachtsis. On the number of Russell's socks or $2 + 2 + 2 + \cdots =?$ *Comment. Math. Univ. Carolin.*, 47(4):707–717, 2006. Online copy available at https://www.emis.de/journals/CMUC/pdf/cmuc0604/herrtach.pdf.

[HW41] Witold Hurewicz and Henry Wallman. *Dimension Theory*. Princeton Mathematical Series, v. 4. Princeton University Press, Princeton, N. J., 1941.

[HW08] G. H. Hardy and E. M. Wright. *An introduction to the theory of numbers*. Oxford University Press, Oxford, sixth edition, 2008. Revised by D. R. Heath-Brown and J. H. Silverman.

[HY88] John G. Hocking and Gail S. Young. *Topology*. Dover Publications Inc., New York, second edition, 1988.

[IK94] Eugene Isaacson and Herbert Bishop Keller. *Analysis of numerical methods*. Dover Publications, Inc., New York, 1994. Corrected reprint of the 1966 original [Wiley, New York; MR0201039 (34 #924)].

[Inc44] E. L. Ince. *Ordinary Differential Equations*. Dover Publications, New York, 1944.

[Isb64] J. R. Isbell. *Uniform spaces*. Mathematical Surveys, No. 12. American Mathematical Society, Providence, R.I., 1964.

[Iwa49] Yosikazu Iwamoto. A proof that π^2 is irrational. *J. Osaka Inst. Sci. Tech. Part I.*, 1:147–148, 1949.

[Jah03] Hans Niels Jahnke, editor. *A history of analysis*, volume 24 of *History of Mathematics*. American Mathematical Society, Providence, RI, 2003. Translated from the German.

[Jam82] Robert C. James. Bases in Banach spaces. *Amer. Math. Monthly*, 89(9):625–640, 1982.

[Jam99] I. M. James, editor. *History of topology*. North-Holland, Amsterdam, 1999.

[Jec73] Thomas J. Jech. *The axiom of choice*. North-Holland Publishing Co., Amsterdam, 1973. Studies in Logic and the Foundations of Mathematics, Vol. 75.

[Jec03] Thomas Jech. *Set theory*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2003. The third millennium edition, revised and expanded.

[JLJ98] Jürgen Jost and Xianqing Li-Jost. *Calculus of variations*, volume 64 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1998.

[Joh02] Warren P. Johnson. The curious history of Faà di Bruno's formula. *Amer. Math. Monthly*, 109(3):217–234, 2002.

[JP95] H. Joyce and D. Preiss. On the existence of subsets of finite positive packing measure. *Mathematika*, 42(1):15–24, 1995.

[JT80] William B. Jones and Wolfgang J. Thron. *Continued fractions*, volume 11 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley Publishing Co., Reading, Mass., 1980. Analytic theory and applications, With a foreword by Felix E. Browder, With an introduction by Peter Henrici.

[Jz75] A. P. Juˇzakov. The application of the multiple logarithmic residue to the expansion of implicit functions in power series. *Mat. Sb. (N.S.)*, 97(139)(2(6)):177–192, 317, 1975.

[Kad66] M. Ǐ. Kadec′. Topological equivalence of all separable Banach spaces. *Dokl. Akad. Nauk SSSR*, 167:23–25, 1966.

[Kan03] Akihiro Kanamori. *The higher infinite*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, second edition, 2003. Large cardinals in set theory from their beginnings.

[Kap68] Irving Kaplansky. *Rings of operators*. W. A. Benjamin, Inc., New York-Amsterdam, 1968.

[Kap71] Irving Kaplansky. *Lie algebras and locally compact groups*. The University of Chicago Press, Chicago, Ill.-London, 1971.

[Kap00] Robert Kaplan. *The nothing that is*. Oxford University Press, New York, 2000. A natural history of zero.

[Kat52] Miroslav Katětov. On the dimension of non-separable spaces. I. *Čehoslovack. Mat. Ž.*, 2(77):333–368 (1953), 1952.

[Kat04] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, third edition, 2004.

[Kec95] Alexander S. Kechris. *Classical descriptive set theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

[Kel42] J. L. Kelley. Hyperspaces of a continuum. *Trans. Amer. Math. Soc.*, 52:22–36, 1942.

[Kel63] J. C. Kelly. Bitopological spaces. *Proc. London Math. Soc. (3)*, 13:71–89, 1963.

[Kel75] John L. Kelley. *General topology*. Springer-Verlag, New York, 1975. Reprint of the 1955 edition [Van Nostrand, Toronto, Ont.], Graduate Texts in Mathematics, No. 27.

[KF75] A. N. Kolmogorov and S. V. Fomīn. *Introductory real analysis*. Dover Publications Inc., New York, 1975. Translated from the second Russian edition and edited by Richard A. Silverman, Corrected reprinting.

[Khi97] A. Ya. Khinchin. *Continued fractions*. Dover Publications Inc., Mineola, NY, russian edition, 1997. With a preface by B. V. Gnedenko, Reprint of the 1964 translation.

[KK03] Robert Kaplan and Ellen Kaplan. *The art of the infinite*. Oxford University Press, Oxford, 2003. The pleasures of mathematics.

[Kle55] V. L. Klee, Jr. Some topological properties of convex sets. *Trans. Amer. Math. Soc.*, 78:30–45, 1955.

[Kle59] Victor Klee. Some new results on smoothness and rotundity in normed linear spaces. *Math. Ann.*, 139:51–63 (1959), 1959.

[Kle02] Stephen Cole Kleene. *Mathematical logic*. Dover Publications Inc., Mineola, NY, 2002. Reprint of the 1967 original [Wiley, New York; MR0216930 (36 #25)].

[Kle04] Felix Klein. *Elementary mathematics from an advanced standpoint*. Dover Publications, Inc., Mineola, NY, 2004. Arithmetic, algebra, analysis, Translated from the third German edition by E. R. Hedrick and C. A. Noble, Reprint of the 1932 translation.

[Kli72] Morris Kline. *Mathematical thought from ancient to modern times*. Oxford University Press, New York, 1972.

[KM76] Kazimierz Kuratowski and Andrzej Mostowski. *Set theory*. North-Holland Publishing Co., Amsterdam, revised edition, 1976. With an introduction to descriptive set theory, Translated from the 1966 Polish original, Studies in Logic and the Foundations of Mathematics, Vol. 86.

[KM78] A. Kanamori and M. Magidor. The evolution of large cardinal axioms in set theory. In *Higher set theory (Proc. Conf., Math. Forschungsinst., Oberwolfach, 1977)*, volume 669 of *Lecture Notes in Math.*, pages 99–275. Springer, Berlin, 1978.

[KN76] John L. Kelley and Isaac Namioka. *Linear topological spaces*. Springer-Verlag, New York-Heidelberg, 1976. With the collaboration of W. F. Donoghue, Jr., Kenneth R. Lucas, B. J. Pettis, Ebbe Thue Poulsen, G. Baley Price, Wendy Robertson, W. R. Scott, and Kennan T. Smith, Second corrected printing, Graduate Texts in Mathematics, No. 36.

[Kne60] Martin Kneser. Beispiel einer dimensionserhöhenden analytischen Abbildung zwischen überabzählbaren Mannigfaltigkeiten. *Arch. Math. (Basel)*, 11:280–281, 1960.

[Knu76] Donald E. Knuth. Mathematics and computer science: coping with finiteness. *Science*, 194(4271):1235–1242, 1976.

[Koe91] Teun Koetsier. *Lakatos' philosophy of mathematics*, volume 3 of *Studies in the History and Philosophy of Mathematics*. North-Holland Publishing Co., Amsterdam, 1991. A historical approach.

[Kok49] J. F. Koksma. On Niven's proof that π is irrational. *Nieuw Arch. Wiskunde (2)*, 23:39, 1949.

[Kön97] Heinz König. *Measure and integration*. Springer-Verlag, Berlin, 1997. An advanced course in basic procedures and applications.

[Kor53] P. P. Korovkin. On convergence of linear positive operators in the space of continuous functions. *Doklady Akad. Nauk SSSR (N.S.)*, 90:961–964, 1953.

[Kör83] T. W. Körner. The behavior of power series on their circle of convergence. In *Banach spaces, harmonic analysis, and probability theory (Storrs, Conn., 1980/1981)*, volume 995 of *Lecture Notes in Math.*, pages 56–94. Springer, Berlin, 1983.

[Kör04] T. W. Körner. *A companion to analysis*, volume 62 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2004. A second first and first second course in analysis.

[KP71] K. Kunen and J. B. Paris. Boolean extensions and measurable cardinals. *Ann. Math. Logic*, 2(4):359–377, 1970/1971.

[KP86] Steven G. Krantz and T. D. Parsons. Antisocial subcovers of self-centered coverings. *Amer. Math. Monthly*, 93(1):45–48, 1986.

[KP02a] Steven G. Krantz and Harold R. Parks. *The implicit function theorem*. Birkhäuser Boston Inc., Boston, MA, 2002. History, theory, and applications.

[KP02b] Steven G. Krantz and Harold R. Parks. *A primer of real analytic functions*. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Birkhäuser Boston Inc., Boston, MA, second edition, 2002.

[Kra05] Steven G. Krantz. *Real analysis and foundations*. Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2005.

[Kra11] Steven G. Krantz. *The proof is in the pudding*. Springer, New York, 2011. The changing nature of mathematical proof.

[Kra15] Steven G. Krantz. *How to teach mathematics*. American Mathematical Society, Providence, RI, third edition, 2015.

[Kre89] Erwin Kreyszig. *Introductory functional analysis with applications*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1989.

[KS74] Y. Katznelson and Karl Stromberg. Everywhere differentiable, nowhere monotone, functions. *Amer. Math. Monthly*, 81:349–354, 1974.

[KS20] Asaf Karagila and Philipp Schlicht. How to have more things by forgetting how to count them. *Proc. A.*, 476(2239):20190782, 12, 2020.

[KT66] J. F. C. Kingman and S. J. Taylor. *Introduction to measure and probability*. Cambridge University Press, London, 1966.

[KT01] Kyriakos Keremedis and Eleftherios Tachtsis. Compact metric spaces and weak forms of the axiom of choice. *MLQ Math. Log. Q.*, 47(1):117–128, 2001.

[Kun70] Kenneth Kunen. Some applications of iterated ultrapowers in set theory. *Ann. Math. Logic*, 1:179–227, 1970.

[Kur66] K. Kuratowski. *Topology. Vol. I.* New edition, revised and augmented. Translated from the French by J. Jaworowski. Academic Press, New York, 1966.

[Kur68] K. Kuratowski. *Topology. Vol. II.* New edition, revised and augmented. Translated from the French by A. Kirkor. Academic Press, New York, 1968.

[KV53] J. L. Kelley and R. L. Vaught. The positive cone in Banach algebras. *Trans. Amer. Math. Soc.*, 74:44–55, 1953.

[KV84] Kenneth Kunen and Jerry E. Vaughan, editors. *Handbook of set-theoretic topology.* North-Holland Publishing Co., Amsterdam, 1984.

[KW04] Eberhard Kirchberg and Wilhelm Winter. Covering dimension and quasidiagonality. *Internat. J. Math.*, 15(1):63–85, 2004.

[L55] M. H. Löb. Solution of a problem of Leon Henkin. *J. Symbolic Logic*, 20:115–118, 1955.

[L63] H. Läuchli. Auswahlaxiom in der Algebra. *Comment. Math. Helv.*, 37:1–18, 1962/63.

[Lak76] Imre Lakatos. *Proofs and refutations.* Cambridge University Press, Cambridge, 1976. The logic of mathematical discovery, Edited by John Worrall and Elie Zahar.

[Lak78] Imre Lakatos. Cauchy and the continuum: the significance of nonstandard analysis for the history and philosophy of mathematics. *Math. Intelligencer*, 1(3):151–161, 1978.

[Lan51] Edmund Landau. *Foundations of Analysis. The Arithmetic of Whole, Rational, Irrational and Complex Numbers.* Chelsea Publishing Company, New York, N.Y., 1951. Translated by F. Steinhardt.

[Lan02] Serge Lang. *Algebra*, volume 211 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, third edition, 2002.

[Lar66] D. G. Larman. On the exponent of convergence of a packing of spheres. *Mathematika*, 13:57–59, 1966.

[Lar67a] D. G. Larman. A new theory of dimension. *Proc. London Math. Soc. (3)*, 17:178–192, 1967.

[Lar67b] D. G. Larman. On Hausdorff measure in finite-dimensional compact metric spaces. *Proc. London Math. Soc. (3)*, 17:193–206, 1967.

[Lau87] Detlef Laugwitz. Infinitely small quantities in Cauchy's textbooks. *Historia Math.*, 14(3):258–274, 1987.

[Lau89] Detlef Laugwitz. Definite values of infinite sums: aspects of the foundations of infinitesimal analysis around 1820. *Arch. Hist. Exact Sci.*, 39(3):195–245, 1989.

[Lax02] Peter D. Lax. *Functional analysis.* Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2002.

[LDR19] Enrico Le Donne and Séverine Rigot. Besicovitch covering property on graded groups and applications to measure differentiation. *J. Reine Angew. Math.*, 750:241–297, 2019.

[LeC73] L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[LeV56] William Judson LeVeque. *Topics in number theory. Vols. 1 and 2*. Addison-Wesley Publishing Co., Inc., Reading, Mass., 1956.

[Lic62] W. B. R. Lickorish. A representation of orientable combinatorial 3-manifolds. *Ann. of Math. (2)*, 76:531–540, 1962.

[Liv02] Mario Livio. *The golden ratio*. Broadway Books, New York, 2002. The story of phi, the world's most astonishing number.

[Liv09] Mario Livio. *Is God a mathematician?* Simon & Schuster, New York, 2009.

[LL97] Elliott H. Lieb and Michael Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1997.

[LM54] P. D. Lax and A. N. Milgram. Parabolic equations. In *Contributions to the theory of partial differential equations*, Annals of Mathematics Studies, no. 33, pages 167–190. Princeton University Press, Princeton, N. J., 1954.

[LN00] George Lakoff and Rafael E. Núñez. *Where mathematics comes from*. Basic Books, New York, 2000. How the embodied mind brings mathematics into being.

[Lor01] G. G. Lorentz. Who discovered analytic sets? *Math. Intelligencer*, 23(4):28–32, 2001.

[LS97] F. William Lawvere and Stephen H. Schanuel. *Conceptual mathematics*. Cambridge University Press, Cambridge, 1997. A first introduction to categories, Revised edition of the 1991 original.

[LS98] Jouni Luukkainen and Eero Saksman. Every complete doubling metric space carries a doubling measure. *Proc. Amer. Math. Soc.*, 126(2):531–534, 1998.

[Lub53] M. Lubański. An example of an absolute neighbourhood retract, which is the common boundary of three regions in the 3-dimensional Euclidean space. *Fund. Math.*, 40:29–38, 1953.

[Luc00] J. R. Lucas. *The conceptual roots of mathematics*. International Library of Philosophy. Routledge, London, 2000. An essay on the philosophy of mathematics.

[Lun49] A. Lunc. A bicompactum whose inductive dimension is greater than its dimension defined by means of coverings. *Doklady Akad. Nauk SSSR (N.S.)*, 66:801–803, 1949.

[Luu98] Jouni Luukkainen. Assouad dimension: antifractal metrization, porous sets, and homogeneous measures. *J. Korean Math. Soc.*, 35(1):23–76, 1998.

[LW98] E. J. Larson and L. Witham. Leading scientists still reject god. *Nature*, 394:313, 1998.

[LW08] Lisa Lorentzen and Haakon Waadeland. *Continued fractions. Vol. 1*, volume 1 of *Atlantis Studies in Mathematics for Engineering and Science*. Atlantis Press, Paris, second edition, 2008. Convergence theory.

[Mad88] Penelope Maddy. Believing the axioms. I. *J. Symbolic Logic*, 53(2):481–511, 1988.

[Mae84] Ryuji Maehara. The Jordan curve theorem via the Brouwer fixed point theorem. *Amer. Math. Monthly*, 91(10):641–643, 1984.

[Mar76] Donald A. Martin. Hilbert's first problem: the continuum hypothesis. In *Mathematical developments arising from Hilbert problems (Proc. Sympos. Pure Math., Northern Illinois Univ., De Kalb, Ill., 1974)*, pages 81–92. Amer. Math. Soc., Providence, R.I., 1976.

[Mas77] William S. Massey. *Algebraic topology: an introduction*. Springer-Verlag, New York-Heidelberg, 1977. Reprint of the 1967 edition, Graduate Texts in Mathematics, Vol. 56.

[Mas91] William S. Massey. *A basic course in algebraic topology*, volume 127 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991.

[Mat95] Pertti Mattila. *Geometry of sets and measures in Euclidean spaces*, volume 44 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995. Fractals and rectifiability.

[May06] R. Mayans. The Chebyshev equioscillation theorem. *J. Online Math.*, 6, 2006. http://www.maa.org/publications/periodicals/loci/joma/the-chebyshev-equioscillation-theorem.

[McM62] D. R. McMillan, Jr. Some contractible open 3-manifolds. *Trans. Amer. Math. Soc.*, 102:373–382, 1962.

[Mel66] Z. A. Melzak. Infinite packings of disks. *Canad. J. Math.*, 18:838–852, 1966.

[Men26] Karl Menger. Über umfassendste n-dimensional Mengen. *Proc. Akad. Amsterdam*, 29:1125–1128, 1926.

[Men28] Karl Menger. *Dimensionstheorie*. Teubner, Leipzig, 1928.

[Men15] Elliott Mendelson. *Introduction to mathematical logic*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, sixth edition, 2015.

[Mic51] Ernest Michael. Topologies on spaces of subsets. *Trans. Amer. Math. Soc.*, 71:152–182, 1951.

[Mik54] J. Mikusiński. Sur la méthode d'approximation de Newton. *Ann. Polon. Math.*, 1:184–194, 1954.

[Mil66] Kenneth S. Miller. *An introduction to the calculus of finite differences and difference equations*. Dover Publications, Inc., New York, 1966.

[Mil78] John Milnor. Analytic proofs of the "hairy ball theorem" and the Brouwer fixed-point theorem. *Amer. Math. Monthly*, 85(7):521–524, 1978.

[Mil97] John W. Milnor. *Topology from the differentiable viewpoint*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Based on notes by David W. Weaver, Revised reprint of the 1965 original.

[Mir56] H. Mirkil. Differentiable functions, formal power series, and moments. *Proc. Amer. Math. Soc.*, 7:650–652, 1956.

[Mit65] Barry Mitchell. *Theory of categories*. Pure and Applied Mathematics, Vol. XVII. Academic Press, New York-London, 1965.

[ML92] H. Movahedi-Lankarani. On the inverse of Mañé's projection. *Proc. Amer. Math. Soc.*, 116(2):555–560, 1992.

[ML98] Saunders Mac Lane. *Categories for the working mathematician*, volume 5 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1998.

[Moi49] Edwin E. Moise. Grille decomposition and convexification theorems for compact metric locally connected continua. *Bull. Amer. Math. Soc.*, 55:1111–1121, 1949.

[Moi90] Edwin E. Moise. *Elementary geometry from an advanced standpoint*. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA, third edition, 1990.

[Moo82] Gregory H. Moore. *Zermelo's axiom of choice*, volume 8 of *Studies in the History of Mathematics and Physical Sciences*. Springer-Verlag, New York, 1982. Its origins, development, and influence.

[Mor39] Anthony P. Morse. The behavior of a function on its critical set. *Ann. of Math. (2)*, 40(1):62–70, 1939.

[Mor40] C. B. Morrey, Jr. Functions of several variables and absolute continuity, II. *Duke Math. J.*, 6:187–215, 1940.

[Mor47] Anthony P. Morse. Perfect blankets. *Trans. Amer. Math. Soc.*, 61:418–442, 1947.

[Mor54a] Dietrich Morgenstern. Unendlich oft differenzierbare nicht-analytische Funktionen. *Math. Nachr.*, 12:74, 1954.

[Mor54b] Kiiti Morita. Normal families and dimension theory for metric spaces. *Math. Ann.*, 128:350–362, 1954.

[Mor93] Robert Edouard Moritz. *Memorabilia mathematica*. MAA Spectrum. Mathematical Association of America, Washington, DC, 1993. The philomath's quotation book, 1140 anecdotes, aphorisms and passages by famous mathematicians, scientists and writers, Reprint of the 1914 original.

[Mor96] Marston Morse. *The calculus of variations in the large*, volume 18 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 1996. Reprint of the 1932 original.

[Mor09] Frank Morgan. *Geometric measure theory*. Elsevier/Academic Press, Amsterdam, fourth edition, 2009. A beginner's guide.

[Mos80] Yiannis N. Moschovakis. *Descriptive set theory*, volume 100 of *Studies in Logic and the Foundations of Mathematics*. North-Holland Publishing Co., Amsterdam, 1980.

[MQ94] T. S. Michael and T. Quint. Sphere of influence graphs: edge density and clique size. *Math. Comput. Modelling*, 20(7):19–24, 1994.

[MR68] R. M. F. Moss and G. T. Roberts. A creeping lemma. *Amer. Math. Monthly*, 75:649–652, 1968.

[MTW73] Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler. *Gravitation*. W. H. Freeman and Co., San Francisco, Calif., 1973.

[Mun66] James R. Munkres. *Elementary differential topology*, volume 1961 of *Lectures given at Massachusetts Institute of Technology, Fall*. Princeton University Press, Princeton, N.J., 1966.

[Mun75] James R. Munkres. *Topology: a first course*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1975.

[Mun84] James R. Munkres. *Elements of algebraic topology*. Addison-Wesley Publishing Company, Menlo Park, CA, 1984.

[MW99] John N. McDonald and Neil A. Weiss. *A course in real analysis*. Academic Press Inc., San Diego, CA, 1999. Biographies by Carol A. Weiss.

[MZ52] Deane Montgomery and Leo Zippin. Small subgroups of finite-dimensional groups. *Ann. of Math. (2)*, 56:213–241, 1952.

[MZ74] Deane Montgomery and Leo Zippin. *Topological transformation groups*. Robert E. Krieger Publishing Co., Huntington, N.Y., 1974. Reprint of the 1955 original.

[Nag65] Jun-iti Nagata. *Modern dimension theory*. Bibliotheca Mathematica, Vol. VI. Edited with the cooperation of the "Mathematisch Centrum" and the "Wiskundig Genootschap" at Amsterdam. Interscience Publishers John Wiley & Sons, Inc., New York, 1965.

[Nag70] Keiô Nagami. *Dimension theory*. With an appendix by Yukihiro Kodama. Pure and Applied Mathematics, Vol. 37. Academic Press, New York-London, 1970.

[Nat61] I. P. Natanson. *Theory of functions of a real variable. Vol. I, II*. Translated from the Russian by Leo F. Boron. Frederick Ungar Publishing Co., New York, 1961.

[New88] *The world of mathematics. Vol. 1–4*. Tempus Books, Redmond, WA, 1988. A small library of the literature of mathematics from A'h-mosè the Scribe to Albert Einstein, Presented with commentaries and notes by James R. Newman, Revised reprint of the 1956 original.

[Niv47] Ivan Niven. A simple proof that π is irrational. *Bull. Amer. Math. Soc.*, 53:509, 1947.

[Niv56] Ivan Niven. *Irrational numbers*. The Carus Mathematical Monographs, No. 11. The Mathematical Association of America. Distributed by John Wiley and Sons, Inc., New York, N.Y., 1956.

[NN01] Ernest Nagel and James R. Newman. *Gödel's proof*. New York University Press, New York, revised edition, 2001. Edited and with a new foreword by Douglas R. Hofstadter.

[Nyi84] Peter Nyikos. The theory of nonmetrizable manifolds. In *Handbook of set-theoretic topology*, pages 633–684. North-Holland, Amsterdam, 1984.

[Okr21] Arika Okrent. *Highly Irregular*. Oxford University Press, 2021.

[Osg97] William F. Osgood. Non-Uniform Convergence and the Integration of Series Term by Term. *Amer. J. Math.*, 19(2):155–190, 1897.

[Osg00] W. F. Osgood. Zweite Note über analytische Functionen mehrerer Veränderlichen. *Math. Ann.*, 53(3):461–464, 1900.

[Osg03] William F. Osgood. A Jordan curve of positive area. *Trans. Amer. Math. Soc.*, 4(1):107–112, 1903.

[Oss78] Robert Osserman. The isoperimetric inequality. *Bull. Amer. Math. Soc.*, 84(6):1182–1238, 1978.

[Oxt71] John C. Oxtoby. *Measure and category. A survey of the analogies between topological and measure spaces.* Springer-Verlag, New York, 1971. Graduate Texts in Mathematics, Vol. 2.

[Oxt61] John C. Oxtoby. Cartesian products of Baire spaces. *Fund. Math.*, 49:157–166, 1960/61.

[Par62] L. A. Pars. *An introduction to the calculus of variations.* John Wiley & Sons Inc., New York, 1962.

[Par05] K. R. Parthasarathy. *Introduction to probability and measure*, volume 33 of *Texts and Readings in Mathematics*. Hindustan Book Agency, New Delhi, 2005. Corrected reprint of the 1977 original.

[Pas70] B. A. Pasynkov. Bicompacta with distinct dimensions. *Dokl. Akad. Nauk SSSR*, 192:503–506, 1970.

[Pea75] A. R. Pears. *Dimension theory of general spaces.* Cambridge University Press, Cambridge, England, 1975.

[Ped89] Gert K. Pedersen. *Analysis now*, volume 118 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1989.

[Pel77] M. J. Pelling. Classroom Notes: Formulae for the Arc-Length of a Curve in R^N. *Amer. Math. Monthly*, 84(6):465–467, 1977.

[Pen05] Roger Penrose. *The road to reality.* Alfred A. Knopf, Inc., New York, 2005. A complete guide to the laws of the universe.

[Pét38] H. Pétard. A Contribution to the Mathematical Theory of Big Game Hunting. *Amer. Math. Monthly*, 45(7):446–447, 1938.

[Phe66] Robert R. Phelps. *Lectures on Choquet's theorem.* D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1966.

[Phe74] R. R. Phelps. Dentability and extreme points in Banach spaces. *J. Functional Analysis*, 17:78–90, 1974.

[Pie01] Jean-Paul Pier. *Mathematical analysis during the 20th century.* Oxford University Press, New York, 2001.

[Pod77] Klaus-Peter Podewski. Topologisierung algebraischer Strukturen. *Rev. Roumaine Math. Pures Appl.*, 22(9):1283–1290, 1977.

[Poi96] Henri Poincaré. *Science and method.* Key Texts: Classic Studies in the History of Ideas. Thoemmes Press, Bristol, 1996. Translated by Francis Maitland, With a preface by Bertrand Russell, Reprint of the 1914 edition.

[Pól90a] G. Pólya. *Mathematics and plausible reasoning. Vol. I.* Princeton University Press, Princeton, NJ, 1990. Induction and analogy in mathematics, Reprint of the 1954 original.

[Pól90b] G. Pólya. *Mathematics and plausible reasoning. Vol. II.* Princeton University Press, Princeton, NJ, second edition, 1990. Patterns of plausible inference.

[Pom92] Ch. Pommerenke. *Boundary behaviour of conformal maps*, volume 299 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1992.

[Pon63] V. I. Ponomarev. Paracompacta: their projection spectra and continuous mappings. *Mat. Sb. (N.S.)*, 60 (102):89–119, 1963.

[Pot04] Michael Potter. *Set theory and its philosophy*. Oxford University Press, New York, 2004. A critical introduction.

[Pou61] Ebbe Thue Poulsen. A simplex with dense extreme points. *Ann. Inst. Fourier (Grenoble)*, 11:83–87, XIV, 1961.

[Pre83] D. Preiss. Dimension of metrics and differentiation of measures. In *General topology and its relations to modern analysis and algebra, V (Prague, 1981)*, volume 3 of *Sigma Ser. Pure Math.*, pages 565–568. Heldermann, Berlin, 1983.

[PS92] Harold R. Parks and Richard M. Schori. A Jordan arc in $\mathbf{R}^m$ with positive m-dimensional Lebesgue measure. In *The problem of Plateau*, pages 229–236. World Sci. Publishing, River Edge, NJ, 1992.

[Pug15] Charles C. Pugh. *Real mathematical analysis*. Undergraduate Texts in Mathematics. Springer, Cham, second edition, 2015.

[Ram01] A. G. Ramm. A simple proof of the Fredholm Alternative and a characterization of Fredholm operators. *Amer. Math. Monthly*, 108:855–860, 2001.

[Ray02] Jean Saint Raymond. Local inversion for differentiable functions and the Darboux property. *Mathematika*, 49(1-2):141–158 (2004), 2002.

[Red87] Pietro Redondi. *Galileo: heretic*. Princeton University Press, Princeton, NJ, 1987. Translated from the Italian by Raymond Rosenthal.

[Rei48] E. R. Reifenberg. A problem on circles. *Math. Gaz.*, 32:290–292, 1948.

[Rei62] M. Reichbach. The power of topological types of some classes of 0-dimensional sets. *Proc. Amer. Math. Soc.*, 13:17–23, 1962.

[Rei70] Constance Reid. *Hilbert*. With an appreciation of Hilbert's mathematical work by Hermann Weyl. Springer-Verlag, New York, 1970.

[Rem91] Reinhold Remmert. *Theory of complex functions*, volume 122 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991. Translated from the second German edition by Robert B. Burckel, Readings in Mathematics.

[Rem98] Reinhold Remmert. *Classical topics in complex function theory*, volume 172 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998. Translated from the German by Leslie Kay.

[Rib92] Paulo Ribenboim. Fields: algebraically closed and others. *Manuscripta Math.*, 75(2):115–150, 1992.

[Rib00a] Paulo Ribenboim. *My numbers, my friends*. Springer-Verlag, New York, 2000. Popular lectures on number theory.

[Rib00b] Paulo Ribenboim. *My numbers, my friends*. Springer-Verlag, New York, 2000. Popular lectures on number theory.

[Ric63] Ian Richards. On the classification of noncompact surfaces. *Trans. Amer. Math. Soc.*, 106:259–269, 1963.

[Rit48] Joseph Fels Ritt. *Integration in Finite Terms. Liouville's Theory of Elementary Methods*. Columbia University Press, New York, N. Y., 1948.

[Rob66] Abraham Robinson. *Non-standard analysis*. North-Holland Publishing Co., Amsterdam, 1966.

[Rob77] James W. Roberts. A compact convex set with no extreme points. *Studia Math.*, 60(3):255–266, 1977.

[Rob06] Siobhan Roberts. *King of infinite space*. Walker and Company, New York, 2006. Donald Coxeter, the man who saved geometry, With a foreword by Douglas R. Hofstadter.

[Rog70] C. A. Rogers. *Hausdorff measures*. Cambridge University Press, London, 1970.

[Rog80] C. A. Rogers. A less strange version of Milnor's proof of Brouwer's fixed-point theorem. *Amer. Math. Monthly*, 87(7):525–527, 1980.

[Rol76] Dale Rolfsen. *Knots and links*. Publish or Perish, Inc., Berkeley, Calif., 1976. Mathematics Lecture Series, No. 7.

[Ros72] Maxwell Rosenlicht. Integration in finite terms. *Amer. Math. Monthly*, 79:963–972, 1972.

[Roy68] Prabir Roy. Nonequality of dimensions for metric spaces. *Trans. Amer. Math. Soc.*, 134:117–132, 1968.

[Roy88] H. L. Royden. *Real analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[RR85] Herman Rubin and Jean E. Rubin. *Equivalents of the axiom of choice. II*, volume 116 of *Studies in Logic and the Foundations of Mathematics*. North-Holland Publishing Co., Amsterdam, 1985.

[RS72] C. P. Rourke and B. J. Sanderson. *Introduction to piecewise-linear topology*. Springer-Verlag, New York, 1972. Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 69.

[RS09] Eleanor Robson and Jacqueline Stedall, editors. *The Oxford handbook of the history of mathematics*. Oxford University Press, Oxford, 2009.

[RSN90] Frigyes Riesz and Béla Sz.-Nagy. *Functional analysis*. Dover Publications Inc., New York, 1990. Translated from the second French edition by Leo F. Boron, Reprint of the 1955 original.

[Ruc82] Rudy Rucker. *Infinity and the mind*. Birkhäuser Boston, Mass., 1982. The science and philosophy of the infinite.

[Rud69] Mary Ellen Rudin. A new proof that metric spaces are paracompact. *Proc. Amer. Math. Soc.*, 20:603, 1969.

[Rud76] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York, third edition, 1976. International Series in Pure and Applied Mathematics.

[Rud87] Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[Rus73] T. Benny Rushing. *Topological embeddings*. Academic Press, New York, 1973. Pure and Applied Mathematics, Vol. 52.

[Rus92] T. B. Rushing. Hausdorff dimension of wild fractals. *Trans. Amer. Math. Soc.*, 334(2):597–613, 1992.

[Ś58] S. Świerczkowski. On a free group of rotations of the Euclidean space. *Nederl. Akad. Wetensch. Proc. Ser. A 61 = Indag. Math.*, 20:376–378, 1958.

[Sak64] Stanisław Saks. *Theory of the integral*. Second revised edition. English translation by L. C. Young. With two additional notes by Stefan Banach. Dover Publications Inc., New York, 1964.

[Sar42] Arthur Sard. The measure of the critical values of differentiable maps. *Bull. Amer. Math. Soc.*, 48:883–890, 1942.

[Sav94] Richard P. Savage, Jr. The paradox of nontransitive dice. *Amer. Math. Monthly*, 101(5):429–436, 1994.

[SB10] Khodr Shamseddine and Martin Berz. Analysis on the Levi-Civita field, a brief overview. In *Advances in p-adic and non-Archimedean analysis*, volume 508 of *Contemp. Math.*, pages 215–237. Amer. Math. Soc., Providence, RI, 2010.

[Sch88] H. A. Schwarz. Über ein die Flächen kleinsten flächeninhalts betreffendes Problem der Variationsrechnung. *Acta Soc. Sci. Fennicae*, XV:318–355, 1888. Festschrift zum Jubelgeburtstag des Herrn Karl Weierstrass, 1885.

[Sch52] J. A. Schatz. Review of "On a theorem of Gelfand and Neumark and the B*-algebra," by M. Fukamiya. *Math. Reviews*, MR0054176 (14,884a), 1952.

[Sch72] Horst Schubert. *Categories*. Springer-Verlag, New York-Heidelberg, 1972. Translated from the German by Eva Gray.

[Sch97] Eric Schechter. *Handbook of analysis and its foundations*. Academic Press Inc., San Diego, CA, 1997.

[Sco61] Dana Scott. Measurable cardinals and constructible sets. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys.*, 9:521–524, 1961.

[Seg47] I. E. Segal. Irreducible representations of operator algebras. *Bull. Amer. Math. Soc.*, 53:73–88, 1947.

[Sei00] Charles Seife. *Zero*. Penguin Books, New York, 2000. The biography of a dangerous idea.

[SG77] G. E. Shilov and B. L. Gurevich. *Integral, measure and derivative: a unified approach*. Dover Publications Inc., New York, english edition, 1977. Translated from the Russian and edited by Richard A. Silverman, Dover Books on Advanced Mathematics.

[Sha05] Stewart Shapiro, editor. *The Oxford handbook of philosophy of mathematics and logic*. Oxford Handbooks in Philosophy. Oxford University Press, Oxford, 2005.

[She68] R. B. Sher. Concerning wild Cantor sets in E^3. *Proc. Amer. Math. Soc.*, 19:1195–1200, 1968.

[She84] Saharon Shelah. Can you take Solovay's inaccessible away? *Israel J. Math.*, 48(1):1–47, 1984.

[She18] David Sherry. The Jesuits and the method of indivisibles. *Found. Sci.*, 23(2):367–392, 2018.

[Shi89] Andrei Borisovich Shidlovskii. *Transcendental numbers*, volume 12 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1989. Translated from the Russian by Neal Koblitz, With a foreword by W. Dale Brownawell.

[Sie49] Carl Ludwig Siegel. *Transcendental Numbers*. Annals of Mathematics Studies, no. 16. Princeton University Press, Princeton, N. J., 1949.

[Sie65] Wacław Sierpiński. *Cardinal and ordinal numbers*. Second revised edition. Monografie Matematyczne, Vol. 34. Państowe Wydawnictwo Naukowe, Warsaw, 1965.

[Sil71] Jack Silver. The consistency of the GCH with the existence of a measurable cardinal. In *Axiomatic Set Theory (Proc. Sympos. Pure Math., Vol. XIII, Part I, Univ. California, Los Angeles, Calif., 1967)*, pages 391–395. Amer. Math. Soc., Providence, R.I., 1971.

[Sin03] Yakov Sinai, editor. *Russian mathematicians in the 20th century*. World Scientific Publishing Co. Inc., River Edge, NJ, 2003.

[Sin13] Simon Singh. *The Simpsons and their mathematical secrets*. Bloomsbury Press, New York, 2013.

[Sko86] Richard Skora. Cantor sets in S^3 with simply connected complements. *Topology Appl.*, 24(1-3):181–188, 1986. Special volume in honor of R. H. Bing (1914–1986).

[SL58] Curt Schmieden and Detlef Laugwitz. Eine Erweiterung der Infinitesimalrechnung. *Math. Z.*, 69:1–39, 1958.

[Smu92] Raymond M. Smullyan. *Gödel's incompleteness theorems*, volume 19 of *Oxford Logic Guides*. The Clarendon Press Oxford University Press, New York, 1992.

[Sob63] S. L. Sobolev. *Applications of functional analysis in mathematical physics*. Translations of Mathematical Monographs, Vol. 7. American Mathematical Society, Providence, R.I., 1963. Translated from the Russian by F. E. Browder.

[Sok11] Alan D. Sokal. A ridiculously simple and explicit implicit function theorem. *Sém. Lothar. Combin.*, 61A:Art. B61Ad, 21, 2009/11.

[Sol70] Robert M. Solovay. A model of set-theory in which every set of reals is Lebesgue measurable. *Ann. of Math. (2)*, 92:1–56, 1970.

[Sol71] Robert M. Solovay. Real-valued measurable cardinals. In *Axiomatic set theory (Proc. Sympos. Pure Math., Vol. XIII, Part I, Univ. California, Los Angeles, Calif., 1967)*, pages 397–428. Amer. Math. Soc., Providence, R.I., 1971.

[Spa81] Edwin H. Spanier. *Algebraic topology*. Springer-Verlag, New York, 1981. Corrected reprint.

[Spi65] Michael Spivak. *Calculus on manifolds. A modern approach to classical theorems of advanced calculus*. W. A. Benjamin, Inc., New York-Amsterdam, 1965.

[Spi79] Michael Spivak. *A comprehensive introduction to differential geometry. Vol. I*. Publish or Perish Inc., Wilmington, Del., second edition, 1979.

[SS95] Lynn Arthur Steen and J. Arthur Seebach, Jr. *Counterexamples in topology*. Dover Publications, Inc., Mineola, NY, 1995. Reprint of the second (1978) edition.

[SS03] Elias M. Stein and Rami Shakarchi. *Complex analysis*. Princeton Lectures in Analysis, II. Princeton University Press, Princeton, NJ, 2003.

[SS13] Juan Souto and Matthew Stover. A Cantor set with hyperbolic complement. *Conform. Geom. Dyn.*, 17:58–67, 2013.

[ST80] Herbert Seifert and William Threlfall. *Seifert and Threlfall: a textbook of topology*, volume 89 of *Pure and Applied Mathematics*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980. Translated from the German edition of 1934 by Michael A. Goldman, With a preface by Joan S. Birman, With "Topology of 3-dimensional fibered spaces" by Seifert, Translated from the German by Wolfgang Heil.

[Šta71] M. A. Štan′ko. Solution of Menger's problem in the class of compacta. *Dokl. Akad. Nauk SSSR*, 201:1299–1302, 1971.

[Sta78] Harold M. Stark. *An introduction to number theory*. MIT Press, Cambridge, Mass., 1978.

[Ste66] A. K. Steiner. The lattice of topologies: Structure and complementation. *Trans. Amer. Math. Soc.*, 122:379–398, 1966.

[Ste83] Shlomo Sternberg. *Lectures on differential geometry*. Chelsea Publishing Co., New York, second edition, 1983. With an appendix by Sternberg and Victor W. Guillemin.

[Ste04] J. Michael Steele. *The Cauchy-Schwarz master class*. MAA Problem Books Series. Mathematical Association of America, Washington, DC; Cambridge University Press, Cambridge, 2004. An introduction to the art of mathematical inequalities.

[Ste15] Ian Stewart. *Professor Stewart's incredible numbers*. Basic Books, New York, 2015.

[Sti22] John Stillwell. *The story of proof—logic and the history of mathematics*. Princeton University Press, Princeton, NJ, 2022.

[Sto48] A. H. Stone. Paracompactness and product spaces. *Bull. Amer. Math. Soc.*, 54:977–982, 1948.

[Str81] Karl R. Stromberg. *Introduction to classical real analysis*. Wadsworth International, Belmont, Calif., 1981. Wadsworth International Mathematics Series.

[Str87] Dirk J. Struik. *A concise history of mathematics*. Dover Publications Inc., New York, fourth edition, 1987.

[Str13] Steven Strogatz. *The joy of x*. Mariner Books, Boston, MA, 2013. A guided tour of math, from one to infinity, Reprint of the 2012 edition.

[Sul94] John M. Sullivan. Sphere packings give an explicit bound for the Besicovitch covering theorem. *J. Geom. Anal.*, 4(2):219–231, 1994.

[SW99] H. H. Schaefer and M. P. Wolff. *Topological vector spaces*, volume 3 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1999.

[Swa94] Charles Swartz. *Measure, integration and function spaces*. World Scientific Publishing Co. Inc., River Edge, NJ, 1994.

[Swi61] William C. Swift. Classroom Notes: Simple Constructions of Nondifferentiable Functions and Space-Filling Curves. *Amer. Math. Monthly*, 68(7):653–655, 1961.

[SZ55] Helmut Salzmann and Karl Zeller. Singularitäten unendlich oft differenzierbarer Funktionen. *Math. Z.*, 62:354–367, 1955.

[Tay06] Michael E. Taylor. *Measure theory and integration*, volume 76 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006.

[Tho92] Carsten Thomassen. The Jordan-Schönflies theorem and the classification of surfaces. *Amer. Math. Monthly*, 99(2):116–130, 1992.

[Tho03] Simon Thomas. The classification problem for torsion-free abelian groups of finite rank. *J. Amer. Math. Soc.*, 16(1):233–258, 2003.

[Thu97] William P. Thurston. *Three-dimensional geometry and topology. Vol. 1*, volume 35 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1997. Edited by Silvio Levy.

[Tie23] Heinrich Tietze. Beiträge zur allgemeinen Topologie. I. *Math. Ann.*, 88(3-4):290–312, 1923.

[Tru97] J. K. Truss. *Foundations of mathematical analysis*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, 1997.

[Tuc46] A. W. Tucker. Some topological properties of disk and sphere. In *Proc. First Canadian Math. Congress, Montreal, 1945*, pages 285–309. University of Toronto Press, Toronto, 1946.

[Tuk40] John W. Tukey. *Convergence and Uniformity in Topology*. Annals of Mathematics Studies, No. 2. Princeton University Press, Princeton, N. J., 1940.

[TW97] F. C. Tinsley and David G. Wright. Some contractible open manifolds and coverings of manifolds in dimension three. *Topology Appl.*, 77(3):291–301, 1997.

[Tym86] Thomas Tymoczko, editor. *New directions in the philosophy of mathematics*. Birkhäuser Boston Inc., Boston, MA, 1986. An anthology.

[Ury25a] Paul Urysohn. Über die Mächtigkeit der zusammenhängenden Mengen. *Math. Ann.*, 94(1):262–295, 1925.

[Ury25b] Paul Urysohn. Zum Metrisationsproblem. *Math. Ann.*, 94(1):309–315, 1925.

[Usp86] V. V. Uspenskiĭ. A universal topological group with a countable basis. *Funktsional. Anal. i Prilozhen.*, 20(2):86–87, 1986.

[Var98] Ilan Vardi. Archimedes' cattle problem. *Amer. Math. Monthly*, 105(4):305–319, 1998.

[vB04] Bruce van Brunt. *The calculus of variations*. Universitext. Springer-Verlag, New York, 2004.

[vD85] Eric K. van Douwen. Horrors of topology without AC: a nonnormal orderable space. *Proc. Amer. Math. Soc.*, 95(1):101–105, 1985.

[vdW30] B. L. van der Waerden. Ein einfaches Beispiel einer nicht-differenzierbaren stetigen Funktion. *Math. Z.*, 32(1):474–475, 1930.

[Vil16] Cédric Villani. *Birth of a theorem*. Farrar, Straus and Giroux, New York, 2016. A mathematical adventure, Translated from the 2012 French original by Malcolm DeBevoise, With illustrations by Claude Gondard.

[vM84] Jan van Mill. An introduction to $\beta\omega$. In *Handbook of set-theoretic topology*, pages 503–567. North-Holland, Amsterdam, 1984.

[vM89] J. van Mill. *Infinite-dimensional topology*, volume 43 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, 1989. Prerequisites and introduction.

[vM01] Jan van Mill. *The infinite-dimensional topology of function spaces*, volume 64 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, 2001.

[vN30] J. v. Neumann. Allgemeine Eigenwerttheorie Hermitescher Funktionaloperatoren. *Math. Ann.*, 102(1):49–131, 1930.

[vN33] J. von Neumann. Die Einführung analytischer Parameter in topologischen Gruppen. *Ann. of Math. (2)*, 34(1):170–190, 1933.

[vR68] A. C. M. van Rooij. The lattice of all topologies is complemented. *Canadian J. Math.*, 20:805–807, 1968.

[vRS82] A. C. M. van Rooij and W. H. Schikhof. *A second course on real functions*. Cambridge University Press, Cambridge, 1982.

[Wac49] Sylvain Wachs. Contribution à l'étude de l'irrationalité de certains nombres. *Bull. Sci. Math. (2)*, 73:77–95, 1949.

[Wag93] Stan Wagon. *The Banach-Tarski paradox*. Cambridge University Press, Cambridge, 1993. With a foreword by Jan Mycielski, Corrected reprint of the 1985 original.

[Wal48] H. S. Wall. *Analytic Theory of Continued Fractions*. D. Van Nostrand Company, Inc., New York, N. Y., 1948.

[Wal60] Andrew H. Wallace. Modifications and cobounding manifolds. *Canad. J. Math.*, 12:503–528, 1960.

[Wee02] Jeffrey R. Weeks. *The shape of space*, volume 249 of *Monographs and Textbooks in Pure and Applied Mathematics*. Marcel Dekker, Inc., New York, second edition, 2002.

[Wei37] A. Weil. *Sur les espaces à structure uniforme et sur la topologie générale*. Actualités Scientifiques et Industrielles [Current Scientific and Industrial Topics], No.551. Hermann & Cie, Paris, 1937.

[Wei62] André Weil. *Foundations of algebraic geometry*. American Mathematical Society, Providence, R.I., 1962.

[Wei76] Clifford E. Weil. On nowhere monotone functions. *Proc. Amer. Math. Soc.*, 56:388–389, 1976.

[Wei07] André Weil. *Number theory*. Modern Birkhäuser Classics. Birkhäuser Boston, Inc., Boston, MA, 2007. An approach through history from Hammurapi to Legendre, Reprint of the 1984 edition.

[Whi35] Hassler Whitney. A function not constant on a connected set of critical points. *Duke Math. J.*, 1(4):514–517, 1935.

[Why50] G. T. Whyburn. Open and closed mappings. *Duke Math. J.*, 17:69–74, 1950.

[Wil]

[Wil04] Stephen Willard. *General topology*. Dover Publications, Inc., Mineola, NY, 2004. Reprint of the 1970 original [Addison-Wesley, Reading, MA; MR0264581].

[WR67] Edmund Whittaker and G. Robinson. *The calculus of observations: An introduction to numerical analysis*. Fourth edition. Dover Publications, Inc., New York, 1967.

[Wun52] W. Wunderlich. Eine überall stetige und nirgends differenzierbare Funktion. *Elemente der Math.*, 7:73–79, 1952.

[Wyl64] C. R. Wylie, Jr. *Foundations of geometry*. McGraw-Hill Book Co., New York, 1964.

[Yam53a] Hidehiko Yamabe. A generalization of a theorem of Gleason. *Ann. of Math. (2)*, 58:351–365, 1953.

[Yam53b] Hidehiko Yamabe. On the conjecture of Iwasawa and Gleason. *Ann. of Math. (2)*, 58:48–54, 1953.

[Yan02] Benjamin H. Yandell. *The honors class*. A K Peters Ltd., Natick, MA, 2002. Hilbert's problems and their solvers.

[Yom83] Y. Yomdin. The geometry of critical and near-critical values of differentiable mappings. *Math. Ann.*, 264(4):495–515, 1983.

[Yor70] James A. Yorke. A continuous differential equation in Hilbert space without existence. *Funkcial. Ekvac.*, 13:19–21, 1970.

[Yos80] Kôsaku Yosida. *Functional analysis*, volume 123 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin-New York, sixth edition, 1980.

[You09] W. H. Young. On Implicit Functions and Their Differentials. *Proc. London Math. Soc.*, S2-7(1):397, 1909.

[Zah47] Zygmunt Zahorski. Sur l'ensemble des points singuliers d'une fonction d'une variable réelle admettant les dérivées de tous les ordres. *Fund. Math.*, 34:183–245, 1947.

[ZD07] Smilka Zdravkovska and Peter L. Duren, editors. *Golden years of Moscow mathematics*, volume 6 of *History of Mathematics*. American Mathematical Society, Providence, RI; London Mathematical Society, London, second edition, 2007.

[Zor09] Vladimir A. Zorich. *Mathematical analysis I, II*. Universitext. Springer-Verlag, Berlin, 2009. Translated from the 2002 fourth Russian edition by Roger Cooke, Paperback reprint [of MR2033094].

Index

$AP(\mathbb{R})$, 2602
$A \otimes \mathbb{K}$, 2523
A^{-1}, 2447
A_+, 2507
A_o^{-1}, 2447
A_{sa}, 2497
$C_0(X)$, 2496
$M_\infty(A)$, 2523
$M_n(A)$, 2523
$W(T)$, 2451
$\mathrm{GL}_1(A)$, 2447
$\mathrm{GL}_1(A)_o$, 2447
$\mathrm{Irr}_{\mathcal{H}}(A)$, 2522
$\mathbb{K}$, 2495
Λ_I, 2510
$\mathrm{Prim}(A)$, 2520
$\mathrm{Prim}^J(A)$, 2521
$\mathrm{Prim}_J(A)$, 2521
$\mathbb{M}_n$, 2495
$\mathcal{H}^\infty$, 2422
$\mathcal{H}^n$, 2422
$\mathcal{H}_\phi$, 2517
$\mathcal{P}(A)$, 2516
$\mathcal{S}(A)$, 2514
$\check{x}$, 2521
ℓ^2, 2405
$\ell^2(S)$, 2405
$\hat{A}$, 2522
π_ϕ, 2517
$\sigma(T)$, 2449
$\sigma_A(x)$, 2437
ξ_ϕ, 2518
e^x, 2447
$exp(A)$, 2447
$w(T)$, 2451
x_+, 2507
x_-, 2507

Abel's Test, 623
Abel's Theorem, 861
Abel, Niels Henrik, 354, 2621
abelian category, 279
abelian group, 354, 2565
abelian semigroup, 354, 2565
absolute G_δ, 1542
absolute neighborhood retract, 1871
absolute retract, 1865
absolute value, 298
absolutely continuous function, 2245
absolutely convergent infinite series, 611
absorbing set, 2293
abstract simplicial complex, 1859
abstraction, 16
AC, 195
accumulation point, 1444
    of a subset of $\mathbb{R}$, 903
adaptive procedure, 1184
Addison, John, 1968
additivity
    countable, 1970
    finite, 1971
adele, 1639
a.e., 1977
affine map
    of convex sets, 2295
affinely independent, 494
Al-Khwarizmi, 2615
Alaoglu's Theorem, 2352
Aleksandrov, Pavel, 1967
aleph, 246

Alexander horned sphere, 1747
algebra
  graded, 529
  of functions, 1948
  of subsets, 1920
  tensor, 528
algebraic number, 191, 412, 426
algebraically independent numbers, 426
almost disjoint, 1977
almost everywhere, 1148, 1977
almost periodic function, 2602
alternating series, 615
Alternating Series Test, 615
amplification
  of a Hilbert space, 2422
  of a representation, 2512
analysis, 2
  complex, 380
analytic continuation, 1390
analytic dimension, 1815
analytic function, 867
  at a point, 885
  complete, 1385
  complex, 867
  real, 867
  total, 1384
analytic function element, 1384
analytic set, 1963–1969
analytic subring, 1814
annihilator, 2601
annulus, 1343
ANR, 1871
antiderivative, 770
Antiderivative Theorem, 1315
  for convex sets, 1314
Antoine's necklace, 1751
Antoine, Louis, 1748
A-operation, 2018
applied mathematics, 83
approximate derivative, 2286
approximate limit, 2284
approximate midpoint property, 960
approximate unit, 2509
  almost idempotent, 2509
  continuous, 2509
  idempotent, 2509
  sequential, 2509
approximately continuous, 2285
approximately differentiable, 2286
approximation dimension, 1825
approximation property, 2460
AR, 1865
arc, 954, 982, 1616, 1747
arc length, 950, 972
Archimedean property, 316
Archimedes, 52, 69, 1108
Archimedes' cattle problem, 69
arcsine, 741
arcwise connected space, 1614, 1616
area
  of a circle, 745
Aristotle, 62
Armstrong, Neil, 95
Artin-Schreier Theorem, 358
Arzelà-Ascoli Theorem
  Classic, 1689
  General, 1687
ascent, 2458
Askey, Richard, 78
associative law, 293, 334
Assouad dimension, 2056
Atiyah, Michael, 83, 89, 107
Atkinson's Theorem, 2465
atom, 268, 449, 1981
atomic statement, 132
atomless measure, 268
attaching, 1524
authority, 14
autonomous differential equation, 1271
average value
  of a function, 1157
axiom, 17, 18, 291
Axiom of Choice, 195, 220, 230–231, 248
  countable, 199
  for finite sets, 201
Axiom of Constructibility, 239
Axiom of Countable Choice, 199
Axiom of Dependent Choice, 200
Axiom of Determinacy, 199
Axiom of Projective Determinacy, 199
Axiom of Regularity, 219

B*-algebra, 2495
Baby Monotone Convergence Theorem, 1973
Baby Tonelli Theorem, 1834
Baire Category Theorem, 1563, 1598
Baire class, 1952, 1959
Baire measurable function, 1601
Baire property, 1600
Baire space, 1596
Baire, René, 1597
Baker, Alan, 428
balanced hull, 2294
balanced set, 2293
Banach *-algebra, 2495
Banach algebra, 2432
Banach Fixed-Point Theorem, 926
    Extended, 928
Banach limit, 2486
Banach space
    reflexive, 2338, 2351
    strictly convex, 2427
    uniformly convex, 2427
Banach, Stefan, 265, 267, 2235, 2241
Banach-Tarski Paradox, 221
    basic version, 545, 555
    general version, 557
    popular version, 545
Barnum, Phineas T., 2624
barreled, 2325
Barwise, Jon, 133
barycenter
    of a measure, 2372
barycenter refinement, 1490
barycentric coordinate, 1851, 1858
base
    for a decimal expansion, 395
    for a topology, 1434
basic open set, 1434
Basic Ratio Test, 600
Basic Root Test, 599
basis
    Hamel, 468
    of a vector space, 468
Bauer simplex, 2379
Belk, Jim, 1746
Bell, Eric Temple, 21, 82, 86, 2668
Bellos, Alex, 12
Beltrami, Eugenio, 45
Benacerraf, Paul, 329
Beppo Levi Theorem, 2118
Berkeley, George, 685
Bernays, Paul, 208
Bernoulli, Daniel, 778
Bernoulli, Jakob (Jacques), 778, 2616
Bernoulli, Johann (Jean), 778, 2616
Bernstein polynomial, 2384
Bernstein, Felix, 190, 2025
Bernstein, Sergei, 880, 2384
Besicovitch covering, 2075
Besicovitch Covering Theorem, 2077
    Euclidean Version, 2079
Besicovitch Partition Theorem, 2075
Besicovitch, A., 891, 1698
Bessel's inequality, 2412
best approximation to an irrational number, 409
between, 930
Between Axiom, 931
Bicommutant Theorem, 2331
bicontinuous mapping, 1503
Big Picard Theorem, 1358
bilinear map, 514
bimodule, 524
binary icosahedral group, 1530
binomial coefficient, 829
Binomial Theorem, 829, 830
Bishop, Errett, 99, 2379, 2648
Bloch, André, 2624
block, 1452
block decomposition, 1825
Boas, Ralph, 202, 898
Bohr compactification, 2603
Bohr, Harald, 52, 836
Bolyai, Janos, 45
Bolzano-Weierstrass Theorem, 572, 906
    in $\mathbb{R}^n$, 989
Bombelli, Raffael, 99
Bonnet, Ossian, 759
Boole, George, 105
Boolean Prime Ideal Theorem, 200
bootstrapping, 1977, 2107
Borel function, 1927
Borel isomorphism, 1928
Borel measurable function, 1944–1952

Borel measure, 2035, 2475
finite, 2189
Borel regular outer measure, 2018
Borel set, 1938–1939, 1947–1952
Borel space, 1927
standard, 1957, 1962
Borel, Emile, 889
Borel-Cantelli Lemma, 411
First, 1975
Borromean rings, 1754
Borsuk conjecture, 1722
Bosch, Robert, 1703
bound variable, 136
boundary
of a cell, 1850
bounded
sequence, 561
set, 168, 2315
bounded above, 168
bounded below, 168
Bounded Convergence Theorem
nonnegative version, 2133
Bounded Inverse Theorem, 2359
Bounded Mean-Square Convergence Theorem, 2148
bounded operator, 2327
bounded variation norm, 2262
bounded-finite measure, 2039
Bourbaki, Nicolas, 123, 208, 1564, 2642, 2657
Bowen, Rufus, 2621
box topology, 1478
Boy's surface, 1522
BPI, 200
branch, 1305, 1385
branch cut, 1305
branch point, 1393
Briggs, Henry, 54
Brouwer Fixed-Point Theorem, 1777–1787
Brouwer, Luitzen, 98, 1705, 1706
Browder, Earl, 2627
Buddha, 14
Bukreev, Boris, 2621
bump function
smooth, 886
bundle
cotangent, 1249
tangent, 1249
Bunyakovskii Inequality, 2400
Bunyakovskii, Viktor, 2400
Burali-Forti Paradox, 227
Busch, Adolphus, 41
busy beaver function, 39

C*-algebra, 2495
abstract, 2495
stable, 2523
C*-axiom, 2495
calculus
multivariable, 987
vector, 987
cancellation, 335, 338, 352
canonical embedding, 2338, 2340, 2350
Cantor set, 1633, 1696, 1730, 1751, 1956
of positive measure, 1956
Cantor, Georg, 181, 190, 368, 434, 2624, 2655, 2668
Cantor-Bendixson Theorem, 1451
Carathéodory extension process, 1995–2021
Carathéodory Extension Theorem, 2010
Carathéodory's Theorem, 2001
Carathéodory, Constantin, 2002
Cardano, Hieronimo, 99
cardinal, 244–264
aleph, 246
continuum, 186
Dedekind-finite, 251
inaccessible, 259
κ-additive, 265
large, 260
limit, 246
Mahlo, 269
measurable, 271
Ramsey, 274
real-valued measurable, 266
regular, 259
Russell, 256
singular, 259
strongly compact, 276
strongly inaccessible, 259
successor, 246
weakly inaccessible, 259
cardinal invariant

of a topological space, 1436
Cartan, Henri, 1465, 1564, 2621
Casorati-Weierstrass Theorem, 1358
catalog, 1952
category, 278
skeletal, 284
Cauchy complete, 1668
Cauchy Criterion
for infinite series, 585
Cauchy criterion, 578
Cauchy data, 1295
Cauchy filter, 1663
Cauchy integral, 1108–1110
Cauchy Integral Formula, 1317
for derivatives, 1320
Cauchy nest, 1663
Cauchy net, 1661
Cauchy problem, 1295
Cauchy product of infinite series, 646
Cauchy sequence, 368, 578
Cauchy system, 1295
Cauchy's Condensation Test, 608
Cauchy's Inequalities, 1321
Cauchy's Test, 598
Cauchy's Theorem, 1327, 1328
for convex sets, 1316
for simply connected sets, 1329
homology version, 1329
Cauchy, Augustin-Louis, 6, 368, 664, 759, 760, 886, 1106, 2400
Cauchy-Lipschitz Theorem, 1275
Cauchy-Riemann equations, 1299
Cauchy-Schwarz Inequality, 2400
Cavalieri, Bonaventura, 1173
Cayley number, 388
CBS Inequality, 2400, 2472
CBS inequality
for states, 2515
CCC, 1846
cdf, 2189, 2259
Čech, Eduard, 1575
cell, 1849
cell complex, 1856
centered collection of subsets, 1465
Chain Rule
one variable, 698–700
chained arrow notation, 240
Chaitin, Gregory, 56, 68
Champernowne's constant, 419
Change of Variable
backward, 2186
forward, 2186
Lebesgue version, 2268
Riemann version, 1160, 1166
change-of-basis matrix, 483
Chapman, C. H., 103
character
of a locally compact abelian group, 2599
characteristic
of a field, 357
characteristic function, 1114, 1915
characteristic polynomial, 511
Chebyshev nodes, 817
Chebyshev polynomial, 819
Chebyshev quadrature, 1177
Chebyshev, Pafnuty, 816
Cheng, Eugenia, 11, 74, 103
Chernavskii, A. V., 1078
Chernoff, Paul, 1563
Chevalley, Claude, 2597
Choice, 195
Choquet's Theorem, 2376
Chrétien, Jean, 57
Church, Alonzo, 120
circular function, 746
circumference, 26
Clairaut's Theorem, 1002
complex, 1406
formal, 1427
Clarkson's Inequality, 2427
classical adjoint, 508
clopen set, 1605
closed face, 2368
Closed Graph Theorem, 2362
for locally compact groups, 2581
closed interval, 314
closed manifold, 1769
closed set, 1443
in $\mathbb{R}$, 901
closed statement, 137–139
closer and closer, 565
closure

of a subset of $\mathbb{R}$, 904
closure of a set, 1444
closure operation, 1449
cluster point
of a net, 1458
of a sequence, 1454
cobounded filter, 272
cofinal condition, 178
cofinal subset, 177
Cohen, Paul, 190
coisometry, 2497
Cole, Frank N., 139
Cole, K. C., 17
collapsing, 1515
column vector, 456
comeager set, 1596
commutative law, 293, 334
compact convergence, 665
compact group, 2579
compact manifold, 1769
compact operator, 2458
compact representation, 1575
compact space, 1437, 1554–1566
compact support, 1573
compact-open topology, 1680
compactification, 1574–1578
one-point, 1569, 1590
Stone-Čech, 1575–1578
Wallman, 1578
compactly contained, 2553
compactly embedded, 2481
compactly generated, 1583
Compactness Theorem of model theory, 200
compactum, 1554
comparable elements, 167
Comparison Test, 591, 614
Extended, 592
General Limit, 595
Limit, 594
complement, 1914
complete
Cauchy, 1668
topologically, 1662
uniform space, 1661
uniformly, 1668
complete analytic function, 1385
complete induction, 172, 302–303
complete measure space, 1976
complete topological group, 2578
completely continuous operator, 2458
completely Hausdorff space, 1499
completely normal space, 1536
completely regular space, 1491–1492
completely separated, 2343
Completeness Axiom, 168, 312, 374
completion
of a measure space, 1977
of a pre-Hilbert space, 2406
of a topological group, 2578
of an inner product space, 2405
complex
cell, 1856
simplicial, 1857
complex analysis, 380
complex differentiable, 1298
complex number, 380–388
complex plane, 382
component, 1609
composite quadrature procedure, 1172
compression
of a measure, 1983
computable function, 39
concatenation, 948, 966
concave function, 765
conditional expectation, 2179
conditionally convergent infinite series, 611
cone, 1852
over a topological space, 1516
conformal equivalence, 1372
conjecture, 67
conjugate exponent, 2470
connected
simply, 1372
connected component, 1609
connected space, 1605
connective, 130
constant, 130
Constant Multiple Rule, 694
complex, 1302
constituent set, 156
Constructive Analysis, 199
constructivism, 98–101

content, 1188
continent of mathematics, 81
continued fraction, 402–420
  convergent, 407
  finite, 403
  term, 404
continuity
  on a set in $\mathbb{R}$, 678
continuous functional calculus, 2504
continuous spectrum, 2449
continuum, 186, 1742
  hereditarily indecomposable, 1742
  indecomposable, 1742
Continuum Hypothesis, 190, 258
  Generalized, 258
contractible space, 1895
contraction
  strict, 926
  strict power, 927
Contraction Mapping Theorem, 926
contraction of an infinite series, 589
contradiction, 62–63
contrapositive, 134, 143
contravariant, 486
controlled metric, 2070
convergence
  compact, 665
  normal
    for series, 854
    of sequences, 665, 1331
  u.c., 665, 1680
convergence class, 1464
convergence test
  Abel's Test, 623
  Alternating Series Test, 615
  Basic Ratio Test, 600
  Basic Root Test, 599
  Cauchy's Condensation Test, 608
  Cauchy's Test, 598
  Comparison Test, 591, 614
    Extended, 592
    General Limit, 595
    Limit, 594
  d'Alembert's Test, 598
  Dirichlet's Test, 623
  Gauss's Test, 607
  Integral Test, 598
  Kummer's Test, 602
  Raabe's Test, 603
    Limit Version, 604
    Second Order, 606
  Ratio Test, 600, 614
  Root Test, 599, 613
  Scaling Test, 622
  Yermakov's Test, 609
convergent sequence, 562
convergent series, 583
converse, 134
convex function, 2204
convex hull, 2294
convex set, 2294
Conway, John H., 1533
coordinate vector, 481
corollary, 154
cosecant, 743
coset, 2567
cosine, 741
cotangent, 743
cotangent bundle
  in $\mathbb{R}^n$, 1249
countable additivity, 1970
Countable Axiom of Choice, 199
Countable Chain Condition, 1436, 1846
countable complement topology, 1441
countable ordered set, 189
countable set, 185
countable subadditivity, 1972, 1996
countably infinite set, 185
counterexample, 66
counting measure, 1971
Courant, Richard, 692
covariant, 486
covariant functor, 280
cover, 1437
covering
  uniform, 1656
covering dimension, 1801
Covering Theorem
  Besicovitch, 2077
Coxeter, Donald, 2657
Cramer's Rule, 510
creeping lemma, 393

critical number, 701
critical point, 701
cross-cap, 1527
cube, 1849
cumulative distribution function, 2189
    for signed Borel measures, 2259
curve
    space-filling, 1696
        ultimate, 1696
cut point, 1383
cyclic group, 346
cyclic vector, 2514

d'Alembert's Test, 598
d'Alembert, Jean, 16
da Vinci, Leonardo, 96
Daniell integral, 1117
Daniell, P. J., 1117
Darboux continuous function, 764
Darboux property, 764
Darboux sum, 1124
    interior, 1126
Darboux, Jean-Gaston, 1109
Davis, Martin, 330
Davis, Philip J., 840
DC, 200
de Leeuw, Karel, 2379
de Moivre, Abraham, 841
decadent rational number, 399
decimal expansion, 395–402
decomposable element, 515
decomposable measure, 1979
decomposition dimension, 1806
Dedekind cut, 359
Dedekind, Richard, 190, 323, 327, 359
Dedekind-finite cardinal, 251
Dedekind-finite set, 182, 251
definition, 144
deleted neighborhood
    of a point of $\mathbb{R}$, 903
δ-ring, 1923
Denjoy integral, 1118
Denjoy, Arnaud, 728
Denjoy-Young-Saks Theorem, 728
dense, 1446
dense-in-itself, 1450
density
    of a set, 2282
    of a topological space, 1447
density function, 2173
Denver, John, 74
Dependent Choice, 200
derangement, 625
derivation, 1414
derivative
    Dini, 725–729
    directional, 1004
    distributional, 2542
    higher, 703
    Lie, 1251
    one-sided, 724–725
    partial, 995
    Radon-Nikodým, 2174
    second, 703
    weak, 2542
derived set, 1450
    higher, 1452
Descartes, René, 9, 60, 79, 123
Descending Chain Condition, 220
descent, 2458
descriptive set theory, 1952
determinant, 496–510
    expansion by minors, 501–502
    geometric interpretation, 504–508
    multiplicative property, 502–503
    of a linear transformation, 504
diagonal property, 1455
diagonalization argument, 185, 1954
Dieudonné, Jean, 108, 1578, 2637, 2657
difference of sets, 1914
difference quotient, 688
differences, 716
differentiable
    function between Banach spaces, 2525
differential
    of a form, 1262
    of a function, 1261
differential 1-form
    on $\mathbb{R}^1$, 1257
differential equation, 1269
    autonomous, 1271
    ordinary, 1269

partial, 1269
differential field, 1417
differential function field, 1419
differential homomorphism, 1427
differential ideal, 1427
diffomorphism, 1427
dilation
of a measure, 1984
dim, 1801
dimension
analytic, 1815
approximation, 1825
Assouad, 2056
covering, 1801
decomposition, 1806
essential, 1792
Hausdorff, 2048
large inductive, 1798
linear (vector space), 470
local, 1807
metric, 2058
nuclear, 1805
of a cell, 1849
of a complex, 1856
orthogonal, 2415
packing, 2052
partition, 1806
small inductive, 1796
dimension-counting theorem, 478
dimensional kernel, 1823
Dini derivative, 725–729
Dini's Theorem, 1595
Dini, Ulisse, 725, 1045
direct limit
of sets, 1622–1626
of topological spaces, 1630–1631
direct sum
of Hilbert spaces, 2422
direct system, 1622
of sets, 1622
of topological spaces, 1630
directed equivalence, 178
directed function, 178
directed set, 169, 177–180
directional derivative, 1004
Dirichlet Problem, 1411
Dirichlet's Test, 623
discontinuity
jump, 2196
discrete metric, 1433
discrete topology, 1433
dishonesty, 123
disk of convergence, 1336
distribution, 2539
distributional derivative, 2542
distributive law, 293, 334
Divergence Criterion, 585
divergent sequence, 562
divergent series, 583
divided difference, 805
higher-order, 806
division ring, 388
Dixmier, Jacques, 2621
domain of discourse, 132, 147
Domain-Straightening Theorem, 1079
Dominated Convergence Theorem
nonnegative version, 2133
dominating function, 2133
dot product, 494
double sum of an array, 635
doubling metric, 2055
Dowker's Theorem, 1804
downward continuity, 1974
drastic revision
of an infinite series, 642
of an unordered sum, 643
dual basis, 484
dual group, 2599
dual space
topological, 2317
Dugundji Extension Theorem, 1879
Dunham, William, 669
dyadic rational number, 358

e, 733–735
edge, 1856
edifice of mathematics, 79
Edison, Thomas, 74
Egorov's Theorem, 1985
Egorov, Dmitri, 2655
eigenvalue, 510
eigenvector, 510

Eilenberg, Samuel, 286
Einstein, Albert, 46, 81
elementary extension, 1423
elementary function, 1423
elementary function field, 1423
endpoint maximum, 703, 727
endpoint minimum, 703, 727
Enflo, Per, 2460
entourage, 1650
envelope
  lower, 1151
  upper, 1151
epimorphism, 279
ϵ-star, 1853
equiconnected, 1868
equiconsistent, 274
equiintegrable, 2142
equipotent, 181
equivalence
  of representations, 2512
equivalence class, 164
equivalence relation, 163
Ermakoff, *see* Yermakov
error, 58
essential dimension, 1792
essential map, 1779
essential singularity, 1357
essential subspace, 2513
Euclid, 45, 308, 343
Euclid's Parallel Postulate, 45
Euclidean Four-Point Axiom
  Weak, 934
Euler, Leonhard, 111, 113, 417, 422, 427, 778, 834, 1110
Euler-Mascheroni constant, 663
even covering, 1656
evenly continuous, 1685
Eves, Howard, 51
Existence and Uniqueness Theorem
  for first-order ODEs, 1275
    global version, 1278
    local version, 1277
    standard version, 1278
  for higher-order ODEs
    global version, 1282
    local version, 1281
    standard version, 1281, 1289
  for systems of first-order ODEs
    global version, 1280
    local version, 1280
    standard version, 1280, 1289
existential quantifier, 135
exponential
  in a differential field, 1421
exponential function, 730–739
exponents
  rules of, 730
exposed point, 2382
Extended Banach Fixed-Point Theorem, 928
Extended Comparison Test, 592
Extended First-Order Transfer Principle, 448
extended metric, 917
extended real line, 324
extended simple language, 150
Extended Simple Transfer Principle, 448
extended topologist's sine curve, 1727
extension
  of a measure, 1984
Extension Axiom, 931
Extension Theorem, 2008
  for μ°, 2012
exterior algebra, 537
exterior power, 530
extrapolation, 804
Extrapolation Property, 933
extreme point, 2295
Extreme Value Theorem, 393, 682

F-space, 2311
Faà di Bruno's formula, 704, 880
Fabius function, 894
face, 1850, 2295
  closed, 2368
Face Partition Theorem, 1780
facial topology, 2380
factorial polynomial, 721, 814
faithful representation, 2593
Fatou's Lemma, 2131
FedererSardThm, 1843
Fermat's Last Theorem, 427
Fermat, Pierre de, 417
Feynman, Richard, 21, 61

field, 295, 356, 455
differential, 1417
differential function, 1419
formally real, 358
nonarchimedean ordered, 375, 437
of fractions, 358
of subsets, 1920
ordered, 297, 356, 436
quotient, 358, 376
real-closed, 323, 358, 377
skew, 388
field axioms, 295
filter, 272, 1465, 1932
Cauchy, 1663
cobounded, 272
finer, 1467
fixed, 1933
free, 1933
κ-complete, 272
neighborhood, 1465
normal, 273
prime, 1934
Filter Extension Theorem, 1933
filterbase, 1465
maximal, 1468
subordinate, 1467
fine uniformity, 1660
finite additivity, 1971
finite ascent, 2458
finite Borel measure, 2189
finite complement topology, 1435
finite descent, 2458
Finite Intersection Property, 1555, 1933
finite measure, 1978
finite set, 182
finite subadditivity, 1972, 1996
Finite-Dimensional Spectral Theorem, 2468
finitely additive measure, 1971
Finslerian length structure, 962
first category set, 1596
first countable space, 1435
First Decomposition Theorem, 1800
First Derivative Test, 762
first differences, 716
First Isomorphism Theorem of Group Theory, 2569
first-order logic, 133
first-order statement, 133
first-order theory, 133
First-Order Transfer Principle, 448
Five Lemma, 1904
fixed filter, 1933
fixed-point theorem
Banach, 926
extended, 928
Brouwer, 1777–1787
flat module, 527
flow under a constant function, 1523
folium of Descartes, 1053
foot, 935
formal Laurent series, 376
formal power series, 375
formal proof, 57
formal statement, 129–147
formalism, 101–108
Forsyth, A. R., 85
forward differences, 814
Four-Color Theorem, 112, 1727
Fourier coefficient, 2411
Fourier series, 664
Fourier, Joseph, 422
fractional linear transformation, 1381
Fraenkel, Abraham, 208
Franklin, Benjamin, 85
Fréchet, Maurice, 2010
Fréchet ordering, 234, 375
Fréchet space, 2311
Fréchet-Urysohn space, 1469
Fréchet-von Neumann-Jordan Theorem, 2402
Fredholm Alternative, 2463
Fredholm index, 2460
Fredholm operator, 2460
Fredholm, Ivar, 2457
free cyclic group, 346
free filter, 1933
free group, 551
free module, 523
free semigroup, 550
free variable, 135
Frege, Gottlob, 13, 100, 101
F_σ, 1938, 1959
Fubini's Differentiation Theorem, 1827

Fubini's Series Theorem, 2199
Fubini's Theorem, 1827
continuous form, 1210
for double summations, 639
for signed and complex measures, 2160
product form, 2156
Riemann integral form, 1210
upper and lower form, 1208
Fubini, Guido, 640, 1827, 2199
full partition, 1222
full real-valued measure, 265
Fuller, Buckminster, 87
fully normal, 1490
fully separated, 2343
function
analytic, 867
at a point, 885
arcsine, 741
Borel, 1927
Borel measurable, 1944–1952
characteristic, 1114
circular, 746
convex, 2204
cosecant, 743
cosine, 741
cotangent, 743
Darboux continuous, 764
directed, 178
exponential, 730–739
Fabius, 894
hyperbolic, 746–751
increasing, 2196
indicator, 1114
inverse hyperbolic, 748
inverse trigonometric, 744–745
locally measurable, 1943
log-convex, 2223
logarithm, 732
measurable, 1927, 1940, 2027
midpoint-convex, 2213
monotone, 2196
multiplicity, 1955
natural logarithm, 735
nondecreasing, 2196
nonincreasing, 2196
rational, 358, 374, 509
regulated, 775, 1164
secant, 743
simple, 1114, 2108
simplicial, 1857
sine, 741
step, 770
strictly increasing, 2196
strictly measurable, 1943
strictly monotone, 2196
tangent, 743
trigonometric, 740–751
functional calculus
continuous, 2504–2505
holomorphic, 2446–2447
functionally dependent, 1019
functionally Hausdorff spaces, 1499
functionally independent, 1019
functor, 280
covariant, 280
Fundamental Principle of Infinite Series
First, 582
Fundamental Principle of Nonnegative Series, 591
fundamental system of neighborhoods, 2305, 2573
Fundamental Theorem of Algebra, 386, 1307, 1398, 1786
Fundamental Theorem of Arithmetic, 308, 343
Fundamental Theorem of Calculus, 583
general HK version, 2163
HK version, 2162
Lebesgue Version 1a, 2252
Lebesgue Version 1b, 2256
Lebesgue Version 2a, 2255
Lebesgue Version 2b, 2255
Version 1, 1153
Version 2, 1156
Fundamental Theorem of Interval Length, 316
fuzzy logic, 99

galaxy, 917
Galileo, 15, 65, 182
Gall, Franz Joseph, 2622
Galois, Évariste, 2621
Gamma function, 834
gap, 58
Gardner, Martin, 427
Gauss's Lemma, 431

Gauss's Test, 607
Gauss, Carl Friedrich, 45, 67, 100, 301, 616, 745, 839
Gauss-Kronrod integration, 1184
Gaussian quadrature, 1176
G_δ, 1548, 1938, 1959
Gelfand transform, 2503
Gelfand's Theorem, 2499
Gelfand-Mazur Theorem, 2444
Gelfand-Naimark Theorem, 2519
Gelfond, Aleksandr, 427
Gelfond-Schneider Theorem, 427
General Complex Analytic Implicit/Inverse Function Theorem, 1088
General Implicit/Inverse Function Theorem, 1058, 2530
General Limit Comparison Test, 595
general position, 493
General Real Analytic Implicit/Inverse Function Theorem, 1086
generalization, 1079
Generalized Continuum Hypothesis, 258
generalized function, 2539
generic set, 1596
geometric realization, 1859
geometric series, 586
germ, 1396, 1397
Gleason, Andrew, 2597
Glöckner, Helge, 2584
gluing, 1517
gluing scheme, 1532
GNS construction, 2517–2518
GNS representation, 2517
Gödel, Kurt, 103, 185, 190, 208, 2624
Gödel constructible universe, 239
Goethe, 102
Goldblatt, Robert, 108, 436
golden ratio, 417
good approximation to an irrational number, 409
Goodstein function, 40
Gordan, Paul, 98
Gosper, R. William, 845
Goursat's Theorem, 1312
graded algebra, 529
graded commutative algebra, 537
gradient, 1017
Gram-Schmidt procedure, 2416
Grassmann algebra, 537
greatest lower bound, 168, 311, 361
Greenberg, Marvin J., 84
Gregory, James, 791, 807
Gregory-Newton interpolation formula, 814
grid, 1189
Griffin, Frank L., 90
Grothendieck group, 354
Grothendieck universe, 276
Grothendieck, Alexandre, 286, 2624, 2643, 2655
group, 346, 2565
  abelian, 354, 2565
  compact, 2579
  cyclic, 346
  dual, 2599
  free, 551
  free cyclic, 346
  Grothendieck, 354
  locally compact, 2579
  locally Euclidean, 2597
group action, 1528, 2574
  faithful, 1528
group algebra, 2592
group-homomorphism, 2566
grouping of an infinite series, 589
groupoid, 287
Gudermannian, 751

Haar measure, 2587
Hadamard, Jacques, 18, 84, 107
Hahn series, 377
Hahn, Hans, 2010, 2101
Hahn-Banach Separation Theorem, 2342, 2343
Hahn-Banach Theorem
  Complex Version 1, 2339
  Complex Version 1A, 2339
  Version 1, 2334
  Version 1A, 2334
  Version 2, 2337, 2339
Hahn-Mazurkiewicz Theorem, 1642
half-open interval, 314
Halmos, Paul, 1927
Hamel basis, 468
Hamilton, William Rowan, 388

handles, 1524
Hanf-Tarski Theorem, 271
Hardy, G. H., 63, 116
harmonic analysis, 2599
harmonic conjugate, 1408
harmonic function, 1407
harmonic series, 592
Hart, W. D., 92
Hartogs cardinal, 249
Hartogs' Theorem, 1403
Hausdorff dimension, 2048
Hausdorff Maximal Principle, 197
Hausdorff measure, 2044
    normalized, 2049
Hausdorff metric, 1550–1551
Hausdorff Paradox, 553
Hausdorff space, 1482, 1485
Hausdorff, Felix, 1938, 2365, 2620
Hausdorffization, 1512
Hawaiian earring, 1729
Heath, Thomas L., 47
Heaviside function
    smooth, 887
height
    of a topological space, 1452
Heine-Borel Theorem, 905
Helly space, 1587
Helmholtz, H., 85
Henstock-Kurzweil integral, 1118, 2161
Herbrand, Jacques, 2621
hereditarily indecomposable continuum, 1742
hereditarily normal space, 1536
hereditary quotient map, 1513
hereditary set, 359
hereditary subcollection, 1924
Hermite polynomial, 813
Hermite, Charles, 422, 427
Hersh, Reuben, 107
higher derivatives, 703
higher differences, 717
higher-order logic, 133
higher-order theory, 133
Hilbert cube, 1696
Hilbert space, 2404
    real, 2423
Hilbert, David, 45, 46, 98, 102, 181, 427, 1522, 2405
Hippasus, 62
HK integral, 2161
Hobbes, Thomas, 129, 2620, 2667
Hölder, Otto, 839
Hölder's Inequality, 2472
    generalized, 2491
Holmes, Sherlock, 54, 61, 62
holomorphic function, 1298
holomorphically simply connected, 1374
homogeneous metric, 2055
homogeneous space, 1530, 2573
homology
    of contours, 1327
homomorphism
    differential, 1427
homotopic, 1894
homotopy, 1894
    of contours, 1323
Homotopy Extension Theorem, 1873
Homotopy Theorem for Contour Integrals, 1324
honesty, 123
hull-kernel topology, 1498
Hurwitz's Theorem, 1368
Hurwitz, Adolf, 424
Huxley, Thomas H., 81
hyperbolic function, 746–751
    inverse, 748
hypergeometric series, 616–619
hypernatural numbers, 443
hyperplane, 2341
hyperreal numbers, 438
hyperspace, 1550

ideal, 205, 353, 2496
    differential, 1427
    maximal, 205
    prime, 205
    primitive, 2520
    principal, 205
identification space, 1515
if and only if, 135
image, 159, 161
imaginary part
    of an element of a C*-algebra, 2497

implication, 134–135
Implicit Function Theorem
Banach smooth version, 2529
complex analytic version, 1087
continuous version, 1047
formal, 379
real analytic version, 1086
smooth version, 1047
Implicit/Inverse Function Theorem
degenerate case, 1085
General, 1058, 2530
General Complex Analytic, 1088
General Real Analytic, 1086
impossibility theorem, 63
improper Riemann integral, 1163
inaccessible cardinal, 259
incomparable elements, 167
increasing function, 2196
Ind, 1798
ind, 1796
indecomposable continuum, 1742
indeterminate form, 325
index
Fredholm, 2460
indicator function, 1114, 1915
indiscrete topology, 1433
induction, 300–305, 331–332
complete, 172, 302–303
transfinite, 171–174, 229–230
inductive limit, 1623
of topological vector spaces, 2325
inductive limit topology, 1630
infimum, 168, 311, 361
infinite broom, 1634
infinite comb, 1759
infinite ladder, 1763
infinite product
of numbers, 648–653
infinite regress, 303–304
infinite series
of numbers, 582–663
absolutely convergent, 611
alternating, 615
complex, 653–654
conditionally convergent, 611
convergent, 583
divergent, 583
hypergeometric, 616–619
nonnegative, 590
infinite set, 182
infinitesimal, 375, 376, 434, 437
initial-value problem, 1271
inner measure, 2017
integer, 305–308, 345–355
integrable function, 2134
nonnegative, 2114
integral
Cauchy, 1108–1110
Daniell, 1117
Denjoy, 1118
Henstock-Kurzweil, 1118, 2161
improper Riemann, 1163
iterated, 1208
Lebesgue, 1112–1117, 2125
Perron, 1118
Riemann, 6–7, 1106–1107, 1111, 1121–1166
in $\mathbb{R}^n$, 1188–1221
integral domain, 355
Integral Test, 598
integralgeometric measure, 2054
integration, 6–7, 1106–1118
Integration by Parts
general, 2195
Lebesgue version, 2266
Riemann version, 1161
Integration by Substitution
Lebesgue version, 2268
Riemann version, 1160, 1166
interior, 1447
of a cell, 1850
interior Darboux sum, 1126
interleaving, 1460
Intermediate Value Property, 764
Intermediate Value Theorem, 680, 1608
internal point, 2293
interpolating polynomial, 804
Lagrange form, 805
Newton form, 807
interpolation, 804
Interpolation Property, 933
interpolatory quadrature procedure, 1172
interpretable, 148

intersection, 1914
interval, 314
  closed, 314
  half-open, 314
  open, 314
interval of convergence, 859
intrinsic metric, 955
intuitionism, 98–101
Invariance of Domain Theorem, 1706, 1795
Inverse Function Theorem
  Banach space, 2529
  complex analytic, 1303
  complex analytic multivariable, 1087
  formal, 378
  multivariable, 1058
  real analytic multivariable, 1086
inverse hyperbolic function, 748
inverse image, 159, 161
inverse limit, 1631
  of sets, 1626–1630
  of topological spaces, 1631–1636
inverse limit topology, 1631
inverse system
  of sets, 1626
  of topological spaces, 1631
inverse trigonometric function, 744–745
inversion, 2571
involution
  Banach algebra, 2495
irrational number, 62, 323, 422–433, 738
isolated point, 1444
  of a subset of $\mathbb{R}$, 903
isolated singularity, 1352
isometry
  in a C*-algebra, 2497
isomorphism
  Borel, 1928
  category, 281
  natural, 283
  of convex sets, 2295, 2367
  of vector spaces, 464
isotopy, 1898
iterated integral, 1208
iterated sum of an array, 637
Ivanov, Sergei, 1837
IVP, 1271
IVT, 680

Jabuka, Stanislav, 1439
Jacobson topology, 1498
Jensen's inequality, 2222
Jones's Lemma, 1496
Jones, Vaughan, 35
Jones, William, 2619
Jordan content, 1198
Jordan curve, 1703
Jordan Curve Theorem, 1704
  in $\mathbb{R}^n$, 1705
Jordan region, 1198
Jordan, Camille, 1704
Jordan-Schönflies Theorem, 1705
jump discontinuity, 2196
Junior Monotone Convergence Theorem, 2112
Junior Tonelli Theorem, 2154

Kaczynski, Ted, 2624
κ-additive cardinal, 265
κ-complete filter, 272
Kasner, Edward, 182
Kelley, John, 1429
Kepler, Johannes, 1173
kernel
  of a homomorphism, 2569
Khinchin, Aleksandr, 411
Klein bottle, 1520
Klein, Felix, 69, 106, 685
Kline, Morris, 9, 12, 90
Knaster bucket-handle, 1636, 1738
Kneebone, G. T., 16, 120, 220
knot, 1706, 1745
Knuth, Donald, 234
Kolmogorov quotient, 1509
Kolmogorov, Andrei, 69, 1967, 2010
Körner, T. W., 91
Korovkin's Theorem, 2385
Kovalevskaya, Sofia, 2668
Kramp, Christian, 2618
Krauthammer, Charles, 86
Kronecker, Leopold, 98, 2648
Kummer's Test, 602
Kuratowski closure operation, 1449
Kuratowski's Lemma, 198

Kuratowski, Kazimierz, 267

l'Hôpital's Rule, 778–789
l'Hôpital, Marquis Guillaume de, 778
Lagrange, Joseph-Louis, 417, 434, 760, 2616, 2618
Lakatos, Imre, 58, 664
Lakes of Wada, 1743
λ-system, 1925
Lambert, Johann Heinrich, 45, 422
language, 147
  extended simple, 150
  simple, 148
Laplace equation, 1407
Laplace operator, 1407
Laplace, Pierre-Simon, 59, 716
large cardinal, 260
large inductive dimension, 1798
largest element, 169
lattice, 169
  in $C_{\mathbb{R}}(X)$, 2388
Laugwitz, Detlef, 434, 664
Laurent series
  formal, 376
Lawvere, F. William, 286
Lax-Milgram Theorem, 2431
LCS topology, 2307
least upper bound, 168, 311, 361
Lebesgue Covering Theorem, 1781
Lebesgue Density Theorem, 2275
  in $\mathbb{R}^n$, 2283
Lebesgue integrable function
  nonnegative, 2114
Lebesgue integral, 1112–1117, 2125
Lebesgue measurable, 2028
Lebesgue measurable function, 2028
Lebesgue measurable set, 2003
Lebesgue measure, 5–6, 2003–2004
Lebesgue number, 1559
Lebesgue outer measure, 1998–2000
Lebesgue Property, 1559
Lebesgue, Henri, 265, 1112, 1705, 1952, 1967, 1968, 2002, 2198
Lebesgue-Stieltjes integral, 2190
Lebesgue-Stieltjes measure, 2190
Lefschetz, Solomon, 2615
left inverse, 459
left regular representation, 2593
Legendre differential equation, 2418
Legendre polynomials, 1176, 2417
Legendre, Adrien, 422, 834
Legendre-Gauss quadrature, 1176
Leibniz, Gottfried, 90, 434, 1108, 1110, 2620
lemma, 154
length metric, 955
length of an interval, 315
length structure, 961
  Finslerian, 962
lens space, 1529
Levi, Beppo, 2118, 2407
Levi-Civita number, 377
Lie bracket, 1252
Lie derivative, 1251
Lie group, 2597–2598
Lie, Sophus, 2597
Lilly, William, 54
lim inf, 572–577
lim sup, 572–577
limit
  approximate, 2284
  direct
    of sets, 1622–1626
    of topological spaces, 1630–1631
  inductive, 1623
  inverse
    of sets, 1626–1630
    of topological spaces, 1631–1636
  of a net, 1456
  of a sequence, 561–565, 1453
  of sets, 1915
  projective, 1627
limit cardinal, 246
Limit Comparison Test, 594
limit inferior, 572–577
limit ordinal, 229
limit point, 1444
  of a subset of $\mathbb{R}$, 903
limit superior, 572–577
limited element, 436
Lindelöf space, 1437
Lindemann, Ferdinand, 85, 422, 425
Line Axiom, 929

line integral, 983
Line Segment Axiom, 929
linear algebra, 455–513
linear combination, 465
linear function, 475
linear functional, 2330, 2344
  positive, 2514
linear homotopy, 1895
linear operator, 475
linear transformation, 475–476
  determinant, 504
linearly independent set, 466
link, 1745
Liouville number, 413
Liouville's Theorem, 1321
Liouville, Joseph, 412
Lipschitz function
  between metric spaces, 2288
Listing, Johann, 2617
Little Picard Theorem, 1321
LL (locally Lipschitz), 2288
Lobachevskii, Nikolai, 45
local base
  for a topology, 1434
local dimension, 1807
local extremum, 700
Local Immersion Theorem, 1081
local maximum, 700, 727
local minimum, 700, 727
Local Submersion Theorem, 1079
local uniformizing variable, 1394
locally compact group, 2579
locally compact space, 1568–1573
locally conical, 1853
locally connected space, 1614
locally convex, 2307
locally Euclidean group, 2597
locally finite cover, 1578
locally finite measure, 2035
locally Hausdorff space, 1485
locally Lipschitz, 2288
locally measurable function, 1943
locally measurable set, 1925
locally path-connected space, 1615
locally second countable, 1438
locally σ-finite, 2013
locally T_0 space, 1483
locally T_1 space, 1484
Lochs, Gustav, 411
Loeb measure, 451
log-convex function, 2223
logarithm, 732
  in a differential field, 1421
  natural, 735
logic
  first-order, 133
  higher-order, 133
  intuitionist, 144
  ordinary, 142
  second-order, 133
logicism, 97
long line, 1771
loop, 1896
Łoś's Theorem
  simple version, 449
low-dimensional topology, 1721
lower bound, 168, 311
lower envelope, 1151, 1228
lower semicontinuous decomposition, 1505
lower set, 359
Loyd, Sam, 25
L^p-space, 2470–2493
Lucas, J. R., 58, 92
Lusin, *see* Luzin
Luzin Separation Theorem, 1965
Luzin's Theorem, 2285
Luzin, Nikolai, 1967, 2240, 2655

Mackey Borel structure, 2522
Mackey topology, 2322
MacLane, Saunders, 286, 565
Maddy, Penelope, 181
Mahlo cardinal, 269
manifold, 1764–1776
  closed, 1769
  compact, 1769
  concrete, 1886
  open, 1769
  smooth, 1886
  topological, 1764–1770
    with boundary, 1767–1769
  unbounded, 1769

manifold with boundary
topological, 1767–1769
map
essential, 1779
mapping torus, 1523
Marcinkiewicz, Józef, 1599
Martin's Axiom, 275
Martin, George E., 51
mathematical dishonesty, 123
mathematical honesty, 123
mathematics
applied, 83
pure, 83
matrix, 455
addition, 455
classical adjoint, 508
determinant, 496–510
expansion by minors, 501–502
geometric interpretation, 504–508
multiplicative property, 502–503
identity, 458
inverse, 459, 508
multiplication, 457
by blocks, 459
transpose, 459
matrix algebra, 458
matrix ring, 458
matrix units, 2523
Max-Min Theorem, 682
maximal element, 169
maximal Hausdorff quotient, 1511
maximal ideal, 205
maximal pseudometric, 918
maximal T_0 quotient, 1508
maximum
endpoint, 727
local, 727
May, Kenneth O., 76, 110
McCarty, D. C., 101
meager set, 1596
Mean Value Inequality, 763, 1311
mean value property, 1410
Mean Value Theorem, 758
for integrals, 1143, 1150, 1158, 2193
mean-square bounded, 2148
measurable box, 2067
measurable cardinal, 271
measurable function, 1927, 1940
Borel, 1944–1951
measurable rectangle, 2062, 2065, 2067
measurable set, 2000
measurable space, 1927
measurable subspace, 1929
measurably separable, 1847
measure, 2–6, 1970–1985
atomless, 268
Borel, 2035, 2475
bounded-finite, 2039
compression, 1983
counting, 1971
decomposable, 1979
dilation, 1984
downward continuous, 1974
extension, 1984
finite, 1978
finitely additive, 1971
full real-valued, 265
Haar, 2587
Hausdorff, 2044
inner, 2017
integralgeometric, 2054
Lebesgue, 5–6, 2003–2004
locally finite, 2035
Loeb, 451
outer, 1996
packing, 2050
probability, 1978
product, 2064
push-forward, 1985
Radon, 2039
restriction, 1983, 1984
semifinite, 1980
σ-finite, 1978
spherical, 2046
tight, 2039
upward continuous, 1973
zero, 1971
measure space, 1970
complete, 1976
completion, 1977
measure zero
in $\mathbb{R}^n$, 1832

Menger sponge, 1734
Menger universal curve, 1734
meromorphic function, 1360
Mersenne, Marin, 139
mesa function
  smooth, 887
Method I, 1998
Method II, 2042
method of undetermined coefficients, 1175
metric, 908
  controlled, 2070
  discrete, 1433
  doubling, 2055
  extended, 917
  homogeneous, 2055
  intrinsic, 955
  length, 955
  quotient, 1510
  strict length, 955
  strictly intrinsic, 930, 955
metric dimension, 2058
metric outer measure, 2034
metric space, 908
  of a measure space, 1989–1990
metrizable space, 1433, 1536–1540
midpoint, 934
Midpoint Axiom, 945
midpoint rule, 1169
midpoint-convex function, 2213
Mill, John Stuart, 60
Milnor, John, 1307, 1788
minimax approximation, 817
minimum
  endpoint, 727
  local, 727
Minkowski functional, 2298
Minkowski's Inequality, 2473, 2493
Minkowski, Hermann, 46
Möbius strip, 1518
Möbius, August Ferdinand, 2622
Möbius, Paul Julian, 2622
model
  standard, 222
  transitive, 222
module, 522
  flat, 527
  free, 523
  projective, 542
modus ponens, 143
Mollerup, J., 836
Mongré, Paul, 2620
Monodromy Theorem, 1392
monoid, 2565
monomorphism, 279
monotone class, 1925
Monotone Class Theorem, 1927
Monotone Convergence Theorem, 2117
  Baby, 1973
  Downward, 2120
  for Riemann integration, 1226
  for series, 2119
  for step functions, 1226
  Junior, 2112
monotone function, 2196
monotone regular outer measure, 2015
monotone sequence
  of functions, 1595
Montel's Theorem, 1349
Montgomery, Deane, 2597
Moore manifold, 1772, 1773
Moore method, 1461
Moore space, 1772
Moore, Calvin C., 1461
Moore, Eliakim Hastings, 1461, 2616
Moore, Gregory H., 203
Moore, Robert L., 1461, 1772
Moore-Smith convergence, 1461
Morera's Theorem, 1320
Morse Lemma
  analytic one-dimensional, 884
  complex one-dimensional, 1362
  one-dimensional, 799
Morse, Anthony P., 1831
Morse, Marston, 86, 799, 884, 1831
Morse-Sard Theorem, 1831
Mostowski collapsing map, 451
MS, 1847
multi-index, 1002
multi-index notation, 1099, 1103
multilinear map, 514
multiple
  of a representation, 2512

Multiplicative Axiom, 197
multiplicity function, 1955
multivariable calculus, 987
multivariable power series, 1099

Nagata-Smirnov Theorem, 1537
Napier, John, 54, 2617
Nash, John, 2624
natural embedding, 2350
natural isomorphism, 283
natural logarithm, 735
natural number, 298–300, 327–344
negation, 139–142
negative part of a function, 2135
neighborhood, 1432
    of a point of $\mathbb{R}$, 903
neighborhood filter, 1465
Nemes, Gergő, 845
net, 1455–1461
    Cauchy, 1661
    interleaving, 1460
    universal, 1468
Neuman, Alfred E., 46
Newman, James, 182
Newton, Isaac, 81, 434, 807, 1108, 1173, 2617, 2620, 2624
Newton-Cotes formula, 1173
    open, 1175
Nikodým Convergence Theorem, 2102
Nikodým, Otto, 2101
Nikolskii, Sergei, 2621
Nine Lemma, 1911
Niven, Ivan, 422
No-Contraction Theorem, 1777
No-Retraction Theorem, 1777
Nöbeling universal space, 1816
nodal set, 273
Noether, Emmy, 2457, 2615
Noether, Fritz, 2457
Noether, Max, 2457
nonanalytic point, 885
    type (C), 885
    type (P), 885
nonarchimedean ordered field, 375, 437
nondecreasing function, 2196
nondecreasing sequence
    of functions, 1595
nondegenerate
    representation, 2513
nonincreasing function, 2196
nonincreasing sequence
    of functions, 1595
nonmeasurable set, 2024–2026
nonnegative infinite series, 590
nonoverlapping, 1189
nonstandard analysis, 434
nonstandard extension
    of a function, 443
    of a relation, 445
    of a set, 441
nonstationary, 949, 967
normal
    element of a C*-algebra, 2497
    fully, 1490
normal convergence
    for series, 854
    of sequences, 665, 1331
normal family, 1348
normal filter, 273
normal space, 1482, 1486–1489
normal subgroup, 2568
nowhere dense set, 1596
NSS group, 2597
nuclear dimension, 1805
null set, 1147, 1975
null space
    of a linear transformation, 477
nullhomotopic, 1895
number
    algebraic, 191, 412, 426
    algebraically independent, 426
    Cayley, 388
    complex, 380–388
    decadent rational, 399
    dyadic rational, 358
    hypernatural, 443
    hyperreal, 438
    integer, 305–308, 345–355
    irrational, 62, 323, 422–433, 738
    Levi-Civita, 377
    Liouville, 413
    natural, 298–300, 327–344

p-adic, 380
prime, 308, 342
rational, 309, 355–358
real, 291–323, 359–380
transcendental, 191, 412, 422–433
von Neumann, 420
number line, 310
numerical analysis, 1167
numerical integration procedure, 1168
numerical integration process, 1168
numerical range, 2451

obvious, 21
ODE, 1269
oganesson, 95
one-point compactification, 1569, 1590
one-sided derivative, 724–725
open box, 1478
open cover, 1437, 1554
decomposable, 1805
open interval, 314
open manifold, 1769
Open Mapping Theorem, 2359
for σ-compact groups, 2584
for locally compact groups, 2581
Smooth, 1080, 1094
open set, 1432
basic, 1434
in $\mathbb{R}$, 901
regular, 1452
operation, 130
operator, 476
bounded, 2327
bounded below, 2449
compact, 2458
completely continuous, 2458
Fredholm, 2460
positive, 2452
strictly singular, 2460
weakly continuous, 2459
operator algebra, 2331
Oppenheimer, Frank, 84
orbit, 1528
orbit space, 1528
order
of a zero, 1351
order ideal, 1932
prime, 1934
order of a cover, 1801
order of precision, 1171
ordered field, 297, 356, 436
nonarchimedean, 437
ordering
complete, 168
equivalent, 167
induced, 167
ordinary, 167
partial, 167–176
strict, 167
total, 167
Ordering Principle, 201
ordinal, 227–236
limit, 229
successor, 229
ordinary differential equation, 1269
ordinary ordering, 167
Oresme, Nicole, 593, 2619
orthogonal dimension, 2415
orthogonal projection operator, 2419
orthogonality
in Hilbert space, 2408
orthonormal basis, 2413
orthonormal set, 2410
oscillation, 1109, 1128–1129, 1541
weighted local, 1129
osculating polynomial, 813
osculatory quadrature procedure, 1172
Osgood's Lemma, 1404
Osgood, William, 1597, 1698
Oughtred, William, 2619, 2620
outer measure, 1996
Borel regular, 2018
Lebesgue, 1998–2000
metric, 2034
monotone regular, 2015
regular, 2015

packing dimension, 2052
packing measure, 2050
p-adic number, 380
parabolic rule, 1175
paracompact space, 1578–1583

paradox
Burali-Forti, 227
Russell's, 193
paradoxical decomposition, 551
paradoxical set, 552
Parallel Postulate, 45
Parallelogram Axiom, 935
Parallelogram Law, 2402
parametrization by arc length, 954, 977
parametrized curve
in $\mathbb{R}^n$, 966
in a metric space, 947
piecewise-smooth, 971
smooth, 971
paranorm, 2308
parentheses, 131
Parseval's relation, 2412
partial derivative, 995
partial differential equation, 1269
partial isometry
in a C*-algebra, 2497
partial ordering, 167–176
strict, 167
partial sum, 582
partially ordered set, 167–176
partition, 1122–1123
between closed sets, 1487
full, 1222
of a rectangle, 1189
partition dimension, 1806
partition of unity, 1580–1583
Pasch, Moritz, 47
path, 1614
in $\mathbb{R}^n$, 968
in a metric space, 949
path component, 1615
path-connected space, 1614
Paulos, John, 15
PDE, 1269
Peano Axioms, 327
Peano's Existence Theorem for First-Order ODEs, 1284
Peano, Giuseppe, 327, 1696
Peary, Robert, 95
Pedersen, Gert K., 19
Pell's equation, 417
Penrose, Roger, 46, 95
Perelman, Grigori, 2624
perfect subset, 1450
perfectly normal, 1552
Perron integral, 1118
Picard Existence Theorem, 1275
Picard iteration, 1277
Picard's Theorem
Big, 1358
Little, 1321
Picard-Lindelöf Theorem, 1275
Pick's Theorem, 1382
piecewise-linear map, 1862
Pierpont, James, 182
π-system, 1925
PL map, 1862
Plato, 58
Platonism, 78, 92–96
Playfair's Axiom, 45
Poincaré homology sphere, 1530
Poincaré, Henri, 20, 45, 46, 76, 84, 95, 104
point, 464
internal, 2293
singular, 885
point finite cover, 1578
point mass, 1971
point of inflection, 767
point spectrum, 2449
pointwise bounded sequence of functions, 1949
Polarization Identity, 2402
pole, 1354
simple, 1354
Polish group, 2582
Polish space, 1548, 1962
Polya, George, 73, 107, 2567
polydisk, 1100
polyhedron, 1852
noncompact, 1855
Pontryagin dual group, 2599
Pontryagin Duality Theorem, 2601
Pontryagin, Lev, 2597
positive
element of a C*-algebra, 2507
linear functional, 2514
operator, 2452
positive part of a function, 2135

Postnikov, Mikhail, 106
Poulsen simplex, 2380
Power Rule, 693, 697
    complex, 1302
power series, 857–866
    formal, 375
    multivariable, 1099
    unordered, 874
pre-Hilbert space, 2406
preimage, 159, 161
premeasure, 2005
prime filter, 1934
prime ideal, 205
prime number, 308, 342
primitive, 770
primitive ideal, 2520
principal ideal, 205
Principle of Induction, 300
Principle of Recursion, 304
Principle of the Excluded Middle, 63
Pringsheim, Alfred, 87, 128, 898, 2640
Prisoner's Dilemma, 153
probability, 6
probability measure, 1978
product σ-algebra, 2062, 2067
product measure, 2064
product of infinite series, 646
Product Rule, 694
    complex, 1302
    dot product, 1011
    multivariable, 1010
product topology, 1476, 1479
projectable region, 1212, 1214
projection, 2419
    in a C*-algebra, 2497
projective limit, 1627
projective module, 542
projective plane, 1522
projective space
    complex, 1529
    real, 1529
proof, 57
    by contradiction, 62–63
    formal, 57
proper map, 1585
Property (N), 2240
Property (S), 2241
proposition, 132, 154
Prüfer, Heinz, 1772
Prüfer manifold, 1772
p-series, 594
pseudo-arc, 1636, 1741
pseudo-base, 1442
pseudo-circle, 1743
pseudocompact space, 1577, 1813
Puiseux series, 376
pure mathematics, 83
pure set, 239
push-forward
    of a measure, 1985
Putnam, Hilary, 102
Pythagoras, 62
Pythagorean Axiom, 935
Pythagorean Theorem, 2409

quadrature procedure, 1168
    composite, 1172
    interpolatory, 1172
    osculatory, 1172
    regular, 1168
    simple, 1172
quantifier, 135–137
    existential, 135
    negation, 141–142
    restricted, 136
    separated, 146–147
    universal, 135
quasi-compact mapping, 1504
quasi-component, 1610
quasi-inverse, 2461
quaternion, 387–388
quotient field, 358, 376
quotient metric, 1510
quotient ring, 353
Quotient Rule, 696
    complex, 1303
quotient topology, 1502

Raabe's Test, 603
    Limit Version, 604
    Second Order, 606
Rademacher's Theorem, 2290

radial set, 2293
radius of convergence, 858, 1335
Radó, Tibor, 1772
Radon measure, 2039, 2189
Radon-Nikodým derivative, 2174
Radon-Nikodým Theorem, 2176
    for positive linear functionals, 2518
Raikov-complete, 2578
Ramanujan constant, 427
Ramsey cardinal, 274
Ramsey, Frank, 2621
range
    of a linear transformation, 477
Range-Straightening Theorem, 1081
Rank Theorem, 1081
Ratio Test, 600, 614
    Basic, 600
rational function, 358, 374, 509
rational number, 309, 355–358
rational ruler, 1758
Reagan, Ronald, 43
real line
    extended, 324
real number, 291–323, 359–380
real part
    of an element of a C*-algebra, 2497
real-closed field, 323, 358, 377
real-valued measurable cardinal, 266
realism, 78, 92–96
rearrangement, 625
Recorde, Robert, 15, 2618
rectangle, 1188
rectangular sum of an array, 636
rectifiable curve, 950, 972
rectifiably path-connected, 956
recursion, 304–305, 331–332
    transfinite, 172–174
reductio ad absurdum, 62–63
Rees, Mina, 19
refinement, 1579
reflexive Banach space, 2338, 2351
region, 1384
regular boundary, 1378
regular cardinal, 259
regular open set, 1452
regular outer measure, 2015
regular point, 1079
regular quadrature procedure, 1168
regular representation, 2593
regular Riemann Surface, 1392
regular space, 1482, 1485–1486
regulated function, 775
relation, 212
    equivalence, 163
relational system, 147
relative closure, 1439
relative extremum, 700
relative maximum, 700
relative minimum, 700
relative topology, 1438
relatively open, 1439
religion, 55
removable singularity, 1353
reparametrization, 948, 967
representation, 2512
    faithful, 2512, 2593
    GNS, 2517
    irreducible, 2512
    nondegenerate, 2513
    of a *-algebra, 2594
    of a group, 2593
    of an algebra, 2593
    trivial, 2593
    unitary, 2594
representing measure, 2372
residual set, 1596
residual spectrum, 2449
residue, 1363
Residue Theorem, 1363
resolvent function, 2438
resolvent set, 2437
resonance, 2364
restricted direct product, 1639
restricted quantifier, 136
restriction
    of a measure, 1929, 1983, 1984
retraction, 1864
revision of an infinite series, 629
    drastic, 642
Ribenboim, Paulo, 84
Riemann Hypothesis, 427

Riemann integral, 6–7, 1106–1107, 1111, 1121–1166
improper, 1163
in $\mathbb{R}^n$, 1188–1221
Riemann Mapping Theorem, 1375
Riemann sum, 1109, 1123
Riemann Surface
of a function, 1385
regular, 1392
Riemann, Bernhard, 6, 45, 629, 1106
Riesz decomposition, 2383
Riesz Representation Theorem, 1117
for L^p, 2481
Riesz, Frigyes, 1448, 2405, 2449
Riesz-Frèchet theorem, 2421
right inverse, 459
rigid set, 256
Rigidity Theorem, 1351
for harmonic functions, 1411
ring, 351, 455
commutative, 351
division, 388
integral domain, 355
matrix, 458
of fractions, 358
of subsets, 1917
ordered, 353
quotient, 353
unital, 351
ring of adeles, 1639
ring of fractions, 358
Robinson, Abraham, 434, 664
Rolle's Theorem, 759
Rolle, Michel, 759, 2619
Romberg integration, 1184, 1187
root
of a number, 321
Root Test, 599, 613
Basic, 599
Roth, Klaus, 414
Rotman, Brian, 16, 220
Rouché's Theorem, 1367
row vector, 456
Rowbottom, Frederick, 273
Rucker, Rudy, 235
Rudin, Mary Ellen, 1579
Runge phenomenon, 815
Russell set, 256
Russell socks, 256
Russell's Paradox, 193, 208
Russell, Bertrand, 21, 79, 87, 97, 105, 116, 185, 197

saddle point, 1040
Saint Raymond, Jean, 1078
Saks, Stanisław, 728, 2101
Sard, Arthur, 1831
saturated, 1503
Scaling Test, 622
scattered, 1450
Schmidt, Erhard, 2405
Schmieden, Curt, 434
Schneider, Theodor, 427
Schönflies, Arthur, 1705, 2405
Schröder, Ernst, 190
Schröder-Bernstein Theorem, 190, 221
for measurable spaces, 1930
Schumacher, Carol, 17
Schwarz Inequality, 2401
Schwarz lantern, 1096
Schwarz Lemma, 1371
Schwarz Reflection Principle, 1361
Schwarz, Hermann, 2401
Scott, Dana, 276
secant, 743
second category set, 1596
second countable
locally, 1438
second countable space, 1435
Second Decomposition Theorem, 1801
second derivative, 703
Second Derivative Test, 767
multivariable, 1036
second differences, 716
second-order logic, 133
second-order statement, 133
second-order theory, 133, 328
section, 1250
self-adjoint
element of a C*-algebra, 2497
semialgebra, 1916
semicanonical form, 2109

semiclosed map, 1513
semifinite measure, 1980
semigroup, 2565
  abelian, 354, 2565
semiring, 1916
sentence, 132
separable topological space, 1446
separated
  completely, 2343
  fully, 2343
separated sets, 1536
separated union, 1440
separation, 2341, 2343
sequence, 560
  bounded, 561
  Cauchy, 368, 578
  convergent, 562
  divergent, 562
  limit, 561–565
  of functions, 664–668
    monotone, 1595
    nondecreasing, 1595
    nonincreasing, 1595
    pointwise bounded, 1949
sequence of first differences, 716
Sequential criterion
  for continuity, 679
  for limits, 673
  for one-sided limits, 676
sequential space, 1469
sequentially closed, 1469
sequentially compact, 1557
Serre, Jean-Pierre, 92
set
  convex, 2294
  locally measurable, 1925
  pure, 239
  rigid, 256
set-theoretic difference, 1914
sex, 84
Shaw, George Bernard, 8
Shelah, Saharon, 275
shrinking of a cover, 1487
Siegel, Carl, 427, 429
Sierpiński carpet, 1730
Sierpiński gasket, 1730
Sierpiński sieve, 1730
Sierpiński space, 1493
Sierpiński triangle, 1730
σ-finite measure, 1978
σ-ring, 1923
σ-algebra
  product, 2062, 2067
σ-algebra, 1920
σ-compact space, 1567
σ-field, 1920
Silver, Jack, 275
simple function, 1114, 2108
simple language, 148
simple pole, 1354
simple quadrature procedure, 1172
Simple Transfer Principle, 448
simplex, 1851
simplicial complex, 1857
  abstract, 1859
simplicial isomorphism, 1857
simplicial map, 1857
simply connected, 1372
Simpson's rule, 1170
Simpson's three-eighths rule, 1185
Simpson, Thomas, 1173
sine, 741
singular cardinal, 259
singular point, 885, 1389
singularity, 1352
  essential, 1357
  pole, 1354
  removable, 1353
skeletal category, 284
skeleton, 284, 1857
skew field, 388
Skolem's Paradox, 223
Skolem, Thoralf, 208
small category, 278
small inductive dimension, 1796
smooth bump function, 886
smooth Heaviside function, 887
smooth mesa function, 887
Smooth Open Mapping Theorem, 1080, 1094
smooth partition of unity, 887
Smooth Urysohn Lemma, 888
Snake Lemma, 1906

Sobolev space, 2562
solenoid, 1635, 1736
Solovay, Robert, 268, 275
sophomore's dream, 855
Sorgenfrey line, 1496
Souslin, *see* Suslin
space-filling curve, 1696
    ultimate, 1696
span, 465
Spectral Mapping Theorem
    for inverses, 2440
    for polynomials, 2439
    for rational functions, 2441
spectral resolution, 2469
Spectral Theorem
    Finite-Dimensional, 2468
spectrum
    of an element of a Banach algebra, 2437
    of an operator, 2449
spherical measure, 2046
Spivak, Michael, 18
square root
    in a C*-algebra, 2507
square sum of an array, 635
square-integrable, 2148
stable algebra, 2523
stable C*-algebra, 2523
stably isomorphic C*-algebras, 2523
Stalin, Josef, 1967
standard Borel space, 1957, 1962
standard model, 222
standard revision of an unordered power series, 875
star, 1855
star refinement, 1490
*-representation, 2594
state, 2514
    pure, 2516
    vector, 2515
state space, 2514
statement
    atomic, 132
    closed, 137–139
    first-order, 133
    formal, 129–147
    second-order, 133
stationary set, 1814
step function, 770, 1222
Stewart, Ian, 56, 81, 678, 1270
Stirling numbers, 723
Stirling's formula, 841
Stokes' Theorem
    for rectangles, 1265
Stone-Čech compactification, 1575–1578
Stone, Arthur H., 1579
Stone, Marshall, 1575, 2386
Stone-Weierstrass Theorem, 1814, 2386
Story, William, 2622
Straightening Theorem, 1082
strict contraction, 926
strict length metric, 955
strict ordering, 167
strict partial order, 167
strict power contraction, 927
strictly convex, 2427
strictly decreasing, 2196
strictly decreasing function, 2196
strictly increasing function, 2196
strictly intrinsic metric, 930, 955
strictly measurable function, 1943
strictly monotone function, 2196
strictly singular operator, 2460
Strong Law of Large Numbers, 411
strongly compact cardinal, 276
strongly continuous mapping, 1503
strongly inaccessible cardinal, 259
strongly infinite-dimensional space, 1821
Struik, Dirk, 2621
subadditivity
    countable, 1972, 1996
    finite, 1972, 1996
subarrangement, 625
subbase, 1436
subcomplex, 1857
subcover, 1437, 1554
subdivision
    of a complex, 1857
subgroup, 2567
    normal, 2568
submultiplicative norm, 2432
subnet, 1456–1458
    weak, 1458

subordinate filterbase, 1467
subrepresentation, 2512
subseries, 625
subspace, 464
    Borel, 1929
    complementary, 472
    measurable, 1929
subtree, 1547
successor cardinal, 246
successor ordinal, 229
Suiseth, Richard, 655
sum
    of representations, 2512
sum of an infinite series, 583
Sum Rule, 693
    complex, 1302
    multivariable, 1010
Sum Theorem
    for ind, 1799
summation convention, 540
superset property, 1465, 1932
superstructure, 449
support, 1581
    compact, 1573
    finite, 2478
    of σ-finite measure, 2478
    of almost finite measure, 2479
    of finite measure, 2478
support hyperplane, 2298
supremum, 168, 311, 361
surgery, 1524
    on a knot, 1527
Suslin operation, 1966, 2002, 2018
Suslin scheme, 2019
    monotone, 2019
Suslin set, 1966, 2019
    in $\mathbb{R}^n$, 2020
Suslin, Mikhail, 1967, 2621
swelling, 1488
Swiss cheese, 1732
Sylvester, James J., 82, 685
symmetric difference, 1914

T-finite set, 194
tail
    of an infinite series, 584
tail set, 1459
tangent, 743
tangent bundle, 1890
    in $\mathbb{R}^n$, 1249
tangent space, 1249, 1890
Tao, Terry, 74
tautology, 142
Taylor series, 866, 1341
Taylor's Theorem
    Cauchy remainder formula, 796, 801
    integral remainder formula, 797
    Lagrange remainder formula, 796
    multivariable, 1031, 1032
    single variable, 793
Taylor, Brook, 791, 804
Teichmüller-Tukey Lemma, 198
Teichmüller, Oswald, 198
tensor algebra, 528
tensor product
    of Hilbert spaces, 2423
term
    compound, 131
    formal, 131
    simple, 131
tetration, 234
theorem, 154
Thue, Axel, 414
Thurston, William, 8
Tietze Extension Theorem, 1492
    for $\mathbb{R}$, 906
tight, 2146
tight measure, 2039
Tikhonov plank, 1588
Tikhonov space, 1491
Tikhonov's Theorem, 199, 1562–1566, 1588
Titchmarsh, E. C., 7, 422
Tonelli's Theorem, 2155
    Baby, 1834
    for double summations, 637
    Junior, 2154
Tonelli, Leonida, 640
topological dual space, 2317
topological group
    complete, 2578
    locally compact, 2579
    locally Euclidean, 2597

NSS, 2597
Polish, 2582
topological space, 1432–1442
arcwise connected, 1614
compact, 1437, 1554–1566
compactly generated, 1583
completely collectionwise normal, 1678
completely Hausdorff, 1499
completely normal, 1536
completely regular, 1491–1492
connected, 1605
continuum, 1742
contractible, 1895
discrete, 1433
first countable, 1435
fully normal, 1490
functionally Hausdorff, 1499
Hausdorff, 1482, 1485
hereditarily normal, 1536
indiscrete, 1433
Lindelöf, 1437
locally compact, 1568–1573
locally connected, 1614
locally Hausdorff, 1485
locally path-connected, 1615
locally T_0, 1483
locally T_1, 1484
metrizable, 1433, 1536–1540
normal, 1482, 1486–1489
paracompact, 1578–1583
path-connected, 1614
Polish, 1548, 1962
pseudocompact, 1577, 1813
regular, 1482, 1485–1486
scattered, 1450
second countable, 1435
separable, 1446
sequential, 1469
T_0, 1482
T_1, 1482
T_2, 1482
$T_{2\frac{1}{2}}$, 1499
T_3, 1482
$T_{3\frac{1}{2}}$, 1491
T_4, 1482
T_5, 1536
T_6, 1552
totally disconnected, 1611
totally normal, 1496
totally path disconnected, 1615
totally separated, 1611
Urysohn, 1499
zero-dimensional, 1611
topological vector space, 2303
topologically complete, 1662
topologist's sine curve, 1634, 1726
topology, 1432
countable complement, 1441
defined by a metric, 1433
discrete, 1433
finite complement, 1435
indiscrete, 1433
inductive limit, 1630
product, 1476, 1479
relative, 1438
TVS, 2303
weak, 2346
weak-*, 2351
topos, 286
torus, 1520
total analytic function, 1384
total derivative, 1007
total differential, 1007
total ordering, 167
totally bounded, 1672
totally disconnected space, 1611
totally inhomogeneous, 1759
totally normal space, 1496
totally ordered set, 167
totally path disconnected space, 1615
totally separated space, 1611
transcendental number, 191, 412, 422–433
Transfer Principle, 445
Extended First-Order, 448
Extended Simple, 448
First-Order, 448
Simple, 448
transfinite dimension, 1825
transfinite induction, 171–174, 229–230
transitive closure, 218, 241
transitive model, 222
transitive set, 227, 240

translation, 2571
transpose, 459
trapezoid rule, 1169
tree, 1547
tree of mathematics, 80
triangle inequality, 298, 2473
triangular sum of an array, 636
triangulation
  of a polyhedron, 1857
  of a topological space, 1860
  weak, 1860
trigonometric function, 740–751
  inverse, 744–745
trigonometric polynomial, 2392
trivial representation, 2593
Truss, J. K., 14, 291
truth table, 135
Tucker, A. W., 1783
Tukey's Lemma, 198
Tukey, John, 198, 1656
Turing computable, 39
TVS topology, 2303
Tychonoff, *see* Tikhonov

u.c. convergence, 665, 1680
UFD, 431
Ulam matrix, 267
Ulam, Stanisław, 267
ultimate space-filling curve, 1696
ultrafilter, 272, 1932
  fixed, 272
  free, 272, 1933
  normal, 273
  principal, 1933
Ultrafilter Property, 201
unbounded manifold, 1769
uncountable set, 185
uniform bidual, 2323
Uniform Boundedness Theorem, 2364
uniform covering, 1656
uniform space, 1650
  compact, 1672
  precompact, 1673
  totally bounded, 1672
uniform structure, 1650
  on a topological group, 2576
uniform topology, 2323
uniformity
  covering, 1656
  fine, 1660
uniformly absolutely continuous family, 2100
uniformly complete, 1668
uniformly convex, 2427
uniformly countably additive family, 2100
uniformly integrable, 2142
unilateral shift, 2451
union, 1914
unique factorization domain, 431
unitary
  in a C*-algebra, 2497
unitary representation, 2594
unitization, 2433
universal net, 1468
universal property, 1623, 1627
universal quantifier, 135
universe
  Gödel constructible, 239
  von Neumann, 239
unlimited element, 436
unordered power series, 874
unordered sum, 630
upper bound, 168, 311
upper envelope, 1151, 1228
upper semicontinuous decomposition, 1505
upper set, 359
upward continuity, 1973
urelement, 449
Urysohn Metrization Theorem, 1537
Urysohn space, 1499
Urysohn's Lemma, 1489
  Smooth
    for $\mathbb{R}$, 888
Urysohn, Pavel, 2621

vacuous, 303
Vandermonde matrix, 804
variable, 130
  bound, 136
  free, 135
Veblen, Oswald, 1704
vector, 464
  column, 456

linear independence, 466
row, 456
vector calculus, 987
vector field, 1250
vector space, 462–475
basis, 468
dimension, 470
finite-dimensional, 466
isomorphism, 464
quotient, 473–475
subspace, 464
complementary, 472
topological, 2303
vertex, 1849, 1856
vertical arrow notation, 234
Viéte, François, 648
Vietoris topology, 1551
Vietoris, Leopold, 2621
Villani, Cédric, 55, 56, 1293
Vitali Convergence Theorem
finite case, 2144
general case, 2146
Vitali cover
of a metric space, 2042
Vitali, Giuseppe, 265, 2024, 2101
Vitali-Hahn-Saks Theorem, 2102
volume, 1188
von Neumann algebra, 2331
von Neumann numbers, 420
von Neumann universe, 239
von Neumann, John, 46, 89, 208, 329, 669, 1999, 2069, 2597, 2616, 2624

Wagon, Stan, 221
Wallis, John, 99, 129, 648, 2616, 2618, 2620, 2667
Wallman compactification, 1578
Waring, Edward, 805
Warsaw circle, 1727
weak derivative, 2542
weak equivalence, 284
Weak Euclidean Four-Point Axiom, 934
weak subnet, 1458
weak topology, 2346
weak triangulation, 1860
weak-* topology, 2351
weakly inaccessible cardinal, 259
weakly infinite-dimensional space, 1821
weakly n-dimensional, 1823
wedge sum, 1517
Weierstrass Approximation Theorem, 2384, 2620
general, 2390
Weierstrass M-test, 853
Weierstrass, Karl, 14, 425, 665, 2384, 2617, 2620
weight
of a topological space, 1436
weighted local oscillation, 1129
Weil completion, 2578
Weil, André, 74, 105, 117, 274
well defined, 145
well ordering, 169–171
well-ordered set, 169–171
Well-Ordering Principle, 197
Weyl, Hermann, 88, 105
Whewell, William, 102
Whitehead continuum, 1756
Whitehead manifold, 1757
Whitehead, Alfred North, 57, 108, 128
Whitehead, John H. C., 1756
Whitney, Hassler, 1841
Widmann, Johannes, 2618
Wigner, Eugene, 83
Wikipedia, 123
wild arc, 1747
Wiles, Andrew, 76
Wolf, Abraham, 120
Wright brothers, 96

Yamabe, Hidehiko, 2597
Yegorov, *see* Egorov
Yermakov's Test, 609
Young's Inequality
for products, 2472
General, 2490
Young, Grace Chisholm, 728
Young, William H., 2472

Zahorski's Theorem, 899
Zariski topology, 1498
Zermelo, Ernst, 185, 208
zero content, 1147, 1200

zero measure, 1971
zero set, 887
zero-dimensional space, 1611
ZF set theory, 208
ZFC set theory, 208
Zippin, Leo, 2597
Zorich, Vladimir, 65
Zorn's Lemma, 198, 230–231